

Content & User Behavior in Anonymous Hyperlocal Online Platforms

von der Fakultät 1 - MINT -
Mathematik, Informatik, Physik, Elektro- und Informationstechnik
der Brandenburgischen Technischen Universität Cottbus–Senftenberg genehmigte
Dissertation zur Erlangung des akademischen Grades eines Dr. rer. nat.

vorgelegt von

Jens Helge Reelfs

geboren am 13.02.1985 in Sande

| | |
|-----------------------------|---------------------------------------|
| Vorsitzender: | Prof. Dr. rer. nat. habil. Klaus Meer |
| Gutachter: | Prof. Dr. rer. nat. Oliver Hohlfeld |
| Gutachter: | Prof. Dr. Markus Strohmaier |
| Gutachter: | Prof. Dr.-Ing. Andriy Panchenko |
| Tag der mündlichen Prüfung: | 31.03.2023 |

Faculty of Mathematics, Computer Science, Physics,
Electrical Engineering and Information Technology

**Brandenburgische Technische Universität
Cottbus-Senftenberg**

Chair of Computer Networks

Dr. rer. nat. Thesis

***Content & User Behavior
in Anonymous Hyperlocal
Online Platforms***



Jens Helge Reelfs

✉ helge@reelfs.de

🌐 h.reelfs.de

examined by

Prof. Dr. rer. nat. Oliver Hohlfeld

Prof. Dr. Markus Strohmaier

Prof. Dr.-Ing. Andriy Panchenko

March 31, 2023

ACKNOWLEDGEMENTS

For my parents, providing me with this opportunity - they never had.

Thanks to Oliver - mentor and friend.

Thanks to Niklas, Esso and the Jodel team - providing us with data and first-hand insights.

Thanks to other family, friends and colleagues - keeping me busy.

Thank science.

YES, I AM A CRIMINAL.
MY CRIME IS THAT OF CURIOSITY.
(Hacker Manifesto, 1986¹)

¹https://en.wikipedia.org/wiki/Hacker_Manifesto

KURZFASSUNG

Heutzutage hat sich die alltägliche Kommunikation im digitalen sozialen Leben durch die Pandemie nur noch verstärkt. Kommunikation und Informationen sind verfügbar über verschiedene (direkte) Nachrichtenplattformen. Der immer stärker werdende Einfluss auf den öffentlichen Diskurs und Gesellschaft wird weitläufig akzeptiert. Während die meisten Plattformen auf Nutzerprofile setzen und damit soziale Anerkennung ermöglichen, bleiben andere anonym. Eine neue Art von Anwendung kombiniert Anonymität mit einem starken räumlichen Fokus, der Hyperlokalität. Auswirkungen dieser einzigartigen Kombinationen auf die Plattform weitgehend unbekannt.

In dieser Arbeit zeigen wir eine erste datengetriebene ganzheitliche Sicht auf die Jodel Plattform, die beide Eigenschaften kombiniert. Wir nutzen Ground-Truth-Informationen, um eine Fülle von Gemeinschaften in zwei verschiedenen Ländern zu analysieren: Deutschland und dem Königreich Saudi-Arabien. Das Werk folgt einer Plattform-Perspektive bestehend aus vier großen Bereichen rund um das Individuum.

Wir beginnen mit einem breit angelegten Diskurs zu drei **◆ USER ADOPTION**-Prozessen verschiedener Anwendungen. Nachdem wir Messungen des zur Nutzerakzeptanz der deutschen Anwendung zur digitalen Ermittlung von Kontaktpersonen erörtert haben, zeigen wir auf, wie etablierte Plattformen als Seitenkanal zur Umgehung von Zensur im aktuellen russisch-ukrainischen Hybridkrieg genutzt werden. Ferner zeigen wir insbesondere, dass identische Plattformeigenschaften bei der Nachrichtenanwendung Jodel zu sehr unterschiedlicher Nutzung führen können.

Da jede Online-Plattform auf **◆ USER INTERACTIONS** aufbaut, charakterisieren wir im Weiteren das Nutzerverhalten auf Jodel in beiden Ländern. Wir erörtern strukturelle und Nutzungsunterschiede, die ausschließlich auf das lokale Nutzerverhalten zurückzuführen sind. Mit einem detaillierten Blick in die saudische Gemeinschaftslandschaft schließen wir eine geografische und gesellschaftliche Forschungslücke zur Plattformnutzung.

Darüber hinaus erörtern wir **◆ USER CONTENT** und analysieren die Informationsverbreitung. Wir entwickeln ein mehrdimensionales Klassifizierungsschema für Intentionen (warum) und Themen (was) innerhalb sozialer Medien und stellen unseren Crowdsourced-Vergleich saudi-arabischer Inhalte vor. Mit neuronalen Worteinbettungen als Werkzeug diskutieren wir quantitative und qualitative Erkenntnisse zu Wort-Emoji-Einbettungen, die Semantik widerspiegeln. Ferner interpretieren wir diese Einbettungen algorithmisch und zeigen menschliche Akzeptanz.

Im Hinblick auf **◆ USER MANAGEMENT** geben wir detaillierte Einblicke in verteilte Moderationsprozesse und modellieren die Bedrohung durch missbräuchliche Inhalte. Um plattformweit ein gesundes Klima zu erhalten, erörtern wir weiter die Nutzer-Lebensdauer und modellieren diese anhand von Metadaten. Wir schließen mit einem Entwurf für daten-gestützte QoE-Langzeitanalysen in kontrollierten Massively Multiplayer Online Game Umgebungen (MMOG).

ABSTRACT

Nowadays day-to-day digital communication and social life has only fortified with the ongoing pandemic. People enjoy communication and information across various (direct) messaging platforms and accepted its ever-increasing impact on public discourse and society. While traditional platforms implement user profiles enabling social credit, the landscape also includes anonymity. Yet, a new type of application combining anonymity with a strong spatial focus, hyperlocality, emerged over recent years. To this point, platform implications of both uniquely combined design properties largely remain unknown.

In this thesis, we provide a first data-driven holistic view on Jodel that combines both properties. We leverage unbiased complete ground truth information to dissect a plethora of communities across two different countries: Germany and the Kingdom of Saudi Arabia. This work follows a platform perspective identifying four major essentially important areas revolving around the individual.

That is, we begin with a broad analysis of three **◆ USER ADOPTION** processes along three different applications. After discussing our measurements of the user base adoption of the German COVID-19 digital contact tracing application, we provide evidence of well-established platforms being re-purposed as a side channel to evade censorship in the ongoing Russo-Ukrainian hybrid war. We further showcase that the very same platform ingredients may yield vastly different outcomes on the messaging app Jodel.

While any online platform builds upon **◆ USER INTERACTIONS**, we structurally characterize Jodel behavior across both countries. We discuss structural disparities and detail platform implications - solely induced by local user behavior. An in-depth look into the Saudi community landscape closes a research gap to platform usage in a different society, identifying differences.

Further, we discuss **◆ USER CONTENT** analyzing information diffusion. Taking content to the next level, we developed a multidimensional classification scheme for intents (why) and topics (what) of social media messages and provide details of a crowdsourced campaign for Saudi Arabian contents. With neural word embeddings as a tool for making text tangible and the prevalence of emoji in social media communication, we discuss quantitative and qualitative insights to word-emoji embeddings reflecting semantics. Additionally, we make such embeddings interpretable and provide evidence that our method is well in line with human judgement.

In terms of **◆ USER MANAGEMENT**, we detail insights to distributed moderation processes and model the threat of abusive content. In the long term, platforms need to establish a sustainable, preferably growing, environment. That is, we next discuss user lifetime and possibly early churn factors, while modeling user lifetime from metadata. We finish with a blueprint of data-driven long-term quality of experience analyzes in a controlled massively multiplayer online game (MMOG) environment.

CONTENT



| | |
|---|-----|
| 1 INTRODUCTION | 5 |
| A Related Work | 8 |
| B Research Questions and Contributions | 11 |
| C Limitations | 13 |
| 2 JODEL | 15 |
| A Application and Interactions | 17 |
| B Platform Key Design Properties | 18 |
| C Steering Communities | 19 |
| D Ground Truth Dataset and Corpus | 23 |
| 3 BACKGROUND | 27 |
| A Crowdsourcing | 29 |
| B Machine Learning | 30 |
| C Natural Language Processing | 35 |
| D Social Network Analysis | 37 |
| 4 USER ADOPTION | 43 |
| A Corona-Warn-App Deployment and Engagement | 47 |
| B Platforms as a Sidechannel in Wartimes | 53 |
| C Growth Patterns in Social Media Platforms | 65 |
| 5 USER INTERACTIONS | 77 |
| A Structural Community Analysis | 81 |
| B Cross-Country Differences and Structural Implications | 89 |
| C Platform and User Centric Analysis - Spotighting SA | 97 |
| 6 USER CONTENT | 111 |
| A Information Spreading along Hashtags | 116 |
| B Anonymous Messaging Contents | 130 |
| C The Role of Emoji | 145 |
| 7 USER MANAGEMENT | 179 |
| A Distributed Content Moderation | 183 |
| B User Lifetime Insights and Modeling | 199 |
| C Excursus: Data-Driven Long Term Gaming and QoE | 216 |
| 8 FUTURE WORK | 227 |
| A Community Graph Perspective and Content Diffusion | 227 |
| B Cross-Platform Perspective | 228 |
| C Content Moderation and Radicalization | 228 |
| D User Perspective | 228 |
| 9 CONCLUSIONS | 229 |
| REFERENCES | 231 |

INVOLVED PUBLICATIONS

Parts of this thesis base upon the following (peer-reviewed) manuscripts that have already been published or are under current submission. All my collaborators are among my co-authors, if not mentioned otherwise. Paper execution conceptionally has been led by me.

I would like to thank:

- Timon Mohaupt, for implementing parts of the hashtag analysis (♦ [A: Information Spreading along Hashtags](#)), and executing our crowdsourced campaigns on the POLAR framework (♦ [C.3: Interpreting Emoji](#)).
- Jan Garcia, for creating and laying ground of a machine learning pipeline (♦ [B.4: Modeling User Lifetime](#)).
- Max Bergmann, for adjustig, verifying, and using Jan’s pipeline producing results for our user lifetime analysis (♦ [B.4: Modeling User Lifetime](#)); further thanks for staying with us and supporting us with your AraBERT efforts (♦ [B: Anonymous Messaging Contents](#)).
- Ahmed Soliman, for staying with us, improving our content classification system (♦ [B: Anonymous Messaging Contents](#)).
- Leon Wolter, for implementing, and executing time-consuming parts of our analyses about modeling abusive content on the Jodel platform (♦ [A.4: Modeling Disliked and Abusive Content](#)).

Journals, Conferences and Workshops

- [JH1] J. M. Moreno, S. Pastrana, J. H. Reelfs, P. Vallina, A. Panchenko, G. Smaragdakis, O. Hohlfeld, N. Vallina-Rodriguez, and J. Tapiador. Let me inform you! bypassing russian censorship in war time. In *arXiv*, 2023. *Technical Report*.
- [JH2] J. H. Reelfs, M. Bergmann, O. Hohlfeld, and N. Henckell. Understanding & Predicting User Lifetime with Machine Learning in an Anonymous Location-Based Social Network. In *Companion Proceedings of The 2021 World Wide Web Conference*, International Workshop on Location and the Web, page 324–331, New York, NY, USA, 2021. Association for Computing Machinery.
- [JH3] J. H. Reelfs and O. Hohlfeld. Data-driven study of long-term gaming experience. In *International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2022.
- [JH4] J. H. Reelfs, O. Hohlfeld, and N. Henckell. Anonymous Hyperlocal Communities: What do they talk about? In *Companion Proceedings of The 2022 World Wide Web Conference*, International Workshop on Location and the Web, New York, NY, USA, 2022. Association for Computing Machinery.

- [JH5] J. H. Reelfs, O. Hohlfeld, and N. Henckell. Geographic Differences in Social Media Interactions Exist Between Western and Middle-East Countries. In *Passive and Active Measurement Conference*. Springer, 2022.
- [JH6] J. H. Reelfs, O. Hohlfeld, and I. Poese. Corona-Warn-App: Tracing the start of the official COVID-19 Exposure Notification App for Germany. In *Proceedings of the SIGCOMM'20 Poster and Demo Sessions*, pages 24–26. 2020.
- [JH7] J. H. Reelfs, O. Hohlfeld, M. Strohmaier, and N. Henckell. Word-emoji embeddings from large scale messaging data reflect real-world semantic associations of expressive icons. In *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media*, International Workshop on Emoji Understanding and Applications in Social Media, 2020.
- [JH8] J. H. Reelfs, O. Hohlfeld, M. Strohmaier, and N. Henckell. Characterizing the country-wide adoption and evolution of the Jodel messaging app in Saudi Arabia. *The Journal of Web Science*, pages 1–14, 2022.
- [JH9] J. H. Reelfs, T. Mohaupt, O. Hohlfeld, and N. Henckell. Hashtag Usage in a Geographically-Local Microblogging App. In *Companion Proceedings of The 2019 World Wide Web Conference*, International Workshop on Location and the Web, page 919–927, New York, NY, USA, 2019. Association for Computing Machinery.
- [JH10] J. H. Reelfs, T. Mohaupt, S. Sikdar, M. Strohmaier, and O. Hohlfeld. Interpreting emoji with emoji. In *Proceedings of the The Fifth International Workshop on Emoji Understanding and Applications in Social Media*, pages 1–10, Seattle, Washington, USA, July 2022. Association for Computational Linguistics.

Software and Datasets

- [SD1] J. H. Reelfs. Dataset: Jodel emoji embedding dataset (jeed1488), July 2022. <https://doi.org/10.5281/zenodo.6885406>.
- [SD2] J. H. Reelfs and O. Hohlfeld. Dataset: Data-Driven Study of Long-Term Gaming Experience (Tribalwars), July 2022. <https://doi.org/10.5281/zenodo.6862106>.



INTRODUCTION

Social Media Today

Over the past decades, an abundant amount of online and social platforms emerged. They silently have become a significant accepted part of our day-to-day communication and digital social life, which has only strengthened with recent pandemic years [Feldmann et al., 2021]. That is, people are enjoying communication over a wide variety of (direct) messaging platforms, social networks, or other means of communities, e.g., about parenting, programming, or games, or focussing specific media. Though, social network and platform theory is well understood [Kraut et al., 2012], applied research includes classic online social networks [Mislove et al., 2007, Nazir et al., 2008, Schiöberg et al., 2012, Kairam et al., 2012] as well as more specialized variants such as microblogging [Bollen et al., 2011], picture sharing [Vaterlaus et al., 2016, Cha et al., 2009], or knowledge sharing [Wang et al., 2013]. Besides social networks between friends and acquaintances for sharing personal information or pictures, news aggregators, discussion boards, various focused Q&A websites, or specific media-affine platforms, like pictures or videos only, emerged. Since the uprise of online platforms, research has continually tracked platform evolution and refined methods deepening our today's knowledge of such platforms. E.g., downstream information diffusion [Dow et al., 2013, Cannarella and Spechler, 2014, Yan et al., 2013, Woo and Chen, 2016, Matsubara et al., 2017, Kamath et al., 2013]. It is to be noted that most platforms implement non-anonymity being globally accessible; thus research as well.

Still, specific interest has been structural, emerging social ties allow for modeling networks as graphs, or infer structural implications on information flow. Among many others, e.g., identifying misinformation, filter bubbles, and fighting toxicity has become a major challenge that is usually addressed with hierarchical distributed moderation mechanisms. As such, online platform have an ever-increasing impact on public discourse, and thus society.

From a technical perspective, a platform's first major design decision relies on the level of privacy a platform implements. Most global popular platforms may (implicitly) force users to provide their real name, known identity, or some kind of fixed user account, increasing credibility via social credit [Yu et al., 2015]. However, availability in platforms for entertainment or other communication purposes demonstrate that there also exists the desire and need for anonymity. While Online Social Networks and non-anonymous platforms have been subject to extensive research, only comparably few popular platforms provide anonymity. Nevertheless, research has shown that implementing (semi-)anonymity is

a double-edged sword. Anonymity provides serious positive purposes: removing the link to a user provides disinhibition from any social or legal consequences as shown in e.g., confession boards [Birnholtz et al., 2015]. Yet, anonymity may also tilt into severe toxicity [Papasavva et al., 2020, Zelenkauskaitė et al., 2021], cyber-bullying [Whittaker and Kowalski, 2015], and e.g., alt-right filter bubbles. As such, actual message contents are of interest. Specific topics are highly dependent on a given platform, however, there are essentially two available tools for determining specific, or modeling content: unsupervised, or supervised learning. While unsupervised methods apply (parameterizable) algorithmic measures to provide a potentially useful output labelling, supervised methods rely on human in the loop annotations. Though both methods may provide insights to dataset, unsupervised methods usually tend to be less interpretable due to salient hidden spaces; not only within the realm of neural networks, but likewise when applying an LDA. Without explanations, resulting embedding spaces may still be used to distinguish between entities. For providing direct a better understanding of contents, human annotations are usually gathered for creating a dataset at sufficient quality to apply other representation learning algorithms for a scale-out. While text embeddings find large adaption in Natural Language processing, nowadays day-to-day casual communication on social media consists of more than word. Emoji often replace objects or words, and provide sentiment, i.e., they can transport salient cues in writer’s (self-)identification or interpretation [Barbieri and Camacho-Collados, 2018, Robertson et al., 2018, Robertson et al., 2021b], as proved in e.g., sentiment analysis [Berengueres and Castro, 2017]. However, suitability of social media word embeddings including emoji remain unknown.

This ultimately boils down to one of the greatest challenges for any online platform nowadays, managing and steering user bases and communities. Often applied and scaling reactive distributed moderation schemes have been investigated empirically in e.g., [Trujillo and Cresci, 2022, Lampe et al., 2014], or modeled in e.g., [Stoddard, 2015], while also raising concerns[Lampe and Resnick, 2004, Zhu et al., 2021, Gilbert, 2013]. Actively steering users towards desired behavior e.g., through gamification, or badges [Anderson et al., 2013], and contents becomes an incredibly hard task.

The second major design decision relies on a platform’s implemented spatial scope. Most popular online platforms do not implement any restrictions in content accessibility—apart from possibly local legal requirements. Albeit providing a global scope, research has shown that user activity highly correlates to geospatial distance, forming network clusters [Ugander et al., 2011, Schiöberg et al., 2012, Magno et al., 2012]. Such spatially local clusters naturally reinforce on e.g., location-based social networks (LSBN), which focus on spatially-linked contents, yet remaining globally accessible [Cho et al., 2011, Silva et al., 2019]. However, some platforms implement hyperlocality. That is they spatially link content and also restrict access by current user distance, only enabling communication in proximity of fixed geolocated entities. Research has shown that hyperlocality induces a common ground creating community identities [Guta and Karolak, 2015], and likewise developing own variations in language [Robertson et al., 2020, Hovy and Purschke, 2018], bias [Ferrer et al., 2021], and self-inflicted governance [Fiesler et al., 2018]. Further, we want to stress that the majority of research often focus Western partitions of social media and only begin to include other specific regions, like Asia, or the Middle-East [Reyae and Ahmed, 2015].

In contrast to most other popular global non- and anonymous platforms that have been researched plentifully, anonymity coupled with hyperlocality renders a new type of social media. Though related work has identified many insights to such networks, and alike platforms, many aspects remain unknown. Jodel, a microblogging platform as displayed with an example post in Figure .2, implements such anonymity and hyperlocality simultaneously. Leveraging ground truth information from the operator, we take a holistic view and complement existing work with extensive data-driven studies. Our dataset eliminates

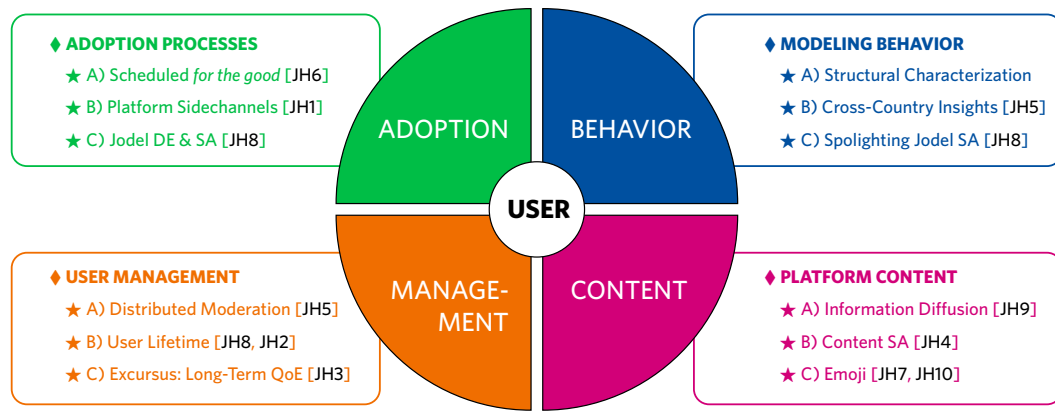


Figure 1: A holistic view: Thesis structure and main research areas. Pressing platform question areas, accompanied by *Jodel*'s overarching main application design properties of 📍 hyperlocality and anonymity 🛡️, challenging social networkness 🌐.

sampling biases and thus adds value puzzle pieces to implications and our understanding of platform design properties at hand. In this holistic view, we define four overarching research questions and subsequent thesis contributions along four major essential key aspects to online platforms as shown in Figure .1, which we will detail next.

◆ **USER ADOPTION** . As for User Adoption within (Social) Online Platforms, we are interested in onboarding processes to new and existing online services. We complement the body of prior research by studying *A*) user onboarding processes of a governmental advertised digital contact tracing app in rich demand of whole societies, *B*) reallocating platform communication to a side-channel at nowadays hybrid warfare, and *C1*) and *C2*) evolution of new social and entertainment platform, *Jodel*, being hyperlocal 📍 by design and thus being spatially limited to distinct communities, comparing two different country's landscape of local communities.

◆ **USER INTERACTIONS** . With *Jodel*, we showcased how even the very same platform ingredients to a messaging platform may lead to much different outcomes in user adoption. Thus, we will now take a deep dive to user interactions, the essential and key driving elements, for any online community platform. While related work on primarily globally available platforms provides rich insights, we complement these perspectives with a data-driven approach comparing not only various independent communities, but also two spatially distinct countries having a different cultural background: Germany and the Kingdom of Saudi Arabia.

◆ **USER CONTENT** . Having discussed adoption patterns and general user interactions on a meta perspective, we next highlight specifics of actual message contents. Based upon and extending metrics from [Kamath et al., 2013], we present and discuss our findings about information diffusion on *Jodel* DE represented by hashtags across the *Jodel* platform in Germany. We next detail actual message contents by creating and conducting crowdsources annotations distinguishing between message intent (why) and topic (what) for *Jodel* SA. Having focused on contents and words, we close the chapter of content with a detailed analysis of emoji usage within today's social media.

◆ **USER MANAGEMENT** . As interesting content may get, moderation becomes an essential survival task of any platform. We provide a characterization of *Jodel*'s distributed moderation system and model of abusive contents. Apart from this acute question, platforms are constantly threatened by *user churn*. Thus, customer lifetime value (CLV) anal-



Figure .2: Jodel example. The Jodel messaging app supports posting short messages containing text and emojis in a user's geographical proximity.

ysis and prediction becomes a key enabler for understanding churn—and reasons thereof. Leaving us with satisfying, but inconclusive results w.r.t. reasons, we propose a new data-driven blueprint for user quality of experience. Leveraging temporally fixed timeframes, i.e., game rounds of an MMOG, we measure gaming dynamics and showcase how data-driven insights can provide strong hypothesis candidates for further classical QoE evaluation—and thus, as a key enabler, help to understand user churn better in the future.

A Related Work

Jodel represents a rare example within the space of online (social) platforms due to its combination of hyperlocality and anonymity. For Jodel or the much alike Yik Yak application, mostly qualitative, but also some empirical research exist. However, specifically the data-driven approaches lack ground-truth information and are scoped by time. Within this thesis, we complement such works with a longitudinal ground truth dataset from the Jodel operator of two countries spanning years, that not only enables in-country community, but also cross-country comparisons, broadening our insights.

We further set out to explore prime examples of very similar platforms sharing (partly) Jodel's unique property combinations of hyperlocality and anonymity. That is, we detail research about the Jodel alike Yik Yak platform, and Whisper which focuses on images. Lastly, we also cover anonymity in a broader sense and provide a research overview of popular platform.

Please note that we provide an overview to general Social Network and Online Platform Analysis in the next chapter [◆ D: Social Network Analysis](#).

A.1 Jodel

We will set out and provide an extensive overview of existing analyses, surveys, and research findings about Jodel itself.

A popular method to gain first insights given limited information are surveys. That is, [Jüttner et al., 2021] surveys 874 participants for reasons to use the Jodel platform. They subdivide the participants into user roles of producers, commenters, raters, and consumers; identifying that "boredom and entertainment are among the most common usage reasons". Similarly, [Kasakowskij et al., 2018] further elaborate on driving factors in participation comparing non-/anonymous usage with the example of Jodel and Instagram. They survey 664 participants, identifying that Jodel as an anonymous platform is more suitable for socializing than creating social credit. With a more specific focus on content quality as an information service, [Nowak et al., 2018] ask 1009 participants about their user experience. They report that 72% of the respondents are students. Furthermore, 92% of the respondents are satisfied with the app while 97% said that they would recommend Jodel. They complement their analysis with an interview with the Jodel founder Alessio Borgmeyer. Though being anonymous, [Zaki, 2018] qualitatively finds evidence in four detailed interviews, that self-identification can emerge from local Jodel communities, that also affects language policies as shown in [Heuman, 2020a]. While presented surveys focused general aspects of

usage, [Elson et al., 2020] investigate social media usage implications for a US Military Academy, highlighting that Jodel is among the top 5 used platforms. They conclude that it potentially has positive and negative effects due to anonymity, embracing also controversial topics. Further, [Seidenschur, 2021] and [Laaksonen and Rantasila, 2021] provide a very recent qualitative analysis on COVID-19 related communication on the platform in Germany and Helsinki, Finland. We complement this body of survey based research with ground truth based empirical classification of the entire user base across two countries.

As mentioned earlier, data-driven studies are hard due to lack of data or the necessary step to sample contents introducing inherent biases and uncertainty. Nonetheless, [Hovy and Purschke, 2018, Purschke and Hovy, 2019] crawl about 2.3M threads and 16.8M posts from 408 German-speaking countries (DACH). They leverage this huge amount of data to show local language and dialect variations correlated to geography. The authors of [Bergau, 2021] use sample of 1.8k posts to create a machine learning algorithm that detects humor. By applying manual feature engineering, they show Support Vector Machines (SVMs) performing quite ok for this task. We complement this work with a data-driven complete Jodel characterization along four essential building blocks from a platform perspective: User Adoption, Interactions, Content, and Management. This way, this thesis provides a very first large scale exploration of the Jodel platform.

A.2 Yik Yak

The Yik Yak application in its functionality and features is almost similar to Jodel; being anonymous and hyperlocal. While the operators reportedly had issues with abusive content, they later changed the app and ultimately shut down [Mahler, 2015]—highlighting the importance of suitable content moderation. App design decisions are discussed e.g., in [Li and Literat, 2017]. In [Williams and Mahmoud, 2018], the authors model and further elaborate on user concerns. Nonetheless, Yik Yak has attracted various researchers to investigate the platform.

That is, many empirical works study posted contents and compare them to e.g., Twitter. Within [McKenzie et al., 2015], leveraging about 665k posts, the authors show that spatial topical clusters emerge from the hyperlocal design, while they significantly differ from Twitter according to an unsupervised Latent Dirichlet allocation (LDA). Likewise, [Saveski et al., 2016] show temporal daily patterns, while also applying an LDA classification across 2.9M posts. In comparison to Twitter, they find more vulgarity (probably due to anonymity). Concentrating on rather controversial contents, [Lee et al., 2017] show that their collected data mostly revolves around dating, sex, friendship, expressing emotion, or personal experience. These findings are in line with [Black et al., 2016] and [Wu et al., 2017], who also highlight the application’s focus on local campus life within US colleges. Further, not being used for political discourse [Vargo and Hopp, 2018], Yik Yak may lead to cyberbullying, while focussing on entertaining contents as found in a rather qualitative study [Clark-Gordon et al., 2017].

Besides empirical data-driven approaches, many researchers used surveys to better understand driving factors behind Yik Yak. By asking 264 participants, [Vaterlaus, 2017] determine that most of the platform users are younger than non-users. They use Yik Yak for informal, entertainment, and socializing purposes. This is confirmed with a qualitative survey with 18 participants in [Frith and Saker, 2017] showcasing that the hyperlocality and local socializing are main driving factors for using the platform. The focus on local communities further provides common grounds for its users as already seen for Jodel, while proactively using e.g., language, to set oneself apart from others (other locations) as shown in [Heston and Birnholtz, 2016b].

While many surveys highlight possible negative effects of anonymity w.r.t. vulgarity, cyberbullying, or adverse content, this property also provides positive effects. A qualitative

survey of 744 posts [West, 2017] reveals that platform environment does not create a particularly bad atmosphere, but may be used as a serious information source. Similarly, [Bayne et al., 2019] elaborate on positive effects and significant value stemming from anonymity. Another example of leveraging this anonymous platform for the good has been extensively displayed within various analyses of platform contents regarding a campus' local libraries in [Robison and Connell, 2017, Lewter and Profit, 2018, Price, 2018].

Nonetheless, platform design decisions with possibly privacy-breaking information leakage remains a major concern [Xue et al., 2016].

A.3 Whisper

The Whisper¹ application also implements anonymity and hyperlocality within their application, while also providing a global view. In contrast to Jodel or Yik Yak, it focuses on sharing image memes (an image with additional text). A first empirical evaluation of contents, interactions, and regional clusters has been published by Wang et al. in 2014 [Wang et al., 2014]. Additionally, [Mondal et al., 2020] presents another empirical evaluation about platform-assigned topics and language across about 90M posts. Information diffusion through the platform has been characterized in [Cao et al., 2012]. General user behavior types, e.g., inactive, or hostile users, may be identified by leveraging actions sequences from clickstreams as presented in [Wang et al., 2016]; they create hierarchical clusters according to user behavior from a clickstream graph representation.

A specific focus on anonymity effects is provided by [Correa et al., 2015], who compare Twitter with Whisper w.r.t. a crowdsourced *anonymity-score* employing Amazon Mechanical Turk. Somewhat expectedly, they identify that certain sensitive topics, such as NSFW, become more frequent on anonymous platforms.

Lastly, we would like to highlight the work by Kang et al. [Kang et al., 2016], who qualitatively survey 18 participants for reasons to use hyperlocal anonymous social platforms, specifically such as Whisper or Yik Yak. Answers include simply entertainment purposes sharing stories or jokes, or deep confessions within the veil of anonymity. Generally, the participants consider anonymous platforms more honest, open, and diverse.

A.4 Anonymity

While we have focused on platforms being quite similar to Jodel, we want to elaborate more on anonymity and privacy within online platforms.

According to a data-driven study by [Stutzman et al., 2013] analyzing about 5k Facebook profiles from 2005 to 2011, users have become more sensitive to privacy. That is, within the early Social Network stage, many users were quite open and willing to openly share their information and posts with anybody. However, over time, they observe a decreasing inclination for sharing data with anybody. That is, users began to restrict profile information, pictures and posts to their social ties.

While the sensitivity to privacy increased, [Leavitt, 2015] identify on Reddit that some users are likely to participating within online platforms in anonymity using throwaway accounts. Further investigation signals that e.g., women are more likely to do so. The urge to participate in anonymity also becomes apparent within semi-anonymous confession boards as researched e.g., in [Birnholtz et al., 2015]. Such confession boards enable sending messages to moderators who then posts them into the public. The authors find that typical topics are taboo or stigmatizing. Within the disguise of this anonymity, inquiring users circumvent negative disinhibition by keeping their identity secret.

Though anonymity can play a positive role for the good, it also introduces downsides. [Whittaker and Kowalski, 2015] report cyberbullying activities among college students on

¹<https://whisper.sh>

social media and identify texting being the top-used method for bullying. They identify specifically targeted attacks being most toxic.

Further, [Bernstein et al., 2011] find that e.g., 4chan, though enabling user accounts, is usually being used in full anonymity by 90% of its post volume. The platform itself appears fast-paced and specifically anonymity leads to stark ephemerality. These findings are confirmed in [Papasavva et al., 2020], who analyze a large corpus over 3.5 years.

Lastly, anonymity removes responsibility as shown for a Yahoo Answers, a Question & Answer platform, as shown in [Kayes et al., 2015]. Nonetheless, the authors identify features and create a classification method identifying abusive users quite well. With Gab, [Fair and Wesslen, 2019] provide a dataset that contains lots of hate comments.

B Research Questions and Contributions

Next, we highlight the specific research questions and major contributions colored along the chapter structure according to the shown in Figure .1.

B.1 ◆ USER ADOPTION

General user adoption processes are inherently hard to observe due to the necessity to include desirably long observation periods. Thus, online platform research relying on temporal developments are often limited in timeframe and by sampling. Only knowing and actively probing the target enables data-driven analyzes; that is, research hardly captures specifically the early emergence of any subject.

To close this gap, we showcase three different platform adoptions processes across three platforms, showcasing their very birth, or specific platform changes w.r.t. user adoption.

That is, we first set out and detail the early adoption of the ★ A) Corona-Warn-App, Scheduled for the good [JH6] . In face of the COVID-19 pandemic and therefore considerable public anticipation, on June 16, 2020, the digital contact tracing app for Germany has finally been released, enabled by the notification API physically relying on Bluetooth beacons. We measure downloads and its popularity throughout the very first days after deployment leveraging netflow traces towards the hosting infrastructure. Thus, we provide insights into daily usage patterns and volume, while also dissecting interest on a spatial dimension across Germany.

Next, we shift our focus to sudden changes in well-established platform usage: With escalated and ongoing Russo-Ukrainian hybrid warfare, digital places have also become battlefields. In this context, specifically Online Social Networks and news platforms have become victim to Russian censorship. Thus, people have become creative using non-blocked ★ B) Platform Sidechannels [JH1] , i.e., re-purposing the communication via a legit platform to inform the Russian population about the escalated ongoing lethal warfare happening in Ukraine. With our contribution, we find evidence of Google Maps and TripAdvisor being used as side channels—and measure reactions from operators, e.g., actively deleting such entities. Based on crowdsourced expert classifications, we further measure and determine amounts of war-related contents on each platform and discuss related metrics.

Having introduced two quite opposite user adoption examples, we now dive into the world of Jodel and its emerging plethora of hyperlocal communities within the veil of anonymity. While empirical works from the very first post barely exist, we highlight the growth and adoption process of Jodel in Germany and the Kingdom of Saudi Arabia ★ C) Jodel DE & SA [JH8] . While Jodel in Germany follows a continuous *organic* growth pattern, much unlike this reference, the KSA user base established all of a sudden. We investigate platform implications and model empirical findings, while providing a strong hypothesis of the circumstances. While in-country communities appear quite similar, the cross-country evolution of the user base and app adoption differs significantly.

B.2 ◆ USER INTERACTIONS

Having showcased that even the very same platform ingredients to a messaging platform may lead to much different outcomes in user adoption, We next focus user interactions, the essential and key driving elements, for any online community platform—discussing structural and empirical meta perspectives.

We explore structural insights to the Jodel platform across both countries in ★ A) [Structural Characterization](#) , incorporating a first analysis dimension distinguishing between communities by size via interaction *volume*. We discuss various structural community insights, platform content and interactions, and user behavior. While the platform allow for postings photos, most users focus on textual messages. Surprisingly, we identify contrastive day/night cycles in application usage between DE and SA. User interactions generally appear to be heavy-tailed.

Further, we provide a rough picture that the anonymous settings does *not* particularly lead to interacting user clusters ❄️. Driven by identified differences in the empirical characterization, we engage those findings in depth via a dedicated ★ B) [Cross-Country Insights \[JH5\]](#) . After highlighting certain significant disparities in interaction distributions between DE and SA adding a *temporal* dimension to our analysis, we elaborate on cross-country disparities in posting and voting volumes and structural implications, identifying how the very same platform properties yield very different behavioral patterns. The Saudi user base prefers actively engaging into discussions, while the German user base more like participates in the voting mechanism.

As an underrepresented subject in literature and lack of insights to such a rich dataset, we empirically characterize and model Jodel’s Saudi user base and communities in ★ C) [Spotlighting Jodel SA \[JH8\]](#) . By dissecting the communities by size *rank*—in favor of a qualitative insight—and *time*, we provide rich insights to many similar, or qualitative distinct distributions found within landscape of spatially distinct community. Other metrics appear to follow scaling effects.

B.3 ◆ USER CONTENT

After focusing on adoption processes and investigate the underlying community structure through user interactions, we next move forward into the realm of message contents. While most analyzed platforms are globally accessible, information spreading characteristics induced by Hyperlocality 📍 as found in Jodel remain unknown.

That is, we explore ★ A) [Information Diffusion \[JH9\]](#) represented by hashtags across the Jodel platform in Germany. We describe information diffusion through various metrics, and provide an ontology of hashtag types distinguished by a temporal and spatial dimension. We showcase examples, and showcase a classification approach with machine learning. Smaller cities around the heavy user bases tend to be highly influenced.

While we have shown information and abstract idea spreading through local communities, we still have not revealed actual message contents, yet. Given the lack of related work, sudden adoption process in Saudi Arabia and political circumstances, we decided to choose Jodel for a crowdsourced ★ B) [Content SA \[JH4\]](#) . We developed and apply a rich classification scheme answering te *intents* (why) and *topics* (what) of a message. We discuss empirical findings of prevalent communication—what do the users talk about, and why they might do so. While we find little evidence of toxicity, people enjoy sharing personal stories and beliefs, entertainment, and (local) information. Further, we also provide evidence of taboo topics benefiting from anonymity.

Having focused on contents, we reason that ★ C) [Emoji \[JH7, JH10\]](#) has acquired a driver’s seat within nowadays casual day-to-day communication. We empirically characterize emoji usage patterns throughout the DE & SA communities insights to emoji usage

within the German and Saudi user bases in [★ C.1\) Social Media Emoji Usage](#). To make emoji understanding machine-feasible, we leverage Word-Emoji embeddings [★ C\) Emoji \[JH7, JH10\]](#), encoding semantic associations through word co-occurrence. We identify countless reasonable associations from emoji to emoji, emoji to text and vice versa; while uncertainties arises for words having multiple meanings, or may have created a specific platform understanding. Building upon this success, we use semantic differentials making our embedding interpretable; with crowdsourced human judgement experiments, we conclude that our approach a capable of [★ C.3\) Interpreting Emoji](#) in line with humans. I.e., emoji can improve interpretability of POLAR embeddings, specifically in interpreting emoji themselves.

B.4 [◆ USER MANAGEMENT](#)

We showcased Jodel’s emergence and evolution, discussed user interactions and cultural differences, and put a specific focus on content diffusion, actual messaging topics and intents, as well as emoji as a major element in casual communication.

Lastly, we are interested in mechanisms that keep the platform running. That is, provide insights to Jodel’s [★ A\) Distributed Moderation \[JH5\]](#) system that regularizes user contents through priming, incentives, community driven distributed voting, and a moderation safeguard ontop. We empirically characterize Jodel’s distributed moderation architecture across the German and Saudi communities. Our modeling efforts of blocked content with a state-of-the-art BERT-alike Masked Language Model for Jodel Germany, we conclude that computational detection of community-specific unwanted content remains a hard task.

Though our thorough perspective on the moderation architecture reveals internals to content regularization, another major task of user management can be seen in [★ B\) User Lifetime \[JH8, JH2\]](#). While adverse content is a tangible measure, creating self-sustained communities require a consistent influx of new users. Due to lack of research of churn factors on Jodel or alike platforms, we fill this gap by empirically characterizing user lifetime and churn within the Jodel DE & SA communities. By successfully modeling lifetime with domain specific engineered features. We showcase and discuss how the model can be used to infer similarities between communities, and how feature importance measure might indicate important churn factors, learning from the model.

As observed, our peek into leveraging empirical findings from e.g., the model importance might turn out to be hard task, we note that Jodel might generally appear quite ephemeral and thus noisy, which makes investigation of experience influence factors even harder. To lay a foundation in overcoming this issue, we showcase a blueprint analysis of an MMOG implementing long-term round structures spanning years to measure [★ C\) Excursus: Long-Term QoE \[JH3\]](#). Empowered by data-driven (pre-)studies, we uncover (unsurprising) correlation between user lifetime and success factors. Nonetheless, we consider our work as a key enabler boosting the hypothesizing process that may require downstream causality analyzes—making the study of long experience timeframes tangible—for a better understanding of user experience, including reasons for churn.

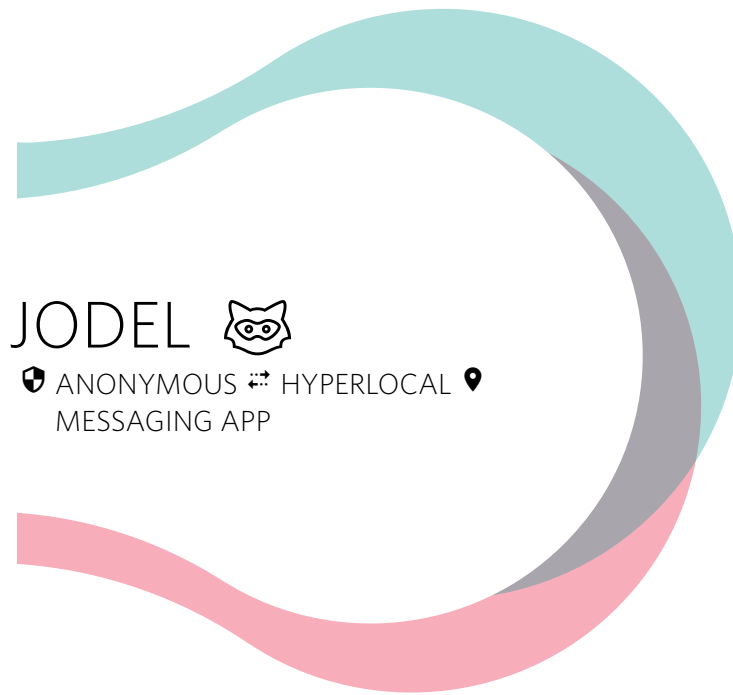
C [Limitations](#)

In this work, with Jodel, we provide empirical evidence that systematic differences in Social Media usage can emerge, given the very same platform ingredients (design features). We identify these differences between unequal countries, whereas the distinct communities within each country generally appear quite homogeneous. Our subject of investigation is a Western country, Germany, and in comparison a Middle Eastern country, the Kingdom of Saudi Arabia. We are fully aware and want to remark that our two country sampling points do not generalize towards other platforms or countries, yet the plethora of local

forming communities due to app design in combination with ground truth data from the operator allows for a complete picture for the two countries under investigation. Nonetheless, specifically related work on similar platforms, and thorough qualitative insights and interviews conducted throughout this research endeavor suggest that many of our findings in fact likely generalize, or at least lets us learn that platforms might be used very differently depending on context, which needs special attention. Parts of this work analyze other platforms, which also represent only a limited set of sampling points—limitations likewise apply. Yet, a broader scope certainly represents a very interesting lever for future work.

Structure

We start with this chapter introducing the overall topic, a general introduction to related work, and the major research questions and contributions of this thesis. Afterwards, we provide insights to the Jodel applications, and background knowledge to methods and tools used within this thesis, while also providing an overview into social network research. After addressing our research questions in a dedicated chapter each, we wrap up with interesting future research directions and open questions, as well as a final conclusion.



In a nutshell. We next briefly describe the application studied in this thesis. The Jodel messaging app is mobile only and enables anonymous and hyperlocal communication. That is, there are no user profiles, while specific users are only enumerated within a conversational thread. Further, hyperlocality links threads to a geolocation. Only in case of spatial proximity, users may communicate within such threads.

CONTENT

| | | |
|----------|---|-----------|
| A | Application and Interactions | 17 |
| B | Platform Key Design Properties | 18 |
| B.1 | Realnames, Pseudonymity and Anonymity  | 18 |
| B.1.1 | To Be, or Not to Be?!  | 19 |
| B.2 | Location-Basedness and Hyperlocality  | 19 |
| C | Steering Communities | 19 |
| C.1 | Social Priming | 20 |
| C.2 | Content Presentation | 20 |
| C.3 | Gamification | 21 |
| C.4 | Enabling Topical Subcommunities - Channels | 21 |
| C.5 | Safeguarding the Communities | 22 |
| D | Ground Truth Dataset and Corpus | 23 |
| D.1 | Legal and Ethics | 23 |
| D.2 | Dataset Description | 24 |

Jodel - Anonymous Hyperlocal Messaging

Jodel¹ is a mobile-only messaging application which we show in Fig. A.1. It is location-based and establishes local communities relative to the users' location. Within these communities, users can *anonymously* post both images² and limited textual content (i.e., microblogging) and reply to posts forming discussion threads. Posted content is referred to as *Jodels*. They are only displayed to other users within close (possibly dynamic, usually up to 20-30km)³ geographic proximity. Further, all communication is *anonymous* to other users since no user handles or other user-related information are displayed.

JODEL is a place where *everyone has a voice*. Our platform is aimed at encouraging people to interact with each other *locally* in *meaningful* ways. It doesn't matter who you are or where you come from, **what matters is what you have to say!** This is a place where you'll find lots of new stuff to love as you get to know your community. We aspire for our communities to be *helpful and friendly* so that **everybody** here can have a *good time* with **#GoodVibesOnly!**
We count on **your** help and positive engagements.

JODEL, *Support/FAQ*
<https://support.jodel.com/hc/en-us/articles/360009286874-Jodel-Values>

Platform Content The design of the application provide individuals with an anonymous community which may have considerable impacts. First, due to anonymity, there are no social ties, which set the focus solely on content that would hopefully be appreciated by others through some kind of platform feedback mechanism. More importantly, anonymity removes constraints in topical variety, i.e., people may open up and share personal experiences, seek or provide for help; while also allowing for controversial discussions. Though anonymity is often believed to promote toxic content, the proven sustainability of Jodel until today shows that smart steering of communities can create helpful, friendly, funny and at least entertaining environments.

Besides enjoying a reportedly rather homogeneous user groups among pupils, students, and rather younger professionals; as anonymous as the platform appears, as much common ground is imposed on each user due to Jodels hyperlocality. All interactions on the platform usually happen in geographically close proximity.

From personal experience being Jodel moderator over years across the world, common posted and discussed topics among others are jokes, memes, (cat) pictures, local questions, school/university related content, love/sex and relationships.

Worldwide Usage Jodel as a technical system does not restrict its usage. However, attracting sufficient users in a certain region to kick-start a new community is a chicken-and-egg situation. Though also being used e.g., in the United States, Jodel has established vivid communities within Europe and the MENA countries. While naturally expanding and purposely engaging certain regions, Jodel specifically mentions activity for Germany, Austria, Switzerland, France, Iceland, Norway, Sweden, Finland, and Saudi Arabia, Bahrain,

¹Jodel, German for yodeling, a form of singing or calling.

²The ability to post videos and subscribe to channels was added after the end of our dataset.

³Within our observation period, this value had been about 10-30km.

and Qatar⁴.

① This manuscript bases upon datasets from **Germany** and **Saudi Arabia** only (D: [Ground Truth Dataset and Corpus](#)), limiting our focus to these two countries; while Germany is the biggest and oldest market, the Saudi user base is the second largest within our available timeframe from the very first posts in summer 2014 accumulating to billions until autumn 2017.











Jodel's Motivation and History Reportedly⁵ the motivation for Jodel has been Facebook's (at least subjective) decline in users sharing information and shifts to portraying the best possible image of oneself. This combination has led to creating an anonymous platform allowing for anyone sharing information again—without possibly incurring negative consequences. However, due to this implementation only considering friends within users contacts, the predecessor application has likewise suffered from little activity and content.








The launch of Jodel in 2014 then kept anonymity, but also introduced the concept of hyperlocality that connects users in geographically close proximity.

Monetization Jodel may be called a successful startup as of today. Having matured, Jodel runs considerable infrastructure and currently employs about 60 people. Alike many startups, they first focused developing and establishing their product; as any platform however, Jodel does not run for free. That is, over recent years, one can observe Jodel's declared efforts to become financially self-sustainable⁶ in a classical two-way approach: 1) introducing advertisements within the application feeds, and 2) more recently adding paid premium features for users⁷.

A Application and Interactions

Next, I focus on the Jodel application interface in Fig. A.1. Note that subsequent sections provide detailed discussions.

Due to establishing local communities, the current users' location  is displayed at . Within these communities, users can post  both images and textual content that is shown as a core element at . Such posts declare a geographical anchor. Additional meta information like the current relative distance  to the post is shown at , or the amount of replies  within a forming discussion thread is displayed at . No further information about posting users is given, rendering all communication *anonymous* . Only *within* a single discussion thread, users are enumerated and represented by an ascending number in their post order, to enable referencing to other users. Posts are displayed to the users in three different feeds  up to a defined amount of threads.

Employing a community-driven filtering and moderation scheme to avoid adverse content. That is every user can increase or decrease a post's vote score by up-  (+1)  or downvoting  (-1) . Posts reaching a cumulative vote score   of (-5) are not displayed anymore. Further, every post can be flagged  as abusive for subsequent moderation.

A (non-public) lightweight gamification approach rewards users with *Karma*  points for their well-behavior as shown at .

⁴<https://about.jodel.com/>

⁵<https://jodel.com/blog/the-story-of-jodel/>

⁶<https://jodel.com/blog/ad-column/>

⁷<https://advertising.jodel.com>

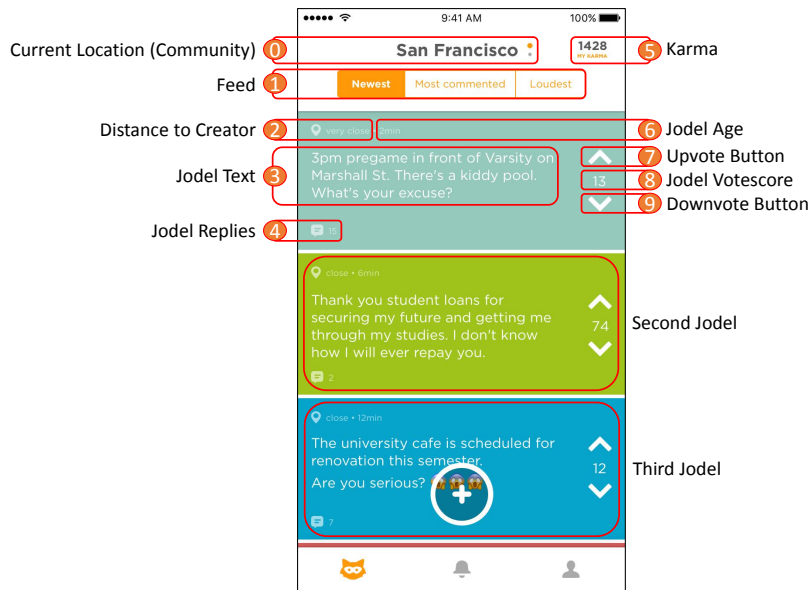


Figure A.1: Jodel iOS mobile application.

B Platform Key Design Properties

First, we want to present specific characteristic design features of the Jodel application that sets it apart from many other platforms. Jodel connects two specific properties: *Anonymity* 🛡️ and *Hyperlocality* 📍, which we describe in detail within subsequent sections.

To provide an overview of possibly well-known platforms, we present a selected set of available property combinations within Table B.1.

- 📄 We provide more information of discussed platform design features and implications within the Related Work section (cf. [A: Related Work](#)), while anonymity and hyperlocality are a recurring part of our investigations.

| | 📍 (Hyper-)Local, Location-Based | 🌐 Global |
|---------------------|---------------------------------|------------------------------------|
| 🛡️ Anonymous | Jodel, YikYak, Whisper | 4chan, 8chan |
| 👤 Pseudonyms | Yelp, | Twitter, Instagram, Snapchat, 9gag |
| 👤 Realname | <i>neighborhood forums/OSN</i> | Facebook, LinkedIn |

Table B.1: Prominent Platform Examples in the Realm of (Non-)Anonymity and (Non-)Hyperlocality. While the space between being anonymous and using real names is fluid, most platforms enforce at least a user account and publicly associate contents and interactions to it. While Facebook enforces real names to new accounts, users on large tech platforms tend to provide their profile with real names albeit not being a necessity. Location-Based OSN relate to specific locations and often revolve around users' local proximity, however we call strictly Location-Based OSN *Hyperlocal*; that is, displayed app content always depends on the current user's location.

B.1 Realnames, Pseudonymity and Anonymity 🛡️

There are various types of non-/anonymity available to platforms. Some (try to) enforce real names, such as Facebook, others are mostly used with real names, such as Twitter, while others usually provide at least some kind of pseudonym/username, such as classic online discussion boards.

Though this decision might appear without stark consequences for a platform, this is particularly not the case. It has been shown that anonymity might lead to toxic environments

as experienced on 4chan, if not properly moderated. More importantly, the implementation of anonymity lacks user profiles, thus not allowing for social ties as they are quite common for any other platform. This leads to rather ephemeral encounters with no sustained social credit, which in turn raises the question of why users participate apart from pure entertainment.

B.1.1 To Be, or Not to Be?! ❄️

As a pressing follow-up to anonymity, another question is to which extent a platform create (Online) Social Networks (OSN).

Social Networks are generally defined as social structures with ties and interactions; OSN describe online platforms sharing the key aspect of users interacting with each other and creating social ties. The analysis of such networks is nothing new [Wasserman et al., 1994], however, OSN in particular allow for large scale analyses of patterns, influences, and various user behavior. We will provide a broad introduction to the research area in chapter [3: BACKGROUND](#).

B.2 Location-Basedness and Hyperlocality 📍

Research has determined certain platforms as Location Based Social Network (LSBN), if it set focus on certain locations as a content-hierarchy, i.e., making specific geolocations main interaction entities. Thereby, it does not restrict the usage according to the user's current location.

In contrast, we call platforms that restrict content visibility to the user's device location *hyperlocal*. As for Jodel, the in-app content presentation adds relative location information within certain discretized categories: *here*, *very close*, *close*, *far*, or *hometown* (cf. [A.1 4](#)). *Here* means the post was shared within about 1km proximity; *very close* about 2km, whereas *close* represents a 10km range and *far* being beyond that radius⁸. In later years, Jodel introduced the *Hometown* feature that allows users to pick one very community as a home community; the app allows for switching to the home community's feed irrespective of the current reported location.

C Steering Communities

For Jodel, it is of utmost importance and most critical to their business keeping the communities growing, expand into new countries or domain domains, and establish as sustainable environment.

Only providing a technical platform without any constraints or safeguards may result in quite toxic environment as observed for 4chan [Papasavva et al., 2020]; Though most users of such platform are aware of its particularities and apparently enjoy discussing (at times controversial) contents, this is arguably not compatible to a broader audience.

After reported trouble [Mahler, 2015] with adverse and abusive content, the Jodel-alike Anonymous Hyperlocal Messagging App YikYak (cf. [A.2: Yik Yak](#)) first tried abandoning anonymity changing their application, while ultimately discontinuing their platform.

Given the very example, the need for steering communities comes very appealing—Jodel sets several corner stones: Applying social and technical measures.

⁸<https://support.jodel.com/hc/en-us/articles/360001040974-Here-Very-close-Close-Far-Hometown>

C.1 Social Priming

To manage user behavior and content, external framing is very important. That is, Jodel demands Open and Humble Clever and Bold Playful and friendly⁹ behavior between its users.

In particular¹⁰:

- **“Positive and friendly:** Jodelers are always positive and nice to each other. Good vibes only!”
- **“Helpful and supportive:** Jodelers help each other out. Do good and may the Karma be with you!”
- **“Colorful and diverse:** Our different colors represent the diversity of people and topics in our community. We celebrate and embrace diversity. Variety is the spice of life.”
- **“Original and creative:** Be your original and unique self, share your own amazing thoughts. We value creativity and new ideas. Just be yourself!”
- **“Respectful and human:** Remember that you are interacting with real people, not just a screen. Treat others the way you want to be treated: friendly and with respect.”
- **“Jodelahuiiiiiiii:** Never ever underestimate the importance of having fun together. Don’t take life too seriously, smile and enjoy the ride.”

Platform Rules Nonetheless, there is need for specific rules that one wants to impose on the communities, setting an absolute lower baseline of acceptable contents. This task of defining and creating such rules is *incredibly* hard, let alone a meaningful categorization.

For Jodel, this is mostly not breaching anonymity of non-public individuals. Others are somewhat common sense in a friendly and respectful environment: no hate and harassment, no fake news, no spamming, no pushy sexual behavior (users are in close proximity), and illegal topics. These rules expand likewise to other parts of the platform like posting Photos/Videos, and Channel Usage.

<https://support.jodel.com/hc/en-us/articles/360000774153--Jodel-Guidelines>

1) The disclosure of personal information (and that of others) 2) Harassment and/or (inciting) violence 3) Discrimination 4) Spamming / Spoilers 5) Negative Vibes and Bad Tone 6) Sexually explicit content 7) Pushy Sexual Behaviour 8) Trolling 9) Fake News 10) Third party applications and content 11) Other and Illegal

C.2 Content Presentation

The application presents three major modes of content presentation ordered differently (note that this is identical for channels) as follows, cf. A.1 ①.

Newest ② The newest feed presents most recent Jodel posts in a community. It resembles a timeliness factor showing and advocating most recent interactions.

Most Discussed ③ Within the most discussed feed, the application displays most discussed topics, i.e., posts having most replies. The selection is additionally windowed by a defined time period to prevent displaying too old contents. It resembles an additional activity factor showing threads actively keeping the community busy providing an additional point to jump into the Jodel universe asking for participation.

⁹<https://about.jodel.com>

¹⁰<https://support.jodel.com/hc/en-us/articles/360001048074-Our-moderation-system->

Loudest 📢: Lastly, the loudest feed present contents most appreciated by the community. It provides a direct list of current hot topics. Likewise to the discussed feed, this selection is windowed by a defined time period.

C.3 Gamification

Karma ↗ To steer user activity into a positive direction, Jodel employs individual Karma scores, cf. A.1 ⑤. They represent a non-public social currency that enforces positivity w.r.t. participation. That is, *positive* interactions get overall rewarded (posting and receiving positive votes; voting, whereas later changes reward upvoting and adding karma costs for downvoting), whereas *negative* interactions incur costs upon Karma: content performing bad w.r.t. a given community, i.e., receiving downvotes, reduces scores. Getting specific contents blocked incurs extra punishment through Karma.

Users accumulating too much negative Karma are banned temporarily or permanently.

Becoming a Moderator ★ In the realm of Jodel, Urban Legends draw a pictures of unlimited power being a Jodel-selected Moderator (cf. C.5 Community Moderators 🗉★), while also complaining about *subjectively unfair* moderation decisions. While it is true that moderators actively influence moderation decisions, they are not local to *their* community breaking with most possible biases and ties.

Also primed by Jodel, unlocking the moderation functionality within the app may be seen as the *Final Boss* or achievement.

C.4 Enabling Topical Subcommunities - Channels

Over time, Jodel realized that a single community works well with fewer participants. However, due to distributed moderation/voting (cf. C.5 Two-Staged Distributed Majority Voting and Moderation) niche topics may have a hard time as they may be filtered out by the community mainstream/majority. Identifying this obstacle, Jodel introduced channels—after the observation timespan of our dataset. They enables users to specifically create or join certain channels named by topic according to local interest, e.g., dedicated to #GNTM (Germany’s Next Topmodel), Fitness, University or School, without disturbing the broader Jodel community in the main feed. These channels are also displayed within the same feed types as described before (most recent, most discussed, and loudest, cf. C.2: Content Presentation).

I.e., channels open up space for more diversity and niche contents; which also triggers establishing new common grounds w.r.t. accepted content/general behavior within such subspaces—framed into imposed platform rules. E.g., though adding images of unambiguous persons or nudity are generally not allowed to preserve user privacy and anonymity and enforce non-nude content. As such, active communities usually immediately vote such contents out, or at least it might be flagged for moderation. However, a German channel called *@bodyselfie* encourages exactly this—always playfully testing bounds (against Jodels image prefilter cf. C.5: Pre-filtering Images/Videos 🗉) risking a block and associated Karma loss (cf. C.3: Karma ↗), but through ironically happening discussion likewise self-regulating the community by examples on the extreme—appreciating any slightest breach upon platform measures, but mostly raising awareness and experiencing a baseline of sufficiently anonymized presentations. Ultimately, some specific contents are commonly appreciated like animal/cat pictures 🐾.

① Channels have been introduced to Jodel after our dataset observation time period (cf. D Ground Truth Dataset and Corpus) and thus cannot be examined.

C.5 Safeguarding the Communities

Employing a Distributed Moderation scheme, Jodel provides the users with a democratic way of determining the appreciation of content. Besides soft steering (cf. C.1 Social Priming) this process, the platform employs design and technical measures to enforce hard rules (cf. C.1 Platform Rules) to protect the community from adverse content and spamming.

Two-Staged Distributed Majority Voting and Moderation A major interaction type is liking or disliking content. It provides each user an easy and immediate way of expressing a like or dislike; whereas joining a discussion would require more time. This results in a democratic mainstream voting - however typically biased by user voting participation, which usually is inhomogeneous. In-app content presentation via feeds further strengthens the mechanism of mainstream appreciation in the *loudest* section (cf. C.2 Content Presentation).

COMMUNITY CONTENT VOTING 🗳️👍 While voting (cf. A.1 Jodel iOS mobile application. 7 / 9) provide a tool to participants actively determining what is liked—or disliked, Jodel’s rules will likely trigger users to dislike content disobeying them. As such, the voting filters spam, offensive and abusive content.

However, some cases are very explicit or may not be resolved relying on the community to vote, or may be buried deep within discussion threads being too long or old for catching any further attention of new users. Thus, an escalated moderation layer is ought to resolve such issues.

COMMUNITY MODERATORS 🚩★ Any user can flag any post that violates the platform’s rules. In case of gathering a thresholded amount of flags, this particular post will enter the Jodel moderation scheme. Here, a majority vote from a set of moderators is obtained, whether the flag is justified. Depending on human moderation decisions, a post may subsequently be blocked (and the user get removed Karma).

For this task, Jodel employs heuristically selected community members by their *well behavior*. Internals from this selection process are unknown, however the Q&A provides the following statement: “Moderating a community is a privilege reserved for our most trusted and positive users. There is a minimum requirement of karma to get into the pool of potential moderators. However, karma is not the deciding factor in the selection process.”¹¹

As mentioned earlier, becoming a moderator may be seen as an achievement by some users. Under the threat of loosing the moderator status, active reminders to (moderator) users for participating in content moderation is employed to trigger better participation. As a reward, participants are provided with a success-rating (agreement with other moderators) and a *picture of the day* after finishing a moderation streak.

Regularization via Karma ⤴ By design the regularization of Karma filters and restricts abusive users. Content is not displayed anymore reaching a cumulative negative score of usually -5. This represents a natural lower bound in downvotes that a post may receive, but provides no upper bound. Further, actions reward users with Karma or community-determined bad interactions get punished; this angle has been made even stricter introducing costs for voting negatively. To increase scores and prevent content from being widely vanished due to voting, this inherently dictates overall positive behavior.

¹¹<https://support.jodel.com/hc/en-us/articles/360001048074-Our-moderation-system->

Pre-filtering Images/Videos 🗣️ There are mechanisms in place to exclude explicit, harmful, and abusive contents, before they get published. This is especially important due to pictures/videos being very expressive and therefore sensible mediums. Urban legends assume that Jodel employs a purely algorithmic approach to image filtering; however Jodel does not reveal how they filter. Nonetheless, this usually incorporate AI models identifying at least candidates for human moderation; or contents may be desk rejected at high model confidence.

Rate Limits We have discussed systematic approaches to encourage positive platform behavior. As an actual hard prevention of spamming, Jodel also implements simple rate limits.

D Ground Truth Dataset and Corpus

A well-known smart person from the past years does not stop telling the story about data science work: You need to be a gold digger and the dataset is your gold mine. From my experience, this is only partly true: When joining research, public datasets within the realm of Online Social Networks (OSN) indeed were mostly nonexistent. Nowadays, the community has collected and released lots of datasets from various platforms. However, most of these examples are incomplete usually due to scraping and therefore sampling and limited to provided front-end information. Most importantly, a dataset is only as interesting as the questions you ask. Serendipity has led me to Jodel; a student in a practical seminar has been investigating this application w.r.t. privacy, which caught my attention. After crawling Jodel data at first, we were able to meet the founding team visiting Aachen (the app's *home-town*) and presented initial empirical insights, which has been perceived very well. A new cooperation was born, and I received the unique chance of investigating ground truth information directly from the operator—without any sampling artifacts.

D.1 Legal and Ethics

I am not affiliated to Jodel. Nonetheless, Jodel generously provided us with a data excerpt of their platform for scientific research purposes. The data has been provided to us under a Non-Disclosure Agreement (NDA). Provided explanations aim to describe challenges that are bound to using such an amazing dataset, and might not be law-proof in their briefness.

With great power comes great responsibility. Provided data is sensitive in various aspects: 1) Data is the new currency in today's internet era. That is, especially obtained platform interaction data represents the heart of the application. Naturally given, Jodel as a company aims to sustain and ultimately somehow monetize their platform—and thus cannot afford a breach of data. 2) More importantly, the data also contains actual message contents which (although being against the Terms of Service) may allow personal identification of individuals. Hence, any data breach is not be acceptable—and would impose negative consequences w.r.t. various juridical views and Jodel's Public Relations. The introduction of the General Data Protection Regulation (GDPR) specifically—and for the better—strengthened rights of customers. While the GDPR implementation arguably might not be optimal, it at least has lead to awareness for data pseudonymization, anonymization, data sparsity, and defines serious responsibilities.

That is, we needed to take state-of-the-art technical and process-related precautions to deal with this dataset responsibly.







Implementing Required Processes Specifically due to GDPR, we must implement certain processes. E.g., we must track which data has been used where and who has been

given access. Further, we must comply to user requested data deletions. Generally inform with Jodel about planned evaluations and discuss results. This process consists of a mixture of communication channels such as asynchronous chat as found in many Start-ups (agile), meet-ups—or video conferencing within the COVID-19 pandemic.

Technically Safeguarding the Data Though an implementation would be possible, due later obligations w.r.t. to data, e.g., emerging from GDPR requirements, we opted for not including the data on our long term storage; providing us with a single point of truth. To safeguard the server, we placed it into an internal network and enabled the firewall—solely accessible via ssh keys. Only authorized governed personnel may access this server. Further, we enabled encryption on file-system level.

Data Sensitivity and Sparsity For executing specific analyses at our University, we provided access to limited sparse subsets of this data through specific user accounts managing access. While most metadata cannot be used to identify persons and thus is less sensitive, content is a different topic with potentially wider consequences. Although users agree to the Jodel Terms of Service when participating on the platform and agree to publishing their contribution into the *public*, this means only the platform itself. As such, we e.g., cannot employ third party instances to analyze or label any data.













D.2 Dataset Description

The basic raw dataset as obtained is simple, i.e., the bare minimum data required to run the application. A relational normalized star datastructure connects the main objects participating in any platform interaction. That is, we have timestamped *interactions* ($n : m$) that link users to posts with a restricted set of interaction *types*:  up- and  downvoting, creating a  new post,  replying,  a registration event occurring whenever Jodel encounters a clean device, or  flagging a post for moderation. Due to the nature of Jodel being hyperlocal, the content is bound to a location—therefore all content require location information. As Jodel organizes conversational threads linearly; this manifests in a simple self-relation ($1 : n$) on the content table: A key to a parent content object suffices that may be empty for posts. Furthermore, a flag whether content is blocked on the platform (due to voting or moderation) is present in addition to a user-level blocked status. The user table also comprises accumulated individual karma scores.

Please refer to Table D.1 for a detailed listing of amounts of present data.

We provide an overview of the data structure as follows; note that I marked primary keys bold, whereas underlining denotes foreign keys.

Interactions

- **interaction id** 
- type in [ upvote,  downvote,  post,  reply,  flag,  registered]
- timestamp   (*minute accuracy, according to related object*)
- content id 
- country  (*demormalized for partitioning*)
- user id 

| 🌐 Country | DE | SA |
|--------------------------|-------------|-------------|
| 📍 Communities | 6.8k | 95 |
| 🌱 Interactions | 3036M | 966M |
| - 👍 upvote / downvote 🗳️ | 2.3G / 465M | 350M / 143M |
| - 🚩 flag | 4.9M | 4.2M |
| 📧 Content | 285M | 469M |
| - 📄 post / reply ↩️ | 49M / 235M | 58M / 411M |
| 👤 Users | 3.6M | 1.2M |

Table D.1: Dataset Statistics.**Content** ✅

- **content id** ✅
- parent id ✅ (*empty if item is post*)
- location 📍
- country 🌐 (*derived from location*)
- blocked ☹️
- accumulate vote score ★ (*materialized*)
- message 📄

User 👤

- **user id** 👤
- karma ↗️
- blocked ☹️

Community Discretization Jodel does not have any specific community boundaries, but a user’s feed always contains content from her currently reported smartphone position. As such, content is always bound to a position and thus is being display to users within a radius of this very location. Used location data is store as geohashes¹², which conveniently allow for coarsening the grid into suitable sizes. This allowed efficient partitioned shortest distance searches over cultural map data from NaturalEarth¹³ assigning all Jodel locations to a specific near city. I argue that this discretization makes sense largely correlating with local urban population clusters. This might overestimate the content within larger cities as seen by its users, however, larger cities subdivide further into districts. Further, due to personal mobility, e.g., home vs. school or work, most users will perceive and reach broader audiences anyway.

Denormalization favoring ClickHouse Aggregation For our data evaluation, we first have tried simple relation databases (MariaDB), which happen to be too slow calculating large aggregates or join; their strength typically relies in indexed access and depending on storage engine ACID compliance. Easy to use semi-structured data lakes, such as mongodb suffered from the same problem, unless one would create an parallelizing framework on top. Long story short, we discovered *ClickHouse*, an Online Analytical Processing (OLAP)

¹²<https://en.wikipedia.org/wiki/Geohash>¹³<https://www.naturalearthdata.com>

data management system, as a large scale resource-friendly and efficient tool for handling aggregates on present amounts of data.

Due to clever compression and data ordering as a columnar storage, ClickHouse asks for denormalization into very wide tables. Although I could have merged all data into a single global table, I opted for adding various derived fields for convenient access; for later speedups, such views were materialized. Still, huge constructs of nested joins across tables might be CPU singlecore bottle necked and specifically memory-hungry operations. Some evaluations required further optimized table engineering decoupling large aggregations into substage materializations.



BACKGROUND

CONTENT

| | | |
|----------|-------------------------------------|-----------|
| A | Crowdsourcing | 29 |
| A.1 | Human Labor | 29 |
| A.2 | Quality Assurance and Incentives | 29 |
| A.3 | Questionnaire Design | 29 |
| A.4 | Evaluation | 30 |
| A.4.1 | Scale Normalization | 30 |
| A.4.2 | Interrater Agreement | 30 |
| B | Machine Learning | 30 |
| B.1 | Approaches | 31 |
| B.1.1 | Unsupervised Learning | 31 |
| B.1.2 | (Semi-)Supervised Learning | 31 |
| B.2 | Learning Blueprint | 31 |
| B.2.1 | Feature Engineering | 31 |
| B.2.2 | Data Balancing | 31 |
| B.2.3 | Data Splitting | 32 |
| B.2.4 | Data Normalization | 32 |
| B.2.5 | Hyperparameter Tuning | 32 |
| B.2.6 | Evaluation | 32 |
| B.3 | Algorithms | 33 |
| B.3.1 | Decision Trees | 33 |
| B.3.2 | Deep Learning | 33 |
| B.4 | Representation Learning | 34 |
| C | Natural Language Processing | 35 |
| C.1 | A Brief History | 35 |
| C.2 | Neural Sequence Embeddings | 35 |
| C.2.1 | Transformers - Leveraging Attention | 35 |
| C.2.2 | BERT | 36 |
| C.3 | Word Embeddings | 37 |
| C.3.1 | Limitations | 37 |
| C.3.2 | Architecture and Training | 37 |

| | | |
|----------|--------------------------------|-----------|
| D | Social Network Analysis | 37 |
| D.1 | User Adoption | 38 |
| D.2 | User Interactions | 38 |
| D.2.1 | Structural Analyzes | 39 |
| D.2.2 | Modeling | 40 |
| D.3 | User Content | 40 |
| D.3.1 | Information Diffusion | 40 |
| D.3.2 | Problematic Content | 41 |
| D.4 | User Management | 41 |

Introduction

Within this background section, we want to highlight some general concepts that find application in this thesis. Please note that we will refrain from re-explaining existing concepts, but rather provide an overview with suitable pointers to in-depth information.

A Crowdsourcing

To acquire any insight to data, there are many technical unsupervised tools available. Yet, understanding textual contents and specific content questions may only be answered by humans, or require at least some labeled data being provided to algorithmically learn from. Crowdsourcing enables to scale-out human annotation tasks to a larger set of users that may receive payment for their efforts and is used that way within this thesis. While the very definition of crowdsourcing primarily focuses on publicizing human labor or other resources to larger user groups, we are interested in the manual labor part. That is, in our realm investigating large datasets, crowdsourcing helps to distribute load across many individuals to save time and potentially cost. Though focusing on quality of experience experiments, [Hoßfeld et al., 2013] discusses best practices. We complement these best practices with a rough overview of essential background information needed to conduct crowd-sourced campaigns as applied in [★ B\) Platform Sidechannels \[JH1\]](#) and [★ B\) Content SA \[JH4\]](#).

A.1 Human Labor

The human labor is conducted by annotators, or called coders, may be of a different types w.r.t. trust. In case the coders are specifically known and possibly doing their work in a controlled environment, we usually talk about trusted coders. Otherwise, using public or unknown resources, they are untrusted. Both cases may have a different impact on necessary quality assurance measures as described later.

Depending on subject under evaluation, specific tasks may further require (trusted) expert coders, not only doing crowdsourcing occasionally, but larger portions including feedback, e.g., group discussion. However, as shown for language tasks, non-expert coders may replace experts by using a larger amount of coders [Snow et al., 2008].

A.2 Quality Assurance and Incentives

Due to humans possibly being mistaken or uncertain, crowdsourced campaigns usually rely on multiple classifications for a single entity, allowing for more confident estimations. However, depending on individual incentive for doing work, coders might be tempted to simply rush through questionnaires to receive a payment. In other occasions, coders might misinterpret tasks and therefore provide unsuitable answers. In any case, a typical—and necessary—quality assurance measure are well-selected test questions (alike re-confirming answers within questionnaires). Such questions are non-trivial as they should not be obvious, but still need to be unambiguous. Should a coder provide too many false answers for these tests, all of her results should be discarded for any further evaluation. However, e.g., demographics inherently cannot be cross-checked, especially for untrusted coders.

It is suggested to estimate coder effort and provide them with at least minimum wage.

A.3 Questionnaire Design

As mentioned before, QA measures usually include a set of test-questions that indicate whether a coder has understood the task and is answering with effort, i.e., non-harmful.

It is generally preferable to create possibly rather multiple easy-to-answer questions instead of a single more complicated question that might create heavier cognitive load. While simple binary answers might suit certain questions well, many crowdsourced campaigns provide the coders with a set of multiple options on a scale, e.g., bad, neutral, good. Such Likert scales typically have different ranges that allow coders to provide various quantized answer options; scale with more than six response options have shown not to increase precision [Simms et al., 2019]. Further, depending on the task, it might be arguable whether the amount of response options should be odd or even, of which the latter might allow for a neutral option. It might also be helpful to add visual cues to the possible answers [Funke and Reips, 2012].

A.4 Evaluation

For evaluating a campaign, any coder responses not passing a defined threshold of correct test questions should immediately be discarded. We describe typical subsequent tasks next.

Depending on goals of a crowdsourcing campaign, chosen data and classification results may be subject to additional significance tests.

A.4.1 Scale Normalization

Mentioned Likert scales may be subject to user heterogeneity. Due to individual, cultural, or language differences, users might tend to use the available extent to varying degree. I.e., while some users might consistently choose a pattern of rather intermediate values, others might prefer extremes. Thus, Likert scales are subject to biases in responses.

Possible solutions to normalize the scales include per-user min-max or std normalization. Other more sophisticated approach take a Bayesian approach [Rossi et al., 2001], or take a more general approach leveraging a multipole representation [Lipovetsky and Conklin, 2018].

A.4.2 Interrater Agreement

Having multiple responses for the same item from multiple coders raises the question of their reliability, i.e., how certain are the coders in their decision. While simple percentages or majority voting schemes might be applied, they do not account for random chance. That is, other metrics are capable of providing more robustness. A common choice Cohen's kappa [Cohen, 1960], which however requires the same items to be classified by the same coders, which may not necessarily be the case. To overcome this issue, research has adopted other alternatives that allow for *holes* within the classifications of various coders, such as Krippendorff's alpha [Krippendorff, 2011, Krippendorff, 2012].

B Machine Learning

Machine learning (ML) describes a part of Artificial Intelligence that tries to model distributions leveraging historical data. There exists a plethora of various methods, of which most rely on statistical methods. All of these approaches have in common, that it is hoped to generalize well to unknown new data, i.e., that data resembles a similar distribution. ML methods are becoming more popular across any research field and applications. While there exists different types of algorithms and goals, we will focus on describing the rough concepts involved in classification tasks as applied in [★ A\) Information Diffusion \[JH9\]](#), [★ C\) Emoji \[JH7, JH10\]](#), and [★ A\) Distributed Moderation \[JH5\]](#).

B.1 Approaches

ML is a highly active field of research with lots of variants having evolved over time. Approaches may be attributed with different rough categories of un-, self- and (semi-)supervised, or e.g., reinforcement learning. Of which we will detail only the former briefly.

B.1.1 Unsupervised Learning

In case the algorithm is not using any labels for data, but trying to create structure or meaning from the data nonetheless, we talk about *unsupervised* learning. Most of such algorithms are usually fed specific parameters that impact outcomes. Simple examples are nearest neighbor clustering, requiring max cluster distance and number of cluster parameters; or for determining topics within provided text e.g., Latent Dirichlet Allocation, similarly to other clustering approaches requires at least a predetermined amount of target topics. As these algorithms usually rely on computational metrics, the quality needs to be assessed separately before results make specifically sense to humans. Yet, such technically determined metrics may be sufficient for, e.g., determining some kind of similarity.

Though arguably not necessarily being unsupervised learning techniques in a narrow sense, methods of finding data representations are likewise a result of rather technical measures.

B.1.2 (Semi-)Supervised Learning

In contrast to unsupervised learning, supervised learning leverages existing labels of data. That is, algorithms are specifically trained to model a function of input data to corresponding labels. The general assumption is that historical (labeled) training data then generalizes to unknown inputs.

Between un- and supervised learning, a semi-supervised approach uses only partially labeled data alike in supervised approaches, but incorporates unlabeled data to increase model confidence and generalization capabilities. Self-supervised learning on the other hand usually produces and uses its own pseudo-labels.

B.2 Learning Blueprint

Any machine learning approach requires a specific setup of various stages to fulfill a classification task to perform well. The very base for any applied ML is sufficient high quality data. I.e., before we can model distributions, we need to collect sufficient amounts of representative information that preferably is unbiased—otherwise the model will likewise learn these biases.

B.2.1 Feature Engineering

The first step incorporates making input data tangible for subsequent computation: feature engineering and selection. The goal here is to (manually) determine useful features that can be leveraged to learn data distributions. While the space of possible features is huge, feature selection tries to determine the most valuable inputs, e.g., applying simple heuristic like focusing on large input variances that might contain more information than others.

B.2.2 Data Balancing

Depending on the gathered dataset, labels for classification might be imbalanced, introducing a bias. However, in most cases, a balanced data for training should be used. While, if applicable, augmentation might artificially enlarge a dataset, there exists several methods like over or under sampling of data to overcome this issue.

B.2.3 Data Splitting

After selection a suitable (balanced) dataset, we split the data (at least) into a training and test set; additionally, the data might be split into an additional validation subset. While the training subset is used to teach the model, the test and validation subset are used to measure model quality and detect possible overfitting to provided training data. In case the amount of data is comparably high, a single split might be seen as sufficient, as often seen in Deep Learning applications. However, specifically with smaller datasets, a cross-validation can help to determine model generalization capabilities by creating independent dataset-splits and creating models for each of these subsets. If all of these independent splits achieve comparable performance, a dataset bias can be ruled out.

B.2.4 Data Normalization

For many learning algorithms, it is beneficial to normalize the input feature values, i.e., to adjust the range of input values. While there are many methods to implement normalization, standard normalization, or min-max normalization are among the most commonly used alternatives. The standard norm shifts values by the data mean and scales them by the standard deviation, whereas a min-max scaling distributes values by apparent value extreme values. Both methods usually target a destination range from zero to one.

Regardless of the applied method, it is important to scale training and test/validation sets independently. Though the impact might arguably small in most case, information from the test or validation set may leak into training data otherwise—giving the algorithm an unwanted advantage.

B.2.5 Hyperparameter Tuning

Most machine learning algorithms have various algorithmic parameters (hyperparameters) that might be adjusted. Generally, it is unknown which parameters provide best results. Though experience may provide a starting point, best suited parameters are subject to tests. That is, a simple grid search, i.e., a cross product between all parameter options to be tested, is conducted. However, this process might be more efficient when using smarter approaches, such as Bayesian optimization within this search. Note that the final choice of model parameters may be a trade-off between possibly better performance, but computationally more expensive model configurations.

B.2.6 Evaluation

Depending on the task, a Machine Learning algorithm's quality can be measured by various metrics. For regression tasks, the Mean Squared Error (MSE), the Mean Absolute Error (MAE), or R2 scores might be used. The evaluation of classification tasks is carried out with confusion matrices and derived metrics. That is, we pairwise count how often a specific model estimates correct or wrong labels. As for binary cases, this might result in true positives (correct matches), true negatives (correct non-matches), or wrongly assigned labels in false positives, or false negatives. These metrics generalize to multi-label classifications.

Among most popular derived metrics, many researchers provide an algorithm's *Accuracy* describing the total fraction of correctly predicted labels. However, depending on the task, this aggregation is not sufficient to paint the full picture. Other metrics like *Precision* (fraction of true positives to overall positive predictions), or *Recall* (fraction of true positives to true positives and false negatives) provide deeper insights into misclassifications. The *F1* score, i.e., geometric mean between Precision and Recall, provides an overall picture.

Note that the different metrics become important to understand model implications in practice due to often occurring imbalanced data.

B.3 Algorithms

In the realm of machine learning, many algorithms and approaches exist. While the simplest form of learning from data, might be simple histograms enabling a best guess for the most occurring label or approximating functions with various regression methods, others apply smarter methods. While a single algorithm might perform well, accuracy may be improved by leveraging multiple approaches, i.e., ensemble multiple models. Due to discussing most classical methods alone would be quite extensive, we will detail the example of Decision Trees (DT)—and the use of multiple DTs into Random Forests (RF) as an example of such ensembles.

B.3.1 Decision Trees

Decision Trees have a long history in supporting decisions [Breiman et al., 1984, Loh, 2014]. The main idea is to split the space of available input features hierarchically into different subtrees of which the leafs correspond to a certain outcome, or w.r.t. classification label. This idea is quite similar to a generalized binary search and thus is comparably cheap in inference. Creating such trees is a non-trivial task as its structure needs to be optimized for best results. Main parameters are tree features, i.e., how many features should be incorporated to a decision, tree depth, i.e., how many decisions should lead to an outcome, and when or how to split a node into and how many samples should still be required to make a reasonable additional decision step.

B.3.1.1 Random Forest

As mentioned before, ensembles of models may improve prediction quality. Though these classifiers do not necessarily have to be of the same type, Random Forests represent ensemble of decision trees. Due to being initialized randomly, decision tree instances usually appear to be different depending on the used seed. By training several DTs possibly in parallel, each being an *estimator*, albeit increasing computational costs, we can leverage multiple DTs to improve prediction quality by conducting a majority vote from all used instances.

B.3.1.2 Feature Importance

Describe earlier, feature selection is a crucial task in creating efficient models. There generally exist two type of metrics that may be applied: a priori and a posteriori. While a priori heuristics deduce some kind of information gain from the input feature set, such as variance (higher variance may indicate richer information gain), posteriori methods are calculated from create models.

In the case of decision trees, the Gini index is a common method to provide insights into the model's decision-making process. This index calculates the extent to which a certain feature is involved into decision processes as a probability. The Gini index is calculated by the sum of squares of probabilities for each class.

B.3.2 Deep Learning

The term Deep Learning refers to a huge class of various architectures finding widespread use in state-of-the-art machine learning approaches. These methods rely on Artificial Neural Networks that adapt the biological process of neurons in our brain built from only few basic building blocks. With improving computational capabilities, the field of deep learning has attracted much research and application focus. Thus, we will only scratch the surface and explain the very basic functionalities.

B.3.2.1 Neurons

An artificial neuron basically has possibly multiple inputs and outputs, while internally conducting some computation, i.e., representing a mathematical function. This function might simply add inputs, or even execute non-linear functions, e.g., specific activations. Bundling many neurons allows the approximation of any function [Leshno et al., 1993].

Though a single neuron appears simple, large neural networks can become complex and thus also approximate more complex functions. Advanced methods that add feedback loops might e.g., add *memory* capabilities with Long short-term memory (LSTM) [Hochreiter and Schmidhuber, 1997] constructs, or Gated recurrent units (GRU) [Gers et al., 2000].

B.3.2.2 Loss Function and Optimization

To train a neural network, we generally need an objective for optimization in a mathematical form: a loss-function determining the quality of a network's weights for some input. Though non-gradient, or synthetic gradient algorithms [Chen et al., 2017, Jaderberg et al., 2017] exist, most applications rely on gradient based methods for training.

Gradient-Based Algorithms. By feeding a data point into the network, we can calculate all resulting weights and determine the current loss. The general idea now pursues finding local minima within the whole loss landscape. To achieve this, gradients through the network are calculated and weights can be updated to improve the loss value towards such a local minimum. A standard optimization algorithm for this task is Stochastic Gradient Descent (SGD). Many other variants, e.g., Adam, may provide better performance, i.e., they may converge faster and thus need less computation.

Model Hyperparameter Search. As discussed earlier for other machine learning techniques, most algorithms can be tuned with hyperparameters. This is the very same in deep learning. Most important parameters are e.g., batch size (how many data items do we process per optimization step), or the learning rate (determining the gradient descent step length). Depending on the used optimization algorithm, there might be more parameters available to tune learning and model performance.

To overcome overfitting issues, one might implement dropout (random re-initializing single weights), or complete layer re-initialization. There exist countless other advanced techniques depending on network architecture, type and application, e.g., weight decay.

B.3.2.3 Geometric Deep Learning

[Bronstein et al., 2021] only recently provided a generic formalization approach to *many* of today's deep learning approaches and basic architectures. They compile a framework of suitable abstractions to e.g., Convolutional Neural Networks, Graph Neural Networks, Transformers, or Recurrent Neural Networks. Stemming from this, it becomes obvious how and why these different types of networks are more or less powerful/expressive, or computationally expensive than others.

Though their work may not yet have reached the broad mainstream, we strongly believe that this formal top-down approach helps in understanding deep learning and specifically connecting the dots between seemingly different basic approaches. However, due to this topic being *very* extensive, we will refrain going into further detail.

B.4 Representation Learning

Often times representation learning is also referred to as feature learning. The very goal is to transform inputs into more suitable representations for subsequent machine learning

approaches, such as classification. I.e., instead of manual feature engineering as discussed before, features are learned, which still might be accomplished using either approaches of un- or supervised learning. Such approaches have become popular especially within deep learning, but are not restricted to it. That is, many approaches embed input information into dense vector spaces with specific optimization functions, embeddings.

We will detail a popular approach to embed words or tokens within the next section according to word co-occurrence. Other embedding approaches are specifically prevalent in NLP, such as BERT [Devlin et al., 2019a] and alike modern Language Models. However, such embeddings can create powerful tools that e.g., encode graph nodes [Grover and Leskovec, 2016], or allow for logical queries [Ren et al., 2020, Ren and Leskovec, 2020, Hamilton et al., 2018].

C Natural Language Processing

Natural Language Processing (NLP) has become an extensive field in research and industry over past decades. It originates from linguistics, but computational methods have found heavy impact. The main goal is to understand and make use of human language in interaction between humans and computers, or technically mediated human interactions. Adjacent areas revolve around natural language understanding and generation where boundaries have become fluid from an algorithmic perspective. In this thesis, NLP methods find application in ★ B) Content SA [JH4], ★ C) Emoji [JH7, JH10], and ★ A) Distributed Moderation [JH5].

C.1 A Brief History

While very early works up into the 1990's focus on symbolic representations, ontologies, and grammars, statistical methods have become more popular specifically due to the rich availability of digitalized texts. With availability and increasing computational power, later methods extended on such approaches introducing machine learning methods. I.e., manual feature engineering and probabilistic methods gained traction, e.g., LIWC¹ as used in [Tausczik and Pennebaker, 2010], or leveraging linguistic properties.

However, with the advent of deep learning, Neural NLP has radically changed the field of NLP with ever-increasing capabilities. Transformer-based [Vaswani et al., 2017] architectures that have emerged in the realm of NLP are still state of the art. Though and specifically because representing an inherently computationally costly method, the key idea of leveraging attention has spread across almost all deep learning domains with great success.

C.2 Neural Sequence Embeddings

Deep learning has unlocked a new era of NLP. By leveraging representation learning techniques, it is no longer necessary to manually develop extensive features. That is, sequences of words (tokens) are fed into these networks to learn contextualized representations. Earlier approaches used recurrent neural networks (RNNs), such as bidirectional encoding LSTM networks, such as Embeddings from Language Models (ELMo) [Peters et al., 2018], with great success [Huang et al., 2015, Chiu and Nichols, 2016, Ghosh et al., 2016, Zampieri et al., 2019].

C.2.1 Transformers - Leveraging Attention

However, [Vaswani et al., 2017] has proven to become a breakthrough introducing the Transformer architecture, not only in language modeling. Though transformers only do

¹<https://www.liwc.app/>

set matching, adding positional encoding to the input, the set matching artificially becomes sequence matching. Note that we refrain from discussing this architecture in detail as it is unimportant to understand its basic functionality—the interested reader will find much detail in the cited original paper, or specific books and webpages on this very topic.

On a very high level, transformers represent a *soft* hashmap. Similarly to many other architecture types, the transformer is created from an encoder and decoder. Both types leverage multi-headed (self-)attention blocks, which are responsible to correlate input sequences to itself, or between encoder and decoder sequences incurring quadratic complexity. While the encoder does not do much more, the decoder introduces masks, which blank out certain tokens from its input sequence. These blanks are then to be estimated by the model—this is how the learning happens.

Reducing Complexity. As powerful as attention has proven to be, the attention mechanism with a quadratic complexity introduces new challenges. I.e., sequence lengths are very restricted by today’s hardware to only hundreds of tokens. This apparent shortcoming has been subject to various research efforts to reduce complexity with e.g., localized attention [Choromanski et al., 2020], linearly approximating attention [Wang et al., 2020], adding the frequency domain [Lee-Thorp et al., 2021], or learning attention masks [Tay et al., 2021]. However, full attention still finds widespread use as of today.

C.2.2 BERT

Among others like GPT [Radford et al., 2018], one of the very first presented language model using transformers was introduced with Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2019a]. In contrast to earlier works using bidirectional LSTM architectures, new architectures allow complete sequences to be handled at once instead of sequentially applying computational steps. Furthermore, a key advantage is that the model has access to the complete input sequence instead of only the previous or next element, allowing for better contextualization. The key idea of BERT is to leverage transformers with masked language modeling (MLM), which blanks out elements of the decoder input and reconstructing the input to learn representation in an unsupervised fashion, creating embeddings of the input.

Such models require huge computational efforts in learning from very large corpora using millions of sentence—pretraining the model. This pretraining only needs to be done once and may then be fine-tuned to specific tasks by adding only an additional layer e.g., for classification, question answering, or next sentence prediction.

Explainability. Due to its success and popularity within NLP, research has tried to better understand model behavior and its decision-making processes. That is, [Tenney et al., 2019] show that BERT models approximately re-invent a classical NLP pipeline w.r.t. linguistic information. [Clark et al., 2019] provides more detail showing that certain attention heads fulfill similar tasks focusing on specific linguistic properties, positional offsets, or specific sentence separators. Others propose leveraging attention flow to better understand decisions [Abnar and Zuidema, 2020].

BERT-Evolution. Research has shown how to improve BERT training and provided lots of variants w.r.t. training, such as RoBERTa [Liu et al., 2019], or used corpora, publicly providing us with very capable models of various model sizes, languages [Scheible et al., 2020, Antoun et al., 2020], and multilingual models. The Hugging Face² community provides central access to a plethora of BERT-like models.

²<https://huggingface.co/models>

C.3 Word Embeddings

While discussed neural embeddings recognize context, i.e., incorporate complete sequences, an earlier method focuses on distinct input words (or n-grams) only. The optimization goal of word embeddings is to embed words into a dense vector space according to word co-occurrence within input sentences from a corpus. That is words often appearing in proximity of each other within the training data will be located close to each other within this embedding space as well. Thus, they have shown to successfully model semantics and may even model linear combinations to a certain extent [Mikolov et al., 2013c] as we likewise show for word-emoji embeddings in ★ C) [Emoji \[JH7, JH10\]](#) .

There exist several methods of which Word2Vec was one of the first examples [Mikolov et al., 2013a], while Global Vectors for Word representation (GloVe) [Pennington et al., 2014] provide a similar outcome. Due to input words denoting distinct input tokens and thus, stemming and stop word elimination might be desirable, [Bojanowski et al., 2017a] promote leveraging subword information in terms of n-grams.

C.3.1 Limitations

Word embeddings that map single words into a vector space are quite simple due to disregarding context, which makes them easy to understand and easy to use. However, specifically lack of context implies the major limitation that such embeddings cannot distinguish between multiple semantics or meanings of words as they occur in language. Nonetheless, the widespread successful application and computationally low-cost inference shows that this problem can be managed.

C.3.2 Architecture and Training

Continuous Bag of Words. Quite similar to a feed-forward neural network, the CBOW methods uses for each input word several surrounding words in a sliding window fashion. By aggregating these input words, the network then tries to reconstruct the input word [P., 2021].

Continuous Skip-Gram. By essentially inverting the direction of the CBOW architecture, another approach resembles a triplet-loss [McCormick, 2016, P., 2021]. The model architecture consists of a hidden layer and an output softmax. That is, from an input sentence, for each of the contained input words, tuples with surrounding words are mined. Then, they are fed into the network that then tries to predict the probabilities of the other words surrounding the target. This process may be improved with negative sampling.

D Social Network Analysis

Please note that the *somehow* related research space is massive. That is, we detail related work on Jodel and similar platforms. We further provide general directions within literature outlining details that fit to presented contents of this thesis. Within each subsequent chapter, we provide additional in-detail task-specific Related Work.

Social network analysis is not a new idea. In his book [Freeman, 2004], Freeman describes the development within the early days. Though borders between online platforms being a social network may be blurry, Otte et al. argue in [Otte and Rousseau, 2002] that social network analysis (or in a broader sense, user interaction graphs) is a key enabling tool within various fields at unprecedented scale.

D.1 User Adoption

As explained earlier, monitoring growth processes within Online Platforms often fall short without cooperation with the operator and rely on crawling and sampling. Thus, related research within this specific field is comparably poor.

Nonetheless, [Mislove et al., 2008] monitor and evaluate the growth of flickr³, a social image and video platform, having crawled the platform over 104 days. Their final dataset consists of 950k users denoting a growth in users of about 58% within this timeframe. Likewise, social links have increased by 63%, which form rather fast, i.e., about 80% of new links establish in only a single day, being heavy-tailed. They further elaborate on social network graph properties like in/out degrees following a power-law, and graph distances being very short at only 2-3 hops for more than 95% of all links. Additionally, [Kumar et al., 2010] investigate flickr and compare structural finding with Yahoo!360. They identify various components and showcase the prevalence of star structures regardless of network size.

The nowadays already shut down Social Network Google+ has been investigated within [Gong et al., 2012]. After an initial bootstrapping phase, where users were only able to join on invitations and hence linear growth, they identify exponential growth going public. By analyzing various graph properties in the social structure, they identify stark differences that can be attributed to these phases. They also identify power-laws in degree distributions, while social attributes follow a lognormal rule. Finally, they present a generative preferential attachment model depicting network structure characteristics.

While so far discussed literature only covers relatively short timeframes, though also relying on crawling and sampling, [Singer et al., 2014] extend their evaluation of Reddit⁴, a link sharing and discussion platform, to multiple years from 2008 to 2012. Their key observation is an observed exponential growth in post volume across various subreddits; while the main reddit.com subreddit experiences a heavy focus of around 60% in 2008, over time, they observe a major shift of attention to other subreddits. By classifying link types, they observe a major shift to self-references ($\approx 40\%$) against text posts and an increased number of shared images.

Early Social Network implementations usually use a central instance from a single operator. Similar to e.g., XMPP⁵, Diaspora⁶ uses a decentralized approach. This allowed [Bielenberg et al., 2012] to gather information from a plethora of different servers for showcasing various growth patterns. While also encountering typical power-law distribution in users per server, they also show that the influx of new users per server is likewise linked to the existing amount of registered accounts. They detail single server user counts over time, identifying a steady slow influx at first. At single changing points, they observe huge additional amounts of new users in short timeframes with an observed following heavier steady growth. Though being a hard question, the authors owe answers of possible reasons for changing evolutions and non-shown other instances, however.

D.2 User Interactions

Possible insights and patterns of user interaction across possible online platforms are countless. In this section, we provide some examples of specific topics under investigation. We detail a large body of research focusing on structural insights mostly applying graph methods to social ties, while also discussing example of interaction modeling.

That is, [Kairam et al., 2012] provide empirical insights the Google+ network, but also implement a qualitative survey about motivating usage factors.

³<https://www.flickr.com>

⁴<https://www.reddit.com>

⁵<https://xmpp.org>

⁶<https://diasporafoundation.org>

The Mastodon⁷ network denotes a new decentralized Social Network similar to Diaspora. In [La Cava et al., 2022], the authors analyze various instances and find a power-law distribution in user counts. Moreover, they categorize users into producers and consumers. While only about few users are active on multiple servers, they push the evaluation to cross instance comparisons of the very same users.

Another approach to reveal user behavior is tracing clickstreams (sequences of interactions). In [Benevenuto et al., 2009], the authors characterize user behavior from such clickstreams primarily on the Orkut platform, and others (Hi5, MySpace, LinkedIn). They show empirical insights to weekly patterns, session lengths and inter-session durations, or inter request times, identifying that users mostly interact with their social ties/friends. The most interested finding is that most user activities are invisible as they do not result in any state change within the system; this passive participation is also called lurking.

Another example of extensively analyzed platform is Reddit⁸. Comparing Twitter, Reddit and 4chan, [Mittos et al., 2020] provide a quantitative study including spatial distributions, but also look into contents. They partly find quite toxic behavior. More specifically, e.g., [Kamarudin et al., 2018] take a detailed look into Reddit users' responses to rape in another qualitative study. Reddit as a platform furthermore establishes domain specific slang with smaller vocabulary compared to Google News as shown in [Ferrer et al., 2021]. Other quantitative data-driven studies show that the volume of posted content within political communities is heavy tailed [Soliman et al., 2019]. However, political interactions at the example of the US 2016 elections have been identified not creating typical echo chambers [De Francisci Morales et al., 2021] as found on other platforms.

Others focus on specifics in user behavior, e.g., on StackOverflow [Yang et al., 2014], a Q&A platform, where the authors characterize expert behavior and introduce a metric that enables their identification.

D.2.1 Structural Analyzes

While presented works focus on specific aspects, more general approaches usually characterize social network structures by graph metrics from emerging social ties. As mentioned earlier, most literature bases upon sampled content as discussed in [Catanese et al., 2011], in which the authors provide graph insights and argue about sampling biases.

An in-depth view on the Microsoft Messenger [Leskovec and Horvitz, 2008] provides global spatial insights to platform usage. Unsurprisingly, major findings indicate that people likely interact with others at the same age, location, and language. Furthermore, cross gender conversations tend to happen more often and typically longer.

Social ties graph structures have been a driving factor for countless research manuscripts within the earlier Social Network analysis days across a plethora of different platforms. All examples have in common that emerging graph structures happen to be small-world scale-free graphs with striking qualitative similarities. While most distributions w.r.t. interactions, participation, and component size usually depict power-law (possible with a cut-off) or lognormal distributions, path lengths are short with densely connected cores and high clustering coefficients. This has been shown for e.g., flickr, LiveJournal⁹, Orkut, Youtube¹⁰, MySpace¹¹, Hi5¹², Facebook¹³, Google+, and Twitter¹⁴ or Mastodon [Mislove

⁷<https://mastodon.social>

⁸<https://www.reddit.com>

⁹<https://www.livejournal.com>

¹⁰<https://www.youtube.com>

¹¹<https://myspace.com>

¹²<https://hi5.com>

¹³<https://www.facebook.com>

¹⁴<https://www.twitter.com>

et al., 2007, Wilson et al., 2009, Ugander et al., 2011, Schiöberg et al., 2012, Magno et al., 2012, Gong et al., 2012, Zignani et al., 2018, Zignani et al., 2019, Raman et al., 2019].

Some of these works also focus spatial distributions showcasing rather local clusters and primarily in-country interactions [Ugander et al., 2011, Schiöberg et al., 2012, Magno et al., 2012].

D.2.2 Modeling

While analysis of platforms social ties and graph structures provides interesting insights, the obvious next step is specific modeling of user behavior. Some of the previously presented works already provide insights to distribution fitting with e.g., lognormal or power-laws, however, these models only scratch the surface being quite abstract. That is, like many others, e.g., [Gong et al., 2012] discuss a generative model leveraging preferential attachment, and incorporating other attributes.

Another early work [Tang et al., 2009] models topic influence within the social graph with what they call Topical Affinity Propagation (TAP). Using multiple topic attributes per user, they model a probabilistic weight distribution across the graph structure. Likewise, [Ferraz Costa et al., 2015] propose a Rest Sleep Comment (RSC) generative self-correlated process model for temporal user interactions.

While already discussed empirical works also show mainly power-law distributions, simpler modeling approaches use lognormal distributions describing Social Network distributions for the Digg¹⁵ platform in [Van Mieghem et al., 2011]. They further show that probabilistic Bernoulli relations affect individual behavior. Yet, others show power-law distributions across posts and comments on Reddit in e.g., [Thukral et al., 2018].

D.3 User Content

Determining types of content is a non-trivial task. There generally exist a plethora of approaches between two extremes: unsupervised and supervised content classification. Unsupervised methods try to find and cluster topics according to technical metrics, e.g., Latent Dirichlet allocation (LDA) using a generative statistical model, or newer approaches leveraging neural embeddings and clustering, e.g., BERTopic [Grootendorst, 2022]. In practice, they tend to be tricky finding suitable algorithm parameters that result in topics making sense. On the other hand, the simplest type of supervised classification is using manual labor, possibly crowdsourced, to classify contents. However, likewise to the unsupervised methods, mining viable and (for one's task) suitable classification schemata may remain an incredibly hard task. Building on top of such manually labelled contents, there exist many approaches in representation learning for creating an automatic classification that leverages manual input for scaling out.

While specific topics of created content on various platforms is very diverse, we opt to subsequently provide an overview of detailed insights w.r.t. contents and their implications for platforms.

D.3.1 Information Diffusion

Online platform live from participation and certain topics become more popular than others by liking, sharing, and discussing, based on social ties and following relations. Ideas, thoughts, and information thus diffuses through such networks.

The authors of [Dow et al., 2013] investigate such behavior on Facebook. They identify only a fraction of posts/photos being exhibited to large scale sharing cascades, while most content remains unnoticed for broader audiences. Other approaches use complex contagion

¹⁵<https://www.digg.com>

models to model information sharing, e.g., [Fink et al., 2016]. Particular focus has also been set on spatial impacts analyzing geographical concentration of topics in e.g., [Yin et al., 2011]. In [Brodersen et al., 2012], the authors do likewise on Youtube, but introduce new metrics that describe the geographical spreading properties. The very same metrics have also been applied to Twitter hashtags in e.g., [Kamath et al., 2013].

Though topic classification or hashtags as a content proxy may be used for tracking information flow through such networks, internet memes depict another example. In [Chen, 2012], the authors analyze meme content on 4chan and its reproduction. [Zannettou et al., 2017] follow the very same idea, but characterize temporal dynamics and inter-arrival times throughout multiple platforms (4chan, Twitter, and Reddit). A follow-up goes more into detail analyzing memes on fringe web communities. They apply image variation clustering to determine their influence. However, [Buntain and Golbeck, 2014] find that single users barely participate in multiple communities.

D.3.2 Problematic Content

Most communication on online platforms remains legit and non-harmful. However, specific regimes attract abusive users. That is, e.g., [Rieger et al., 2021] investigates fringe alt-right communities on 8chan, 4chan, and Reddit. By modeling topics within 70k posts, they identify lots of hate comments. Another qualitative analysis w.r.t. politics on 4chan is provided e.g., in [Ludemann, 2018], whereas e.g., [Parekh et al., 2020] provide data-driven insights into the US presidential election 2016 by analyzing Subreddits about Donald Trump and Hillary Clinton.

Other controversial topics have been analyzed e.g., in [Zelenkauskaite et al., 2021]. The authors analyze hate speech at the example of school shootings and compare user behavior to a control group. Likewise, [Farrell et al., 2019] elaborate on misogyny across Reddit identifying a highly toxic environment. Other far-right topical regimes have been researched e.g., in [Baele et al., 2021] discerning variations across 4chan, 8kun via topical word frequencies and word co-occurrences along about 10k posts. By leveraging negative keywords, [Hosseinmardi et al., 2014] examine the Ask.fm Q&A platform and its specifically abusive users. While such controversial topics also tend to spread misinformation, e.g., [Parekh et al., 2020] investigate how fact checking websites are used as a countermeasure.

Mentioned negative contents usually go hand in hand with exotic or anonymous platforms. However, e.g., Twitter represents a platform where many users tend to use their real identity publicly. In [Zhou et al., 2016], the authors investigate deleted posts. They present user features to determine regrettable contents; they cluster such users by also leveraging sentiment and lexical analysis.

D.4 User Management

As becomes clear not only from previously discussed problematic contents, online platforms are in dire need to implement mechanics that manage and steer user behavior into desired directions. A particular threat is abusive content, but also may be emerging self-reinforcing filter bubbles.

Countermeasures usually rely on filtering or moderation. Due to scaling issues, most nowadays platforms heavily rely on distributed moderation schemes. Such system implement various types of feedback implementations. One of the simplest signals is liking contents as a binary signal for community popularity. This platform input signal may become more fine-grained, e.g., liking and disliking, or one to five stars. Likewise, implementations of displaying popularity to the platform users may vary. E.g., a platform might implement liking and disliking of contents; now they might only show amounts of likes, but refrain from showing the contrary measure of dislikes—thus remove information for the user.

Research has provided various directions to empirically understand such voting mechanisms and user behavior e.g., in [Trujillo and Cresci, 2022, Stoddard, 2015]. It has been shown that the feedback signals of the community often generate community identities and emerging moderation [Fiesler et al., 2018] and social norms [Lampe and Johnston, 2005].

Discussed distributed moderation is only a reactive mechanism, implementing a control loop. Other approaches try to proactively engage its user base setting boundaries (defined platform rules), or reward *positive* behavior. E.g., StackOverflow implements badges that publicly credits its user for certain achievements as shown in [Grant and Betts, 2013]. Doubling down on this approach, [Anderson et al., 2013] discuss optimal badge placement modeling individual user optimization problems. Generally, gamification and social credit create implicit pressure and can influence user *well* behavior [Bosu et al., 2013, Movshovitz-Attias et al., 2013, Kusmierczyk and Gomez-Rodriguez, 2018, Cavusoglu et al., 2015].



USER ADOPTION

In a nutshell. As the success of all technical online platforms highly depend on their user base, we are interested in user **adoption processes** showcasing three distinct applications. We elaborate on the spatially simultaneous and ever-increasing usage of the *Corona-Warn-App* for digital contact tracing for the good at its very anticipated launch at times of a world-wide pandemic. Next, we take a deep dive into *Tripadvisor* and *Google Maps* elaborating on these platforms being used as a side channel to evade censorship in current Russo-Ukrainian war tiems - and discuss moderation by the operator. We further discuss user behavior w.r.t. war-related contents within the Online Social Network *Vkontakte*. We end this chapter with a rich discussion of the growth pattern and user behavior across hundreds of *Jodel* communities within Germany and Saudi Arabia; while the former experiences a gradual organic growth, the latter experienced a kick-start in application usage presumably due to virality on other Social Media.

CONTENT

| | | |
|----------|--|-----------|
| A | Corona-Warn-App Deployment and Engagement | 47 |
| A.1 | Introduction | 47 |
| A.1.1 | Research Questions | 47 |
| A.1.2 | Approach | 48 |
| A.1.3 | Results | 48 |
| A.2 | Related Work | 48 |
| A.3 | Dataset | 49 |
| A.4 | Early Adoption | 50 |
| A.4.1 | Temporal Adoption | 50 |
| A.4.2 | Quick Nation-wide Spread | 50 |
| A.4.3 | Local COVID-19 Outbreaks | 51 |
| A.5 | Conclusions | 51 |
| B | Platforms as a Sidechannel in Wartimes | 53 |
| B.1 | Introduction | 53 |
| B.1.1 | Research Questions | 54 |

| | | | |
|----------|-------|---|-----------|
| | B.1.2 | Approach | 55 |
| | B.1.3 | Results | 55 |
| B.2 | | Related Work | 55 |
| B.3 | | Overview of Russian Censorship | 56 |
| | B.3.1 | Side-channels to bypass censorship | 57 |
| B.4 | | Datasets | 58 |
| B.5 | | Review Activity | 59 |
| B.6 | | Content analysis | 59 |
| | B.6.1 | Labeling War-Related Content | 59 |
| | B.6.2 | Understanding and Measuring Posts | 60 |
| | B.6.3 | Platform Moderation | 62 |
| | B.6.4 | Activity on VK | 63 |
| B.7 | | Conclusions | 64 |
| C | | Growth Patterns in Social Media Platforms | 65 |
| C.1 | | Introduction | 65 |
| | C.1.1 | Research Questions | 66 |
| | C.1.2 | Approach | 66 |
| | C.1.3 | Results | 66 |
| C.2 | | Related Work | 66 |
| C.3 | | The Birth of the Jodel Networks in DE and the KSA | 67 |
| | C.3.1 | Different Adoption Pattern in Germany and the KSA | 67 |
| | C.3.2 | Partitioning the Communities by Volume and Time | 68 |
| C.4 | | Organic Growth Jodel DE | 69 |
| | C.4.1 | Community Interactions Over Time | 69 |
| C.5 | | Adoption of Jodel in the KSA | 70 |
| | C.5.1 | Temporal Adoption | 71 |
| | C.5.2 | Spatial Adoption | 73 |
| C.6 | | Conclusions | 74 |

Introduction

Our first chapter contributes to general **◆ User Adoption** within (Social) Online Platforms. That is, we are interested in onboarding processes to new and existing online services. Depending on the specific question, such analyses are often hard to conduct due to lack of data as most researcher (have to) rely on sampled information, which by itself may introduce biases. While there exists a plethora of different targets to investigate, we focus on different types of user adoption driven by orthogonal essential driving factors. That is, we are interested in examples of *A) user onboarding processes of applications in rich demand of whole societies for the individual and collective good at times of crisis, B) shifts in platform usage from to reallocating its purpose to a side-channel at nowadays hybrid warfare, and in a similar setting to A, C1) and C2) evolution of new social and entertainment platform, Jodel, being hyperlocal 📍 by design and thus being spatially limited to distinct communities, comparing two different country's landscape of local communities.*

Whenever deploying a new application, it is of interested how the app or update is propagated into the (new) user base. [Mcilroy et al., 2016] empirically analyzes optimal strategies for deployment, while [Mcilroy et al., 2016] focus on frequent app updates. While others gathered and investigated information throughout e.g., Google's PlayStore [Viennot et al., 2014], other more technical aspects are of relevance as well, such as the actual distribution of update data, e.g., how to provide such data to millions of users at scale [Singh et al., 2018].

User Adoption digital contact tracing Corona-Warn-App, for the good. In particular, we set out to characterize the early usage patterns of the **★ A) Corona-Warn-App** being used for digital contract tracing of possible COVID-19 infections. While research on this very disruptive event has only started, literature already provides a plethora of interesting and valuable perspectives within the pandemic, amongst *many* others e.g., analyzing Italian mobility patterns [Smolyak et al., 2021], increased Youtube usage [Mejova and Kourtellis, 2021], or general significant shifts observed in internet traffic [Feldmann et al., 2021]. We leverage netflow traffic sample data towards the backend infrastructure to investigate interest on both, a temporal and spatial dimension across Germany. We showcase rich usage right at the startup; local pandemic outbreaks lead to higher country wide usage, presumable due to news media coverage. Research suggests that main forces against app usage is misinformation [Häring et al., 2021], but more importantly serious privacy concerns and trust issues surveyed [Kozyreva et al., 2021, Pape et al., 2021], but also analyzed with data-driven approaches [Dong et al., 2021]. [Kriehn, 2021] suggest that changing the app development to open source has helped to cumbersome at much of mentioned issues.

Evading Censorship by re-purposing well-established platforms to side channels within the Russo-Ukrainian hybrid war. We all are the *sad witness* of a new active war happening in Europe. Hybrid Warfare. We are confronted with large medial misinformation campaigns, fake news, propaganda and restrictions of free speech and information access across the globe; *as if internet had broken its promise.*

In fact, (evading) censorship has been subject to research for decades [Price, 1942, Morgans, 2017, Crandall et al., 2007, Fang, , Bock et al., 2020, Chaabane et al., 2014, Bock et al., 2019]. Russian cyber offenses against Ukraine, NATO and EU countries [Unwala and Ghori, 2016] and internet censorship prior to the war were well-known [Xue et al., 2021, Ramesh et al., 2020, Thomas et al., 2012, Verkamp and Gupta, 2012] including Ukrainian counter blockings [Golovchenko, 2022], and were doomed to be intensified. With the full-scale invasion, both Ukraine [Xynou and Filastò, 2022] and Russia [Meaker, 2022a] extended of governmental-level censorship. To counter hybrid warfare [Haq et al., 2022a], Ukrainian

fact checking projects have been launched [Shuvalova, 2022, Ladygina, 2022]. We find first detailed research on Russian war propaganda Reddit [Hanley et al., 2022], and dataset releases [Fung and Ji, 2022, Shevtsov et al., 2022, Haq et al., 2022b, Park et al., 2022, Zhu et al., 2022, Park et al., 2022].

There are methods for evading censorship, e.g., [Tourani et al., 2015, Fifield et al., 2012]), or Tor implementing anonymity via onion routing [Panchenko et al., 2012, Panchenko et al., 2017], which unfortunately are not yet accessed by the mainstream. As such, we are interested in regular (non-blocked) online platforms at hybrid warfare to evade censorship

★ B) Platforms as Sidechannels . I.e., we use sampled information from Tripadvisor and Google Maps to showcase public information transport w.r.t. the Russo-Ukrainian war via 3rd party applications not being intended for the purpose of general communication, and the reactions to observed (mis-)use in terms of content moderation. Further, we analyze war-related discussed contents on the Online Social Network (OSN) VKontakte.

The very same recipe may yield vastly different results: Characterizing differences in Jodel's user adoption through the German and Saudi community landscape. Though social network analysis is a wide and very active field of research since their early days, very little is known about the early adoption of a new social network. Research has empirically elaborated on growth, development, and state of online social networks [Singer et al., 2014, Kumar et al., 2010, Mislove et al., 2008, Gong et al., 2012, Schiöberg et al., 2012, Wilson et al., 2009, Benevenuto et al., 2009, Jiang et al., 2013]. Factors for successful platforms are studied [Kraut et al., 2012] including the diversity of different actions performed by new users [Karumur et al., 2016] or feedback and its semantic content [Yang et al., 2017]. However, only a few focus on the timeframe and drivers for network growth in particular, while usually relying on sampled information.

Thus, we lastly, engage into our journey of analyzing the Jodel application w.r.t. user onboarding behavior and on-platform interactions

★ C) Growth of Jodel DE and SA on our ground truth dataset. Jodel is a prime example to obtain rich insights due to the forming of *hundreds* of spatially independent communities. We provide a brief overview on how the platform gradually emerged back in 2014 within Germany until late 2017, while portraying a very different adoption behavior within Saudi Arabia that kick-started simultaneously throughout the country (very likely) due to going viral on other Social Media. Therefore, we complement empirical works of the Middle Eastern region [Reyae and Ahmed, 2015] by contributing the first large-scale empirical analysis of the Jodel messaging application in a unique view based on complete ground truth information provided by the network operator itself.








A Corona-Warn-App Deployment and Engagement

On June 16, 2020, Germany launched an open-source digital smartphone contact tracing app ("Corona-Warn-App") to help tracing SARS-CoV-2 (coronavirus) infection chains. It uses a decentralized, privacy-preserving design based on the Exposure Notification APIs in which a centralized server is only used to distribute a list of keys of SARS-CoV-2 infected users that is fetched by the app once per day. Its success, however, depends on its adoption. In this section, we characterize the early adoption of the app using Netflow traces captured directly at its hosting infrastructure. We show that the app generated traffic from all over Germany—already on the first day. We further observe that local COVID-19 outbreaks do not result in noticeable traffic increases.

A.1 Introduction

The Corona-Warn-App [cor, 2020] (CWA) is Germany's official digital contact tracing smartphone app released on June 16, 2020. It aims to trace infection chains by informing users that were exposed to a person later tested positive. Centralized contact *tracking* by apps that report contacts to a central infrastructure raise privacy concerns, which is why a decentralized and privacy-preserving *digital contact tracing* approach (DP-3T) has been proposed [Troncoso et al., 2020]. This concept evolved to the Exposure Notification APIs by Apple [ENA,] and Google [ENG,], of which security and privacy properties were assessed [Baumgärtner et al., 2020]. The CWA uses the decentralized Exposure Notification approach to detect the proximity of other CWA users by collecting pseudonymous identifiers sent via Bluetooth Low Energy, only stored on the phone. Its source code—including the Android and iOS smartphone apps, the backend server, and documentation—is released via Github [Cor, 2020].

Structure [JH6]

-  [A.2: Related Work](#)
-  [A.3: Dataset](#)
-  [A.4: Early Adoption](#)
 -  [A.4.1: Temporal Adoption](#)
 -  [A.4.2: Quick Nation-wide Spread](#)
 -  [A.4.3: Local COVID-19 Outbreaks](#)
-  [A.5: Conclusions](#)

A.1.1 Research Questions

Given the rich public demand for a digital contact tracing application within times of pandemic, we had the rare opportunity to measure the Corona-Warn-App, uncovering the early adoption process. Since widespread adoption is key to the app's success [Ferretti et al., 2020], we question if and how a well-awaited application might find adoption boosts. More specifically, we ask how *interest* in the CWA evolves since day one how local lockdown boost (local) app interest. That is, we analyze traffic patterns and volume dissected by time and the approximated location of origin.

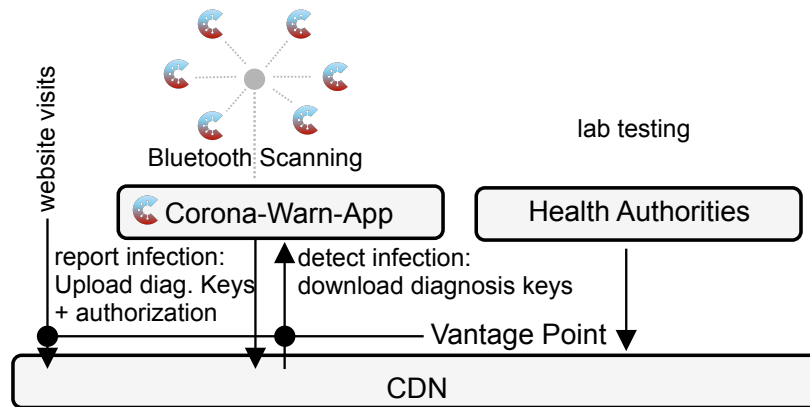


Figure A.1: CWA architecture and vantage point.

A.1.2 Approach

We monitor CWA app and website traffic at its hosting infrastructure as shown in Figure A.1. Phones locally store these received identifiers for 14 days; to protect the user's privacy, all identifiers are volatile by generating new temporary exposure keys every 24 hours. If diagnosed with COVID-19, a user *can* decide to inform others by uploading (parts of her) temporary keys (diagnosis keys) used within 14 days to a central server, verified by health authorities.

By monitoring the API, we observe the first diagnosis keys to be available on June 23 [cwa,]. The CWA regularly downloads shared diagnosis keys from the central server, matches them against the local Bluetooth encounter history, and informs the user of having been exposed to an infected person within the past 14 days if keys match. Shared keys are non-personal identifiable and all contact tracing data never leaves the phone (decentralized approach).

A.1.3 Results

By analyzing traffic data towards the CWA backend infrastructure, we observe an instant usage across all over Germany since the release off the application. Though, with increasing downloads, one might suspect a steady increase in app-usage, after an initial interest peak, the amount of flows declined over the first week. Correlated to new local lockdowns due to Covid-19, the usage again increases significantly. Local pandemic outbreaks further do not seem to affect local app usage in comparison to the whole nation.

A.2 Related Work

The recent pandemic has attracted research in a multitude of different aspects incorporating various methods, views, and countries. We will provide a brief overview about the CWA, while also shortly discussing other pandemic imposed changes to internet infrastructure, or data-driven measurements.

Corona-Warn-App. As we also posit within our contributions here, the success of the digital contact tracing highly depends on user acceptance. Discussed e.g., in [Simon and Rieder, 2021], and according to various survey studies, trust is a major concern in using a tracing application. That is, [Dong et al., 2021] find that about 40% of trust issues are due to misinformation; research generally has determined that lack of application knowledge and application function is a major driver for not using it [Häring et al., 2021]. [Pape et al., 2021] further show positive correlations between education and income to application adoption

in an extensive study with about 4.5k participants. While trust issues are one major concern, [Dong et al., 2021] dissect 70k tweets w.r.t. a negative stance towards the CWA identifying that technical issues and application crashes yield negative app reviews within prominent stores for iOS and likewise Android. [Beierle et al., 2021] confirms that the mood and most content on Twitter is pro-application.

In summary and as shown in [Kozyreva et al., 2021] via another large survey, most pressing concerns towards the application are rooted in security and privacy, possible 3rd party access, and reservations w.r.t. digital contact tracing effectiveness—especially across the non-user group. Nonetheless, application usage is primarily due to protecting the health of oneself, beloved ones, or others. A compact summary and more insights are discussed in [Kriehn, 2021]; yet they also conclude privacy concerns being the major factors for not using the app.

As for development, [Kriehn, 2021] also describe how the CWA development has transitioned into early open sourcing its development—which was particularly helpful in also gaining trust. Further, [Meyer et al., 2021] describe extensive studies on how to optimally parameterize the application.

Mobility and Infrastructure. Without claim of completeness, research has focused many other topics within the pandemic. That is, e.g., [Smolyak et al., 2021] analyzes mobility patterns in Italy within lockdown phases. The leverage data from Facebook forming graph representations of movements; in conclusion, they identify fragmentation and disconnection. Other (non)mobility-patterns especially has lead to a shift towards Home Office and general increase internet usage as shown e.g., on Youtube or Twitter [Mejova and Kourtellis, 2021], while general internet traffic has increased significantly [Feldmann et al., 2021].

In terms of Social Network contents, albeit using a small sample size of 156 posts within Jodel, [Seidenschnur, 2021] identify various active roles of users w.r.t. stance to the pandemic within discussions. Due to much communication and influencing happens online nowadays, Fake News and misinformation is an increasingly important topic. The authors of [Weinzierl et al., 2021] address this issue leveraging Masked Language Model (C.2.2) embeddings combined with additional crafted features put into a Graph Attention network that predict a post’s stance w.r.t. SARS-CoV-2 misinformation.

A.3 Dataset

We obtained sampled Netflow traces from routers connecting the data center hosting the CWA backend (see CDN in Figure A.1). These flows contain web site visits *and* diagnosis key downloads by the app. All client IP addresses are prefix-preserving anonymized. We filter server traffic using 2 IPv4 prefixes mentioned in the CWA backend documentation [Cwa, 2020] and omit IPv6. We verified their usage by resolving the API and web site DNS names (obtained from the app source code) against 10k open DNS resolvers from public-dns.info. As both, app and website, use HTTPS only, we restrict the data to encrypted HTTPS (tcp/443) IPv4 flows from the CDN to the user—resulting in $\approx 3.3M$ matching flows within June 15–25, 2020.

Limitations. Website visits and CWA app API calls are served by the same servers via HTTPS and cannot be differentiated. The routers Netflow cache eviction settings and sampling result in only observing few packets for most flows, making a flow-size based differentiation infeasible. While CWA should periodically download diagnosis keys, energy saving settings prohibit background downloads on some Android and iOS phones, reported on July 24 [DWA, 2020, CWA, 2020] and to be fixed after our study. Periodic request pattern by CWA might thus be used in future work for app identification. Yet, the CWA API DNS name appeared in the Umbrella Top 1M domains [Scheitle et al., 2018] on June 24, 27, July

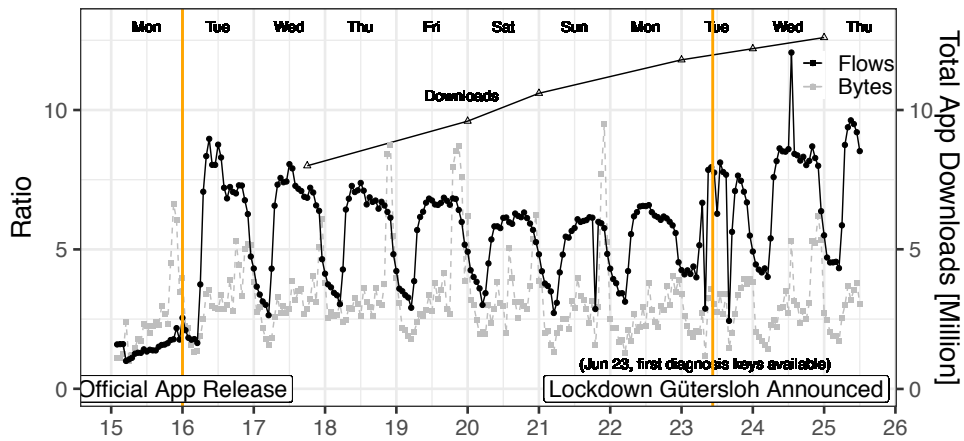


Figure A.2: Aggregated Traffic Observation. Hourly aggregated HTTPS traffic from CWA CDN to users normed to the minimum (left y-axis) and the total app downloads in millions from Google/Apple (right y-axis).

8, 10–11, while the website never appeared—implying *CWA API calls to be more popular than website visits* in OpenDNS and thus *might* dominate the #flows. Flows reveal trends in the interest in CWA and geolocation of destination routers/prefixes enables to study geographic adoption—the scope of this work.

Ethics. The Netflow data provides only flow summaries based on the packet header and does not reveal any payload information. All IP addresses are anonymized; it enables us analyzing aggregates of traffic flows between routers (to identify city-level location information of users) but not individual users. The flow-level statistics do not enable detecting infected users nor deriving *any* user-related information. Our analyses provide aggregated perspectives on the general interest in the app without compromising users’ privacy.

A.4 Early Adoption

A.4.1 Temporal Adoption

We show all HTTPS traffic *from* the CWA CDN to its clients in Figure A.2 (flows and bytes normed to the minimum). It also cumulative shows officially reported downloads from the Apple and Google playstores [App, 2020], starting on June 17; 36 hours after its release, the CWA was downloaded 6.4M times (16.2M total downloads by July 24). With the official release of the CWA on June 16, the traffic immediately increases (7.5x increase of flows on June 16). Interest starts to follow the normal diurnal traffic pattern. After an initial steep traffic increase, it is reduced after a few days, just to re-surge when news in Germany started reporting higher infection rates again and subsequent lockdowns in two districts on June 23 [Deutsche Welle, 2020c] (Gütersloh and Warendorf) followed—widely covered in media.

By knowing that customers of certain ISPs keep the same IP address over time, we studied how regular routing prefixes communicate with the CWA backend (fraction of individual first to last day observed). We observe sustained interest as 50% (75%) of the prefixes occur in 67% (80%) of possible days.

A.4.2 Quick Nation-wide Spread

The success of the CWA app to trace infection chains by contact tracing depends on its adoption and geographic spread. We thus geolocate the request traffic (again both website requests and app API calls—both reflecting interest) within Germany shown in Figure A.3

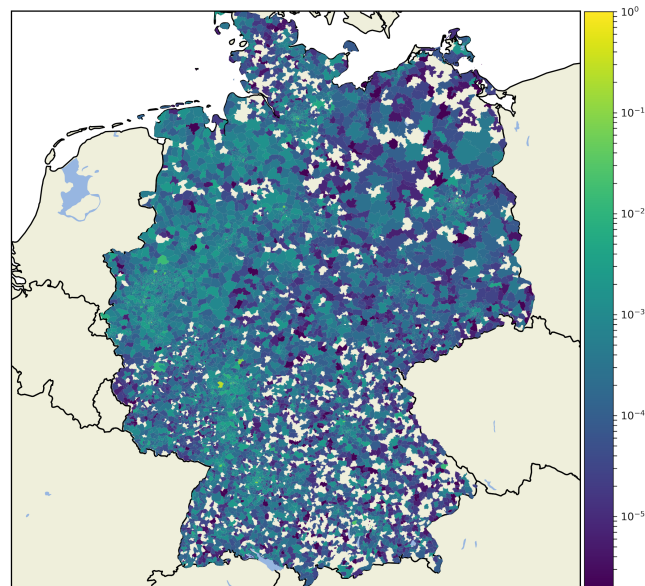


Figure A.3: CWA traffic by district. usage across Germany aggregated over 10 days normalized by max.

by ZIP code areas summed over 10 days. We derive 18% of geolocations from local routers within an ISP that connect customers (ground truth since the router locations are known), while the rest is located by applying the Maxmind geolocation database on routing prefixes. Note that client geolocation *can* be subject to errors; the router city-location can be off the clients location (e.g., in rural areas) and Maxmind’s geolocation can also be subject to inaccuracies at city-level [Poesse et al., 2011]. We observe that almost all districts (shown in the heatmap by ZIP code areas) emit requests to the CWA backend. Notably, evaluating the geographic spread on the first day of the app leads to almost the same observation (not shown). In conclusion, the CWA triggered interest across a almost *all* German districts.

A.4.3 Local COVID-19 Outbreaks

Our measurement period contains two local COVID-19 outbreaks: *i*) in Berlin on June 18 [Deutsche Welle, 2020a], and *ii*) in Gütersloh and Warendorf on June 23 [Deutsche Welle, 2020c]. The latter (June 23) led to few domestic travel restrictions for visitors from these districts [Deutsche Welle, 2020b]. While we observe an increase in usage starting on June 23 (see Figure A.2), this traffic increase also occurs on federal state level simultaneously—not only in the federal state (NRW) being home to the affected districts. In Gütersloh, the traffic increased only very slightly and hardly noticeable (insufficient data for Warendorf). The outbreak in Berlin on June 18 is only visible for users of a single ISP and not in the overall traffic from Berlin-based users. For now, we thus conclude that local COVID-19 outbreaks do not appear to generally increase traffic in only the affected regions. Instead, nation-wide news reports on outbreaks *might* contribute to growing app interest across Germany—an effect worthwhile to investigate in future work.

A.5 Conclusions

In highly demanded public anticipation *for the good*, already on its first day, the CWA app generated substantial interest—manifested in traffic from almost all German districts. Local COVID-19 outbreaks do not appear to increase traffic in the affected regions but can correlate to nation-wide increases. Future work should aim for a better understanding of patterns that drive local adoption, e.g., to which degree app usage is stimulated by news

reports—if and how does news media fire CWA interest; and what will be the long-term app interest.

Given this first example, we observe that users can adopt quite quickly to new applications, especially if it promises fulfilling dire needs.

B Platforms as a Sidechannel in Wartimes

During the first days of the 2022 Russian invasion of Ukraine, Russia's media regulator blocked access to many global social media platforms and news sites, including Twitter, Facebook, and the BBC. To bypass the strict media control set by Russian authorities, pro-Ukrainian groups explored unconventional ways to reach out to the Russian population such as posting war-related content in the user reviews of Russian business available on Google Maps or Tripadvisor. In section, we provide a first analysis of this new phenomenon by analyzing the creative and effective strategies to avoid state censorship. Specifically, we analyze reviews posted on these platforms from the beginning of the conflict to mid May 2022. We measure the channeling of war messages through user reviews as well as in VK, a popular Russian social network. Finally, we analyze the response of platforms in terms of content moderation.

B.1 Introduction

While early platform adoption processes are specifically interesting due to the existing gap in research, adoption processes and usage changes happen throughout any platform's lifecycle. That is, even well-established services may encounter e.g., user churn and shifts in application usage. Such changes primarily root in contents being appreciated, discussed, and shared. Research has shown that platforms struggle mainly with harmful contents e.g., cyber-bullying, or toxicity, misinformation, and emerging self-reinforcing filter bubbles. Likewise, detecting such adverse circumstances and moderation processes in place has been subject to various investigations.

However, in this section, we would like to focus on an unusual happening: With the escalation of the Russo-Ukrainian war in February 2022, internet users have become creative evading Russian internet censorship by re-purposing still available review platforms to side channels informing Russian citizens with information about active warfare.

In 2014, after the revolution of dignity (also known as the Maidan revolution), Russia invaded and subsequently annexed the Crimean peninsula from Ukraine¹. In the same year, Russia, supported by pro-Russian separatist forces started intervention in the Donbas region, violating Ukraine's sovereignty. Russia employed a hybrid approach, deploying a combination of disinformation, irregular mercenary soldiers, regular Russian troops as well as military support to destabilize the region. Despite massive troops deployment in 2021 at the border to Ukraine from its own territory as well as Belarus, Russia deceitfully denied any further attack plans despite serious concerns of the world leaders. On February 24, 2022, Russia launched a wide-ranging attack on Ukraine from multiple fronts that involved airstrikes and missiles across whole Ukrainian territory. This invasion caused Europe's largest refugee crisis since World War II. Massive civil protests against the invasion took place all over the world. In response to the Russian military aggression, most countries imposed economic sanctions against Russia and many provided both humanitarian and military aid to Ukraine.

At this very turning point, Russian authorities implemented (more) measures to control information and promote its own version of the events. As a response, many global social media platforms and other prominent online services were used to raise awareness of war issues and situation happening. Russia's media regulator Roskomnadzor blocked, from very early, the access to Facebook, Twitter [Meaker, 2022a], and many global news sites covering the war. This was confirmed by an OONI report dated March 7, 2022, which provides evidence of Internet censorship by Russian Internet Service Providers (ISPs) based on their network measurement data [Xynou and Filastò, 2022]. In fact, OONI public data reveals a noticeable increase of Internet censorship in Russia from the beginning of the 2022









¹https://en.wikipedia.org/wiki/Russo-Ukrainian_War

conflict. This move was seen both as an attempt to stop the dissemination within Russia of any information not provided by official sources, and also as retaliation for the removal of Twitter and Facebook accounts—allegedly belonging to two pro-Russian disinformation groups [Silverman and Kao, 2022, Collins and Kent, 2022]—and EU bans on Russia’s outlets Russia Today and Sputnik [of the EU, 2022].

Information controls are frequent in time of war, and so are evasive manoeuvres to bypass them. Russia’s censorship efforts were answered with some inventive proposals. On February 28, an account presumably affiliated with Anonymous suggested to employ user reviews in restaurants and other business located via Google Maps to deliver war-related information to the Russian population [@YourAnonNews, 2022a]. Tinder, Tripadvisor and Telegram were also targeted as means to reaching out to the Russian population, thus bypassing the strict media control implemented by Russian authorities [Meaker, 2022b]. On March 4, the squad303 group offered the possibility to target millions of Russian citizens with SMS via the 1920.in site. This service was later improved to also send emails and Whatsapp or Viber messages. Some prominent online service providers responded to these campaigns by actively removing war-related content from their platforms. Google and Tripadvisor placed restrictions on reviews of Russian business, and Google Maps soon stopped accepting new reviews for places located in Russia, Ukraine, and Belarus. They argued that such reviews violate company policies [Hamilton, 2022, Kaufer, 2022, (WSJ), 2022].

The creative use of online services as side-channels to bypass state information controls was anecdotally echoed by the media. Yet, there is no quantitative assessment of the user involvement, effectiveness, and intensity of these campaigns, nor the response by platform operators to moderate content.

Structure [JH1]

-  [B.2: Related Work](#)
-  [B.3: Overview of Russian Censorship](#)
-  [B.4: Datasets](#)
-  [B.5: Review Activity](#)
-  [B.6: Content analysis](#)
 -  [B.6.1: Labeling War-Related Content](#)
 -  [B.6.2: Understanding and Measuring Posts](#)
 -  [B.6.3: Platform Moderation](#)
 -  [B.6.4: Activity on VK](#)
-  [B.7: Conclusions](#)

B.1.1 Research Questions

The creative use of online services as side-channels to bypass state information controls was anecdotally echoed by the media. Yet, there is no quantitative assessment of the user involvement, effectiveness, and intensity of these campaigns, nor the response by platform operators to moderate content. In this section, we fill this gap by measuring and analyzing several datasets purposely collected from Google Maps, Tripadvisor, and VK from February through April, 2022, to showcase out-of-band communication and the general prominence of war-related contents in Russian social networks.

B.1.2 Approach

To address these questions, we study changes in the volume of reviews and moderated content on Google Maps and Tripadvisor, while also monitoring public discourse on the Russian social network VKontakte. We next leverage text-based analysis techniques to label reviews as related to war or not and provide a rich empirical overview of the dataset contents. Finally, we study the reaction of service operators to control or remove war-related information.

This section offers a first, evidence-based analysis of how user reviews on global online services have been used as side-channels to bypass information controls in wartime, as well as operators' efforts to moderate such content.

B.1.3 Results

In this section, we showcase how Tripadvisor and Google Maps were used to bypass state censorship. Specifically, we study how these services were used to deliver information related to the Russian invasion of Ukraine. We identify patterns and campaigns using mentioned platforms as side channels for war-related contents, which has later been shut down by platforms providers implementing content moderation policies and executing (bulk) removals.

Further, we identify and analyze war-related content on a meta-level within the Russian Online Social Network VKontakte, which unsurprisingly showcases that the Russo-Ukrainian war is a lively topic.

Overall, our study reveals how unblocked online platforms were used to circumvent state-wide information controls.

① This work is a result of common efforts between all referenced authors. Major parts of the empirical data analysis, NLP, and crowdsourcing campaign design, setup and evaluation were conducted by me.

B.2 Related Work

Governmental censorship during times of military conflicts has been subject of attention and analysis for a long time [Price, 1942, Morgans, 2017]. The 2014 Russia-Ukraine conflict offered a valuable case study of Russia's information war strategy. Russia started offensive cyber operations against Ukraine not later than 2009 as a part of a broader war campaigns against NATO and EU countries [Unwala and Ghori, 2016]. In 2014, information war operations intensified against Ukraine. While initially aiming at spreading of misinformation and propaganda, later phases of the information war included blocking free access to the Internet. To fight Russian propaganda, Ukraine introduced blocking of several Russian websites in 2017 [Golovchenko, 2022]. With a start of a full-scale attack on Ukraine, both Ukraine and Russia deployed massive extension of governmental-level censorship.

Censorship. The research community has studied in depth the use of censorship techniques. Several recent works have inspected how the Russian government censors the access to different types of services. [Xue et al., 2021] studies the use of throttling on Twitter, in order to pressure the company to apply more restrictive content removal policies.

Misinformation and Propaganda. Analysis of disinformation spreading in the time of war has been a subject of attention not only in the scientific community. Affected by a hybrid war [Haq et al., 2022a], Ukraine has launched several fact checking projects. The texty (Ukrainian for *texts*) project originally emerged to analyze the proliferation of "junk

websites” in the Ukrainian news ecosystem. “News” distributed by such media is often of low quality or even complete fake. Still, in 2018 they were shown to reach over 50 million of visitors per day (for comparison, a popular news media in Ukraine reaches around 15 million of visitors). The up-to-date sample includes *i*) Ukrainian clickbait sites, *ii*) russian sites aimed at Ukraine, *iii*) mainstream russian news sites, and *iv*) mainstream Ukrainian news sites. The data is loaded from RSS-feeds and the analysis is focused on news only on politics, economy, society, and foreign affairs (other categories are excluded by a separate classifier). The proposed/trained classifiers allow for topic detection and estimation of likelihood that an article contains false argumentation or/and emotional manipulation.

Other research focuses on different countries like China [Fang,], or elaborate on the portraiture of Russia within Ukrainian media [Shuvalova, 2022, Ladygina, 2022]. There have been many efforts for also creating, curating datasets for enabling further user behavior analysis on e.g., Weibo citefung2022weibo, Twitter [Shevtsov et al., 2022, Haq et al., 2022b, Park et al., 2022], Reddit [Zhu et al., 2022], and VK [Park et al., 2022].

Though the recent War has only escalated recently (in terms of scientific publication speeds), we find first detailed research on propaganda on Reddit [Hanley et al., 2022]. They conduct research on the infrastructure of Russian disinformation websites, determine topics, and show their success in distributing narratives. A thorough evaluation of various topic-related subreddits shows that e.g., about 40% of `r/Russia`’s content follows disinformation narratives according to determined topics; whereas `r/politics` only experiences a fraction of about 9%.

B.3 Overview of Russian Censorship

Governmental censorship during times of military conflicts has been subject of attention and analysis for a long time [Price, 1942, Morgans, 2017]. The 2014 Russia-Ukraine conflict offered a valuable case study of Russia’s information war strategy. Russia started offensive cyber operations against Ukraine not later than 2009 as a part of a broader war campaign against NATO and EU countries [Unwala and Ghori, 2016]. In 2014, information war operations intensified against Ukraine [Volkova and Bell, 2016]. While initially aiming at spreading misinformation and propaganda, with a start of a full-scale attack on Ukraine, Russian authorities boosted media control by blocking free access to the Internet. This was done in fear of civil protests against the war, preceded by repression and mass arrests of its own population on March 4 2022 [(BBC), 2022]. Words such as *war* and *invasion* were officially banned in Russia’s media. While western news sites and social media networks such as Facebook were blocked, popular Russian social networks like VKontakte (VK) and western ones such as Instagram, Twitter and YouTube remained accessible to limit collateral damage. However, some of them were gradually blocked in Russian ISPs as the conflict evolved [Troianovski, 2022].

This section provides an overview of the breadth and depth of Russia’s Internet Censorship. To this end, we leverage OONI’s web connectivity public data [of Network Interference, 2022]. This data is build on performing web site retrievals from from a control and a test network to identify blocking activities. We focus on a selection of representative web sites in different categories, as shown in Table B.1. To better understand the enforcement of Internet censorship, we manually checked (by a Ukrainian co-author) the list of domains probed by OONI in Russia [Lab and Others, 2014] for relevant Ukrainian news sites. We note that the current OONI probe list misses major news sites from the Ukraine (e.g., korrespondent.net, nv.ua, pravda.com.ua, or rus.lb.ua) and news agencies (e.g., interfax.com.ua, ukrainian.info), which we advocate to be added. Moreover, not all listed sites do appear to be operational news sites anymore (e.g., provokator.com.ua), which we omit.

We observe a large set of sites being blocked in Russia before 2022 as shown in Table B.1.

| category | Website | Start of censorship |
|-----------------------------|-----------------|---------------------|
| Social media | twitter.com | 2022-02-26 |
| | facebook.com | 2022-03-04 |
| | vk.com | not blocked |
| | youtube.com | not blocked |
| International media outlets | dw.com | 2022-03-04 |
| | bbc.com | 2022-03-04 |
| | nbc.com | not blocked |
| | nytimes.com | not blocked |
| | theguardian.com | not blocked |
| Independent media outlets | interfax.ru | 2022-02-26 |
| | currenttime.tv | 2022-02-28 |
| | tvrain.ru | 2022-03-02 |
| Ukrainian media outlets | glavcom.ua | before 2022 |
| | glavnoe.ua | before 2022 |
| | maidan.org.ua | before 2022 |
| | qha.com.ua | before 2022 |
| | hromadske.ua | 2022-02-08 |
| | 24tv.ua | 2022-03-02 |
| | atr.ua | 2022-03-05 |
| | 1plus1.ua | 2022-03-09 |
| | 5.ua | 2022-03-18 |
| | nr2.com.ua | not blocked |

Table B.1: Website censorship in Russia detected by OONI

However, major social media platform Facebook and several news sites, both international and regional, were blocked right after the beginning of the war. Examples include the BBC and DW (a German broadcaster that is available in Russian). Beyond, also independent news channels based in Russia (e.g., tvrain.ru or currenttime.tv) were blocked. As major news and social media websites became inaccessible within Russia, it is plausible that many others news sites for which OONI has no evidence were also blocked. Such intensification of information controls by Russian authorities motivated the creative and novel use of non-blocked side-channels such as reviews in Google Maps and Tripadvisor. As a result, these websites soon become a niche for spreading information about the war. We note that both services were (and still are, at the time of this writing in mid May'22) available to Russian citizens.

B.3.1 Side-channels to bypass censorship

Internet censorship was known to be a practice enforced by Russian authorities prior to the war [Xue et al., 2021, Ramesh et al., 2020, Thomas et al., 2012, Verkamp and Gupta, 2012]. Yet, we observe a noticeable increase of censorship in Russia after the beginning of the 2022 conflict. As a result, major news and social media websites became inaccessible within Russia.

While there exists methods to evade censorship [Fifield et al., 2012], many approaches are not necessarily available to the mainstream. This finding motivates us to study of non-blocked side-channels to bypass censorship. To evade Russian's state control over the messages distributed on major social platforms and news sites, cyber activists proposed using alternative channels to reach out and deliver anti-war and anti-Putin messages to Russian citizens by posting messages on unblocked websites like Google Maps and Tripadvisor. Soon, Alphabet and Tripadvisor started to moderate war-related content published on their services [Kaufer, 2022, (WSJ), 2022]. Activists posted anti-war stickers around neighborhoods and even wrote anti-war messages on banknotes in order to avoid state control [Silina, 2022].

| Dataset | Crawling period | Data period | Size |
|-------------|--------------------------------|--------------------------------|--------------------|
| Google Maps | Mar 4, 2022 - Apr 30, 2022 | Jan 1, 2022 - Apr 30, 2022 | 203,118 reviews |
| Tripadvisor | Mar 12, 2022 - Apr 30, 2022 | May 12, 2021 - Apr 30, 2022 | 5,582 posts |
| VK | Mar 18, 2022 - Apr 30, 2022 | Mar 18, 2022 - Apr 30, 2022 | 1,851,986 posts |

Table B.2: Datasets

B.4 Datasets

We created custom crawlers and collected this information from Google Maps and TripAdvisor. Also, we decided to monitor the activities in VK, the largest Russian social network, to better understand potential social activities around the war and compare trends. This section describes our crawling efforts and provides statistics for the three datasets collected.

Tripadvisor. On March 2, 2022, TripAdvisor’s CEO allowed the use of Ukraine forums to “enable users to share information” about the situation in the country [Kaufer, 2022]. Indeed, a message in certain Russian places showed a message from TripAdvisor staff, indicating that reviews were disabled due to high volume of war-related content, and that users should use the forums to inform about available travel options within Ukraine. Since then, the travel forums for Russia and Ukraine have become a platform for discussion about the war situation. We crawl these forums for two months (from March 12, 2022 to May 12, 2022), harvesting all posts published since the beginning of the war. As some posts are removed by TripAdvisor due to infringement of its ToS, every hour the crawler checks for new posts and collects any new content, thus allowing us to flag removals as well. We also conduct one single crawl to obtain pre-war posts that are one year old, i.e., since May 12, 2021. The dataset contains 5,753 posts made in 1,229 different threads by 1,080 different users.

Google Maps. The Google Maps dataset contains over 200k reviews obtained from almost 50k places located in Russia. We started crawling these reviews on March 4, 2022. We fetch new reviews every 2 hours and update the list of places daily. To discover potential places to crawl, we use a purpose-built Chromium-based instrumented browser that makes use of the “Nearby” search functionality from Google Maps. This functionality lists any places (e.g., hotels, restaurants, museums) found in a given town or its vicinity. We iterate over a set of 321 predefined Russian towns² from where to discover places. In the end, combining these two methods we covered 8,237 different towns. There are 144,706 unique users with at least one posted review in the dataset.

VK. The VK dataset consists of 1.9M posts appearing on the top-50 public communities with the most followers at a given date. It contains publications and replies from 51 different communities, published by 554.5k different users. We started crawling this social network on March 18, 2022. Given the intensity of these communities, we follow a best-effort approach to fetch new posts every 15 minutes, updating the list of communities daily. We acknowledge that, in some posts with high activity, we may miss some replies as VK limits the number of most recent messages that can be accessed.

²https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_Russia_by_population.

B.5 Review Activity

In this section, we study the changes in the number of published and deleted posts per day in Google Maps and Tripadvisor. Due to the best-effort strategy of our VK crawler and the inability to estimate the entire lifespan of every message, we exclude VK messages from this analysis. For new posts, the publication date is directly obtained from the sites. For removals, we are able to precisely guess removal dates for those messages observed during our crawling period. In the case of Google Maps, we calculate them based on the latest time our crawler sees a given review. This yields good results as we daily crawl all reviews (new and previously existing ones) for all monitored places. In the case of Tripadvisor, a removed post is replaced by a placeholder message from the admins indicating the reason for removal. However, the timestamp of this replacement is not provided. Thus, as in the case of, Google Maps, we estimate the removal date as the first timestamp where we observe that a message was replaced by the placeholder message. We conduct hourly crawls, and thus we can estimate the removal date with one-hour precision.

Figure B.1 shows the number of posts that were published and deleted per day in both Google Maps and Tripadvisor, respectively. We refer the reader to [B.6: Content analysis](#) for a detailed analysis of these posts and removals based on their discussion topic. For the two platforms, we see a clear change in volume right after the beginning of the war on 24 February (vertical line in Figure B.1), although they are manifested differently across platforms in terms of intensity. In the case of Google Maps, there is a clear drop in the daily amount of published reviews, suggesting some kind of content moderation, which is consistent to Google policy on not accepting new reviews for places located in Russia, Ukraine, and Belarus. For Tripadvisor, instead, we see a slight increase in the number of posts. As we show in [B.6.2: Understanding and Measuring Posts](#), this is due to a larger volume of war-related posts in these travel forums, which was indeed allowed by forum administrators.

Regarding removals, we observe that the volume and also the frequency of removals is higher for Google Maps than Tripadvisor, as opposed to the new entries where we observe similar patterns.³ This suggests that both sites implement different content moderation policies during war-times. We discuss platform moderation and reasons for removals, when available, in [B.6.3: Platform Moderation](#).

Takeaway. *We observe substantial changes in the number of daily reviews and removals both in Tripadvisor and Google Maps since the beginning of the war. These changes are correlated in time with the blocking of some major social platforms and news sites in Russia, and with a call for using these platforms as side-channels to reach out to the Russian population.*

B.6 Content analysis

To gain further insights about the nature of the posts and removals, we set out to investigate the purpose of the actual posted content. That is, we aim to quantify the amount of posts related to the conflict and whether service providers actively try to moderate the content. We rely on a native speaker to address the language challenges.

B.6.1 Labeling War-Related Content

An initial deep qualitative assessment on GM and TA reveals substantial amounts of war-related posts specifically trying to inform Russians about the war. However, such time-consuming manual labelling process does not scale. To overcome this limitation, we employed unsupervised topic mining techniques [Grootendorst, 2022] (masked LM embed-

³The lack of removals in early May for Tripadvisor is a consequence of an error in the crawler.

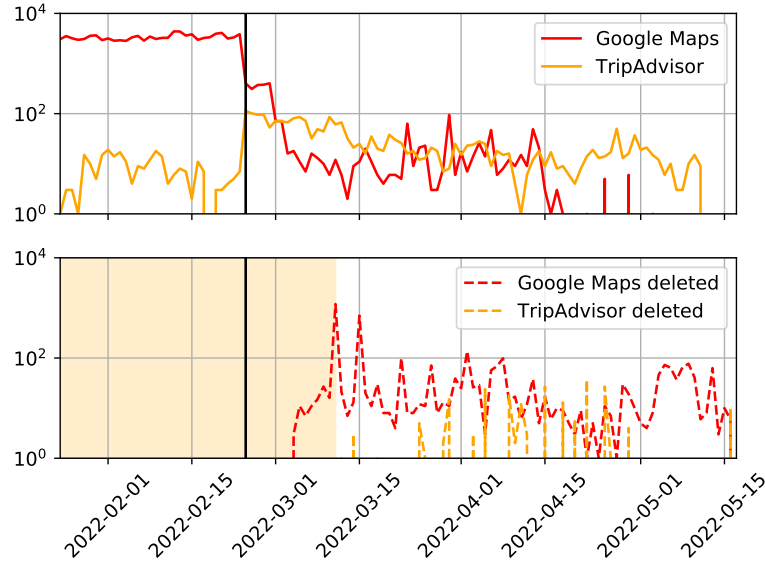


Figure B.1: Number of daily posts. Published (top) and removed (bottom) in Google Maps and Tripadvisor.

ding, dimension reduction, clustering), which indeed identified topics related to war. However, results were very specific and contained huge amounts of false positives (FP).

Thus, we opt for a simpler approach assembling a set of keywords (cf. Table B.4) from our qualitative and data-driven evaluation, aiming at reducing the amount of FP—*better be safe than sorry*. Reusing our qualitative assessment results, we define a threshold of at least three matching keywords to label posts being war-related; a trade-off in favor of precision. We measure quality via a set of lightweight crowdsourced campaigns. Due to predominantly Russian language on both Google Maps and VK posts, we employ *i*) non-native speaking coders using machine translation, and verify results with a second labelling pass from *ii*) native-speaking expert coders. The campaigns were set up for each dataset carefully sampling sets of each 25 posts identified as non-/war related for the time before the war and within war. Furthermore, we focus on non-/deletions—totalling in 750 labels of which 94.7% were consistent between non-/expert coders.

Accuracy Evaluation. Our keyword-matching approach overall works surprisingly well at a precision of 0.94, while the F1 score is 0.85; we observe increased figures in false negatives (FN) as expected. We show detailed classification results in Table B.3.

| Platform | TP | TN | FP | FN | Accuracy | Precision | Recall | F1 |
|----------|-----|-----|----|----|----------|-----------|--------|------|
| VK | 99 | 125 | 1 | 25 | 0,90 | 0,99 | 0,80 | 0,88 |
| GM | 99 | 116 | 1 | 34 | 0,86 | 0,99 | 0,74 | 0,85 |
| TA | 92 | 117 | 16 | 25 | 0,85 | 0,85 | 0,79 | 0,82 |
| Σ | 290 | 358 | 18 | 84 | 0,86 | 0,94 | 0,78 | 0,85 |

Table B.3: Crowdsourcing Classification Results.

B.6.2 Understanding and Measuring Posts

Figure B.2 shows the number of daily reviews and posts published to Google Maps and Tripadvisor. It shows the general trend as well as a category for those messages and posts labeled as war-related. As expected, all war-related documents—with some minor exceptions—are published after the beginning of the war. Between March 15 2022 and the end of April

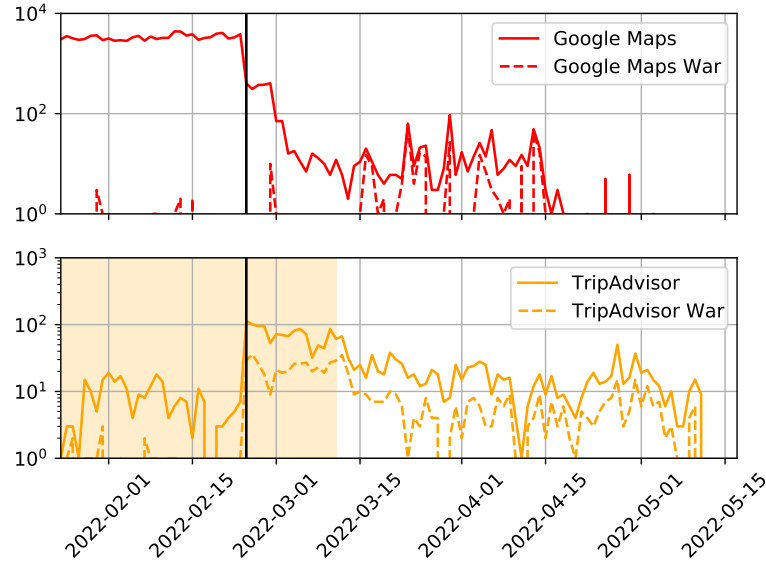


Figure B.2: Number of published "war related" daily posts. Google Maps (top) and Tripadvisor (bottom).

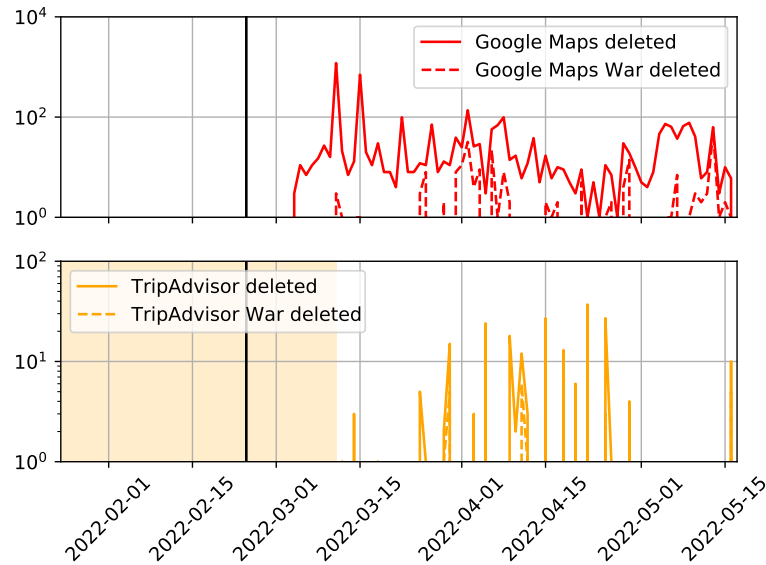


Figure B.3: Number of removed "war related" daily posts. Google Maps (top) and Tripadvisor (bottom).

2022, 46% of reviews published in Google Maps are war-related. Tripadvisor follow a similar ratio, with 31% of posts talking about the war. On Google Maps, we find higher war-related activity between the 23rd and 25th of March. We attribute this peak to the blocking of Google News inside Russia [[@YourAnonNews, 2022b](#)] and to the publication of contact details of Russian companies [[@NIRVANA101, 2022](#)]. We find a second peak between the 13th and 14th of April. In the case of Tripadvisor, while there are dates with more war content than others, we do not find any period concentrating significantly higher activity.

Content analysis. Table B.4 shows the most frequent keywords found in the messages posted in each platform and their frequencies. The list suggests that the posted content is clearly related to the war and its consequences.

| TA | | GM | | VK | |
|------------|-------|-------------|-----|-------------|--------|
| russia | 1.056 | russia | 248 | (putin) | 24.933 |
| putin | 506 | war | 202 | (ukraine) | 22.331 |
| ukraine | 444 | ukraine | 196 | (west) | 21.227 |
| war | 374 | murder | 141 | (azov) | 14.701 |
| ukrainian | 211 | kill | 130 | (donbass) | 12.042 |
| west | 204 | putin | 124 | (war) | 11.122 |
| government | 194 | ukrainian | 118 | (sanctions) | 8.373 |
| invasion | 130 | soldier | 107 | (europe) | 6.904 |
| kill | 117 | truth | 106 | (mariupol) | 6.273 |
| sanction | 113 | government | 102 | (fascism) | 5.860 |
| civilian | 105 | peace | 96 | (hands) | 5.404 |
| sanctions | 101 | (ukraine) | 91 | (nazis) | 5.065 |
| truth | 100 | kill | 90 | (kill) | 4.111 |
| kyiv | 97 | brutal | 81 | (bombs) | 3.962 |
| force | 88 | (hands) | 63 | (blood) | 2.611 |
| border | 85 | (murderers) | 61 | (traitor) | 2455 |
| military | 83 | president | 60 | (Bucha) | 1581 |
| civilians | 82 | thousands | 58 | (bandera) | 1255 |

Table B.4: "war related" Frequent Keywords. Excerpt of most frequent keywords since the beginning of the war (terms in brackets are translated words from Russian).

Duplicated content. Given the public calls for informing Russian citizens about ongoing war events, we analyze if the posted messages contain organized campaigns (i.e., spamming the same text) in addition to personal or individual statements. We look for identical contents related to war (according to our keywords) on all three platforms. We identify several identical posts on Tripadvisor and Google Maps written in English and informing Russian citizens about the war pledging for an end. For Google Maps, we identify 8 posts each being replayed more than 10 times, totalling 188 instances. The same holds for Tripadvisor, though the amount in this case is smaller: 7 posts posted 17 times. In the case of VK, dominated by Russian users, we also find 28 war-related posts replayed 399 times. A qualitative look into the contents provides a wide spectrum of pro-Russian and pro-Ukrainian messages, whereas others primarily pledge to stop the war.

B.6.3 Platform Moderation

Platform operators reacted differently to the use of their services as channels for disseminating war-related information. We next discuss content moderation—or lack thereof—in the platforms analyzed in this study. For reference, Figure B.3 shows the number of war and non-war deleted posts for both Google Maps and Tripadvisor. We remind the reader that measuring VK post removals is not feasible due to the aforementioned crawler limitations.

Tripadvisor. In the case of Tripadvisor, there is evidence of various posts (and entire threads) being removed by forum administrators. Table B.6 shows the number of posts and the reason provided in the placeholder message being left by forum administrators. We focus on messages removed after the war started. Note that up to 121 posts were removed at the author's request according to the metadata offered by the platform. Hate speech and harassment are the two more prevalent categories when posts are removed by the platform. For the 34 threads (215 posts) that have been removed completely, we ignore the reasons for such removals. Out of these, 13 (38%) have a unique post, and 4 (12%) only contain one reply. Meanwhile, 7 threads (20%) have more than 13 replies. We confirm through manual inspection that the reasons for thread removal typically fall into two categories. First, the conversation in the threads completely deviate from its original purpose (e.g., provide objective information about the conflict) towards political discussions or even hate speech. Second, the thread is initiated with the sole purpose of propaganda or another advertise-

| | Threads | Posts | All |
|--------|----------------|------------|-----------|
| Total | 31 | 38 | 69 |
| Mean | 7d 23h 15m | 2d 16h 16m | 5d 1h 19m |
| Median | 2d 16h 44m | 1d 14h 32m | 2d 8h 28m |
| Stdev | 12d 8h 42m 40s | 2d 14h 0m | 8d 20h 5m |

Table B.5: Lifespan of content removed in Tripadvisor.

| Reason | #Posts |
|---------------------------------------|------------|
| Entire thread was removed | 215 |
| Off-topic chat | 147 |
| Removed by author | 121 |
| Harassment to other users | 102 |
| Hate speech or inappropriate language | 89 |
| Self-promotional advertising | 31 |
| Not written in English | 14 |
| Copyright infringement | 6 |
| Multi-account detected | 2 |
| Total | 727 |

Table B.6: Reasons for content removals on Tripadvisor.

ment, and it is removed quickly, sometimes even before it gets any reply. Table B.5 shows the lifespan of the 69 items (i.e., a post or the entire thread) removed by administrators since we started our periodic crawls (March 12, 2022). Posts in general are removed faster than entire threads, but we also observe that those threads without replies (i.e., containing only the OP message) are removed as quickly as the posts. This confirms that active platform moderation is in place and that much of the content removal is linked to the war.

Google Maps. We find evidence of platform moderation for Google Maps, as war-related reviews have a much shorter lifespan than the rest, typically lasting less than 50 days as opposed to those until the end of our crawling. Shortly after the beginning of the war, Google Maps temporarily suspended posting new reviews on Russian places in an attempt to prevent the generation of content that violates company policies [(WS), 2022]. This led to the drastic reduction on the number of daily published reviews, shown in Figure B.1. While war content is not explicitly prohibited in Google Maps, Alphabet alleged that these reviews were considered “off-topic,” a category that is prohibited in Google Maps, justifying their temporal suspension [Help, 2022]. From our data, this temporary banning is still active in April 2022 (only 8 posted reviews per day). Nevertheless, we find 18 war-related reviews that somehow bypassed Google Maps’ moderation.

B.6.4 Activity on VK

While we have presented insights to platforms being used as a side-channel, we are next interested in public discussions in relation to the war. Figure B.4 presents the amount of (conservatively) identified war-related posts within the VK dataset over time. Note that the shaded region has not been actively crawled, but based on data that persisted until starting crawling, it is represented with the steep increase of observed posts at beginning of March. Total observed posts remain at a consistent level (averaging at 38k posts/day), of which war-related posts account for 100 posts/day according to our conservative keyword approach—naturally indicating ongoing public discourse. From a qualitative perspective, most contents are largely in line with Russian propaganda; we also find evidence of more critical opinions—or likewise posts trying to inform the Russian population about the war. Noteworthy, we find banned words under threat of fine in 1,558 instances (832 assault, 668 invasion, 58 declaration of war), whereas *special military operation* appears just 65 times.

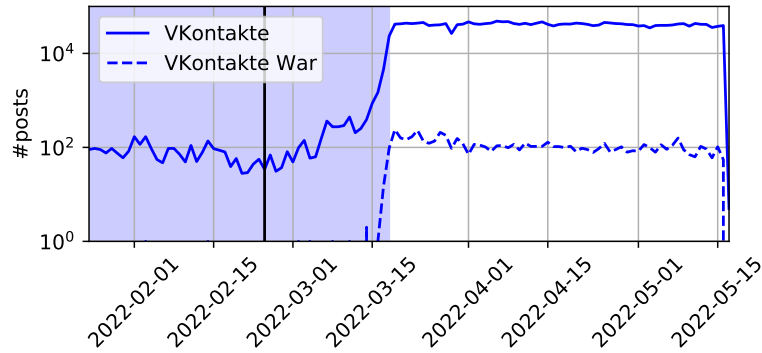


Figure B.4: VK volume of war and non-war posts.

Popularity. As in any Online Social Network (OSN), also VK is driven by user appreciation through views, likes and shares. For all three metrics, we observe a heavy-tailed distribution common within OSN. Comparing war with non-war related threads (initial posts), we observe equal amounts in views on average, while likes and shares fall short for war-related content—also within the upper the 90%-99% percentiles. The same holds true for replies, while the war-related content appears less skewed in the upper percentiles. Focusing on thread lengths, we observe a huge shift towards longer discussions for war-related content (averages 57 vs. 78 replies, similar upper quantile values). We conclude that non-/war content receives equal visibility, whereas the war-related content is discussed by users more extensively.

B.7 Conclusions

This section explores for the first time how users and activists creatively leveraged Internet services such as Tripadvisor and Google Maps to bypass state censorship. Specifically, we study how these services were used to deliver information related to the Russian invasion of Ukraine. Using a dataset collected during the first weeks of the war, we observe the shifts in the pattern of daily user post volume and removals, and also duplicated content suggesting intentional and organized campaigns to disseminate such information. Our automatic content analysis confirms that there is indeed a peak in war-related narratives in the reviews in these platforms. For comparison, we observe a similar pattern in the Russian social network VK.

The use of place and business reviews as side-channels forced platforms to implement content moderation policies. In the case of Tripadvisor, our analyses suggest that administrators do perform intensive content moderation. However, they allow (and indeed encourage users) to discuss and inform about the conflict, mostly to provide information about safe traveling in and out Russia and Ukraine. In the case of Google Maps, there is also evidence of bulk removals leading to a temporary suspension of the reviewing activity that extends up to the time of this writing.

Overall, our study reveals how three non-censored online platforms were exploited to circumvent state-wide information controls. However, its contributions go beyond the analysis of an interesting phenomenon on the Internet, as it provides new insights on human behavior displacement in times of crisis—giving hope to people’s creativity in finding mechanisms to evade governmental censorship. Though to an arguably small extent, we provide evidence of well-established platforms having undergone a shift in usage—and provide further evidence of operators’ responses. This work raises the general question on the role of the Internet in these periods and the effectiveness of Internet censorship.

C Growth Patterns in Social Media Platforms

Social media is subject to constant growth and evolution, yet little is known about their early phases of adoption. To shed light on this aspect, within this section, we empirically characterize the initial and country-wide adoption of Jodel within Germany and Saudi Arabia. Unlike established social media, the studied network Jodel is anonymous and location-based to form hundreds of independent communities country-wide whose adoption pattern we compare. We take a detailed and full view from the operator's perspective on the temporal and geographical dimension on the evolution of these different communities---from their very first the first months of establishment to saturation. This way, we make the early adoption of a new type of social media visible, a process that is often invisible due to the lack of data covering the first days of a new network.

While we identify *organic* growth within Germany over the course of several years, Jodel kick started only with few weeks simultaneously across the country, which we attribute to virality through Social Media.

C.1 Introduction

Showcasing two examples of the vast landscape of user adoption processes, either emerging, or changing, we next will focus on Jodel, a prime example of a new class of emerging anonymous, location-based web applications that *i*) enable users to post anonymously, without displaying user-related information and *ii*) display content only in the proximity of the user's location.

Consequently, adoption patterns can vary not only between countries, but also different cities or rural environments, opening questions on how adoption spreads and the app usage diffuses through a complete country.

Despite the gaining popularity of this novel type of apps, little is known about how they are adopted, or what factors drive adoption and if adoption (success) might be controllable. In particular, the early phases of app adoption are understood poorly even though understanding the social mechanisms behind such diffusion processes are crucial for the design and roll-out of such platforms. This lack of information on early phases of adoption is rooted in a lack of empirical data covering the early phases of a new platform.

This section presents an empirical characterization of the nation-scale adoption of Jodel within Germany and in comparison the Kingdom of Saudi Arabia (KSA), covering the entire adoption phase beginning with the app launch until August 2017. Note that as of today, the platform is still in very popular within both countries. Given that research on these phenomena relies on the cooperation with its operators, this kind of country-wide studies have not been broadly available to our community so far.

As for DE, we provide an overview to multiple years of application activity starting in 2014 of which we will rather briefly describe user base developments, which—as we show—follows an *organic* growth pattern.

As for KSA, our observation period includes the time from the first registered user in March 2015 to the country-wide establishment in August 2017. We focus, however, on the time from the first significant app interactions within the KSA in Aug 2016 until the beginning of Aug 2017. In March 2017 the application experienced a sudden and drastic influx of new users where the usage increased from hardly any to country-wide adoption. From a network perspective, this sudden adoption of a new application represents a change in network traffic in which a new application suddenly appears at a country-wide scale. From an operator perspective, this sudden adoption can be initially looking like malicious use (e.g., to spam the application).

Structure

- 📁 C.2: Related Work
- ♥ C.3: The Birth of the Jodel Networks in DE and the KSA [JH5]
 - 🗨️ C.3.1: Different Adoption Pattern in Germany and the KSA
 - 📄 C.3.2: Partitioning the Communities by Volume and Time
- ↗ C.4: Organic Growth Jodel DE
 - 📊 C.4.1: Community Interactions Over Time
- ↗ C.5: Adoption of Jodel in the KSA [JH8]
 - 🕒 C.5.1: Temporal Adoption
 - 🌐 C.5.2: Spatial Adoption
- 📄 C.6: Conclusions

C.1.1 Research Questions

The aim of this section is to showcase and *empirically characterize* the early adoption phase of Jodel w.r.t. to user adoption. The app is a well-suited network to study this question given the fact that its location-based nature—in which no country-wide communication is possible—enables us to compare the behavior of hundreds of *independent* communities country-wide. The establishment of these communities raises the general question if and how users are onboarding and how the communities evolve over time and with size. We are particularly interested in difference between these communities, but also take the rare chance comparison two countries being culturally different.

C.1.2 Approach

Our data-driven study compares a plethora of independent communities country-wide in Germany, and the Kingdom of Saudi Arabia. Our analysis focuses on two orthogonal dimensions: *i*) temporal, and *ii*) geographic adoption. This way, we can empirically characterize and model the adoption and usage of a new type of social media on a country wide scale, while providing a comparison of two (culturally) different countries.

C.1.3 Results

While we observe a gradual increase in Jodel usage across Germany over years—*organic* growth, the adoption in SA began in Riyadh (capital) and happened in 3 phases, most notably a phase of sudden adoption—in all communities simultaneously—supported by social media influencers advertising Jodel. Surprisingly, the different communities show the same qualitative adoption pattern nationwide.

Furthermore, we identify a heavy-tailed distributions across all communities: Most of them experiencing only few interactions and content, while few others are flooded with new content.

C.2 Related Work

Social network analysis is a wide and very active field of research since their early days. Various perspectives and platforms have been investigated providing understanding with empirical or qualitative studies ([Mislove et al., 2007, Nazir et al., 2008, Schiöberg et al., 2012, Kairam et al., 2012])

However, very little is known about the early adoption of a new social network. Existing works partially provide information about the growth and development of online social networks, such as Reddit [Singer et al., 2014], Yahoo 360 and Flickr [Kumar et al., 2010, Mislove et al., 2008], Google+ [Gong et al., 2012, Schiöberg et al., 2012], Facebook [Wilson et al., 2009] and others [Benevenuto et al., 2009, Jiang et al., 2013] usually relying on sampled information; further, they do not focus on drivers or reasons for network growth in particular. Especially our ground truth information enables us to empirically trace the birth and development of a new community in detail—in this case of DE and the KSA.

While patterns of successful online communities are well understood (e.g., [Kraut et al., 2012]), only few studies investigate *early* pattern of user retention. Identified success pattern include the diversity of different actions performed by new users [Karumur et al., 2016] or the presence of replies (feedback) and its semantic content [Yang et al., 2017].

Further, most literature investigates platforms from the Western hemisphere, while neglecting other regions, such as the Middle East except for rare examples [Reyae and Ahmed, 2015], or Asia.

We complement these works by contributing the first large-scale empirical analysis of the Jodel messaging application adoption within Germany and the Kingdom of Saudi Arabia based on ground truth information.

C.3 The Birth of the Jodel Networks in DE and the KSA

The growth patterns of social networks are less understood given that data captures from the very beginning of a social media platform are typically unavailable. We take the rare chance to begin showcasing the rate in which the Jodel platform established itself in both countries. Our first peek is relevant to better understand network activity and to define a meaningful aggregate layer for comparison, e.g., time slices, for studying cultural shifts in social media usage in the next section.

C.3.1 Different Adoption Pattern in Germany and the KSA

We show the adoption of the Jodel network in both networks by the number of interactions over time in Figure C.1. The figure shows the number of weekly interactions for Germany (solid line) and the KSA (dotted) since the very first interaction till the end of our data set in August 2017. With interaction, we refer to any interaction with the Jodel system, i.e., either posting, replying, or voting.

The adoption of Jodel in Germany is characterized by a slow but rather steady growth of network activity over time, peaking in 2016/2017. This captures the birth of the Jodel network that originated in Germany and then constantly increased in popularity. In contrast to the steady increasing activity in Germany, the adoption of Jodel in the KSA is characterized by a substantial influx of users and an increase in activity at a short time in March 2017. To our best knowledge, the reason for this behavior is that Jodel went viral via social media in Saudi Arabia—in absence of any marketing campaigns of the operator itself and has suddenly turned Jodel KSA into a vivid place throughout the country.

While studying the reasons that were driving these adoption processes is beyond the scope of this section, their adoption processes differ substantially. That is, referring to Table D.1, in only 4 months, Jodel KSA has roughly gathered 1.2 million users, while Jodel Germany over six times longer time period accumulates to 3.6 million. Likewise, the amount of interactions equally scales between the KSA with 1 billion and Germany with 3 billion interactions. This observation and differences in adoption allow for putting aggregates, e.g., comparable time slices, for our study into perspective.

Takeaway Adoption pattern and thus associated traffic can differ substantially.

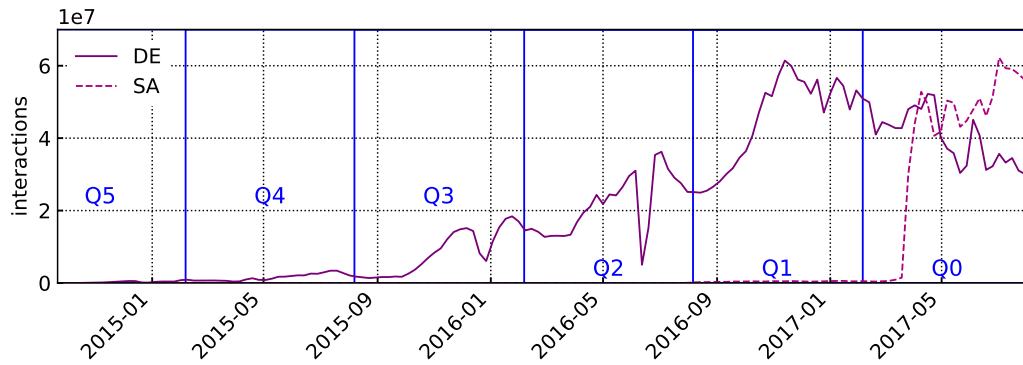


Figure C.1: Observed Jodel Activity in DE and SA. Jodel activity in Germany (DE) and the Kingdom of Saudi Arabia (SA) over our observation time. While the DE network enjoys a steady growth over time, the usage in the KSA has suddenly surged and remains stable.

C.3.2 Partitioning the Communities by Volume and Time

To compare the social media usage in both countries, we need to derive comparable datasets. That is, we need to enable comparisons between two different populations of posts, interactions, and users across a multitude of communities. We thus identify two main dynamics within the lifecycle of communities defining the aggregation dimension: *i)* time and *ii)* per community interaction volume.

C.3.2.1 Slicing by Time

As we are limited in the length of observations, especially for Saudi Arabia, we selected half-year long timeframes backwards from the end of observation. We show these timeframes with vertical bars in Figure C.1 named by country $DE_0, DE_1, \dots, DE_5, SA_0, SA_1$, where the index indicates the partition's age. This simple batching approach creates half-year slices that represent various stages within the community development. In our study, we compare these slices independently in each evaluation to account for the different adoption processes. We have experimented with higher resolutions to enrich our results with more data points (not shown), but our conclusions remain the same for our presented period length. We handle these partitions independently of each other, i.e., early-day users from DE_5 may drop out of the statistics in subsequent partitions due to a lack of interactions.

C.3.2.2 Slicing by Community Interaction Volume

Defining “a” community is not possible on Jodel given that content is displayed relative to the users location and thus differs from user to user. That is, every user might experience a slightly different community to interact with, which cannot be reconstructed from the data. To solve this, we assign each interaction to a nearby major city or district, which generates clusters of interactions that we refer to as communities. This discretization generates an approximation of the individually experienced communities. The resulting approximation is of sufficient accuracy to study and compare the Jodel usage in different parts of the respective countries. Further, the discretization does not normalize for covered area, nor covered population.

We mitigate these inherently hard problems in normalization by simplifying our partitioning approach. By slicing all interactions into quantiles ordered by their corresponding community size, we enable a relative comparison; named is_{q0_25}, \dots representing the corresponding quantile of all interactions, discretized into communities (leading to an approximation). We provide details of this partitioning in Table C.1 describing the amount

| Interactions quantile | #communities | |
|--------------------------|--------------|-------|
| | SA | DE |
| $\approx 75..100$ | 1 | 14 |
| $\approx 50..75$ | 4 | 29 |
| $\approx 25..50$ | 12 | 114 |
| $\approx 0..25$ | 78 | 6,678 |

Table C.1: Community Aggregates. Interaction volume aggregation layer and amount of corresponding communities. Due to the heavy tailed distribution across the community discretization, the upper quantiles consist of fewer communities.

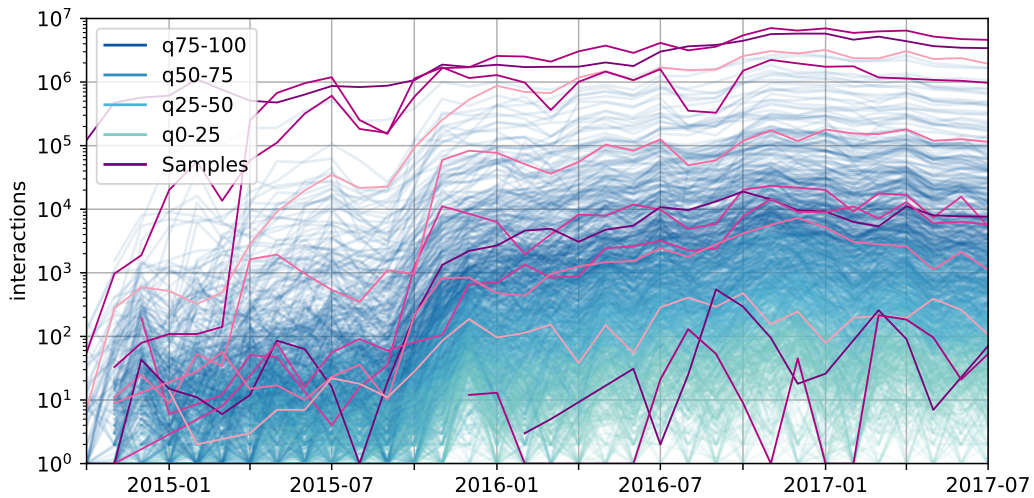


Figure C.2: Monthly Individual Community Interactions over Time, DE. Sample $N=1k$. We observe a steady growth through the communities, albeit a general uplift by late 2015.

of discretized communities per country. That is, e.g., the single largest SA community is the capital Riyadh at about 30% total interaction volume—hence, it is the only community within the set of q_{75_100} . Due to dividing the interaction volume into equal parts, we encounter a heavy-tailed interaction distribution across communities resulting in only few entries within the upper quantiles; the German community size distribution qualitatively matches the SA counterpart, while the latter is largely shifted in volume within magnitudes of fewer communities (not shown).

We will compare the social media usage based on the resulting data sets.

C.4 Organic Growth Jodel DE

The journey of the Jodel application begins back in late 2024 within Germany in the city of Aachen. Being the new kid on the block, Jodel actively set out to various cities with large universities focusing on students as their main target customers.

Within this section, we will briefly showcase and discuss the expansion of the Jodel app in DE and the synchronization between independent communities as shown in Figure C.1.

C.4.1 Community Interactions Over Time

While we have already seen that the overall interaction within Germany have risen over time, we first analyze the amount of platform interactions over the whole observed dataset timespan for individual communities.

C.4.1.1 Individual Communities

In Figure C.2, we showcase a sample of 1k different communities within Germany over time. While the x-axis denotes time, the logarithmic y-axis sums all platform interactions. To provide an idea how the general development took place, we have added a series for each of our sampled communities and colored them according to their final community-interaction quantile (blue). Further, we have added selected samples of a range of communities to exemplify the observed behavior in more detail.

While the large amount of samples does not conveniently allow for distinguishing between specific instances, the overall trend becomes very apparent. In the early days, Jodel usage was dominated by only rather few communities. However, in late 2015, we observe a clear uplift in overall usage by orders of magnitude (at least for some instances). At this point, also the assignment to volume quantiles becomes more stable in comparison to earlier timeframes. We observe consistent amounts of communities and interactions across all interaction regimes; however, the q0-25 communities often experience also few or even no interactions at all.

By taking a closer look into a set of selected Sample communities from various volume-quantiles, we observe a stark increase in usage within early days up to 2016, which more or less saturates later on. We want to highlight that the Jodel usage across communities does not specifically correlate, i.e., their individual interactions over time vary, which we call *organic growth*.

C.4.1.2 Community-Quantiles over Time

Due the community-volume quantile being fixed by overall interactions from the beginning until the end of observation in previous perspectives, we next take a closer look into a time-dependent classification in Figure C.3. That is, we subdivide the community volume quantiles by time as well: Tq0, ..., Tq5 series of which the latter denotes newer data (see Figure C.1).

First, within Figure C.3a, we show the amount of interactions per community. For all communities across any size, we observe significant increases of which the lower quantile communities q00-25 experience most variance. As discussed earlier, it becomes apparent that the amount of interactions settles for Tq0 and Tq1.

Second, we show the amount of communities per quantile over time in Figure C.3b. While Jodel has expanded across Germany, the amount of communities for each of the quantiles increases as expected. The overall heavy-tailed distribution in interactions per community remain over time and even increase.

C.5 Adoption of Jodel in the KSA

We start by analyzing the overall Jodel usage evolution in the Kingdom of Saudi Arabia (KSA). Thereby gaining first insights on our first research question: how fast did the adoption in the KSA occur (cf. C.5.1: [Temporal Adoption](#)) and do adoption patterns differ geographically (cf. C.5.2: [Spatial Adoption](#))? To answer it, we analyze how its users interact with the network in terms of registration and interactions (i.e., content creation and voting events).

A user becomes part of the Jodel network via a device registration event. That is, whenever a newly recognized device starts the Jodel application, the system automatically assigns the device a new user account in the background. We show the geospatial development of Jodel within the KSA in Figure C.4a. The figure shows the app interaction activity as heatmaps on a per snapshot logarithmic scale for four days in 2017: February 1, March 13, March 28 and August 1 (left to right). A lighter/darker color indicates a higher/lower

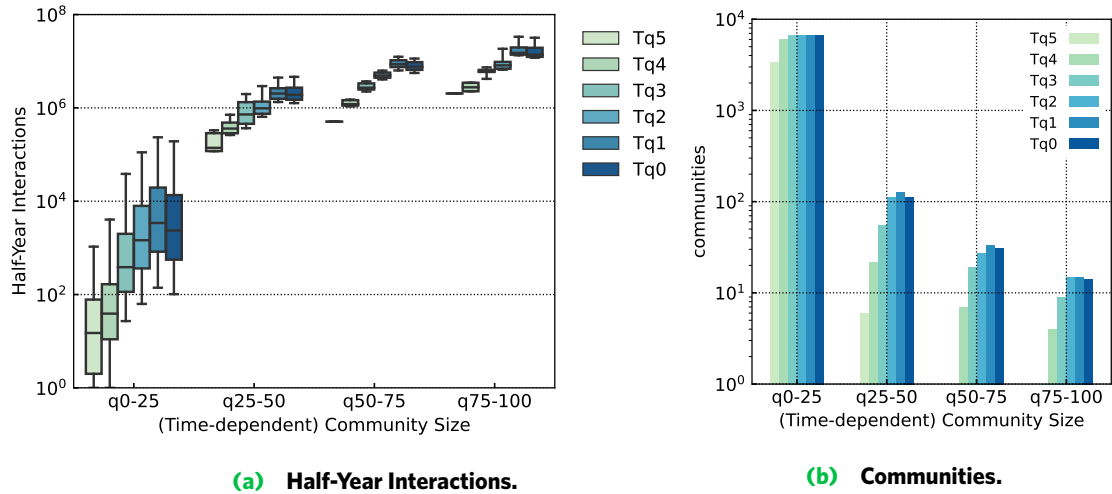


Figure C.3: Half-Year Interactions. The DE Communities sliced into half-year time partitions. (a) *left*: Small communities enjoy only few interactions; we identify an exponential growth along larger communities. (b) *right*: Albeit remaining heavy tailed, the amount of communities per partition over time generally shifts towards a more homogeneous distribution.

amount of activity, respectively. At the beginning of 2017, the capital Riyadh was practically the only city where Jodel was used, while the adoption swept over all major cities later on. We next begin with a detailed study of the temporal phases of Jodel adoption in the KSA (cf. [C.5.1: Temporal Adoption](#)) and then study the adoption pattern of the different Jodel communities (cf. [C.5.2: Spatial Adoption](#)).

C.5.1 Temporal Adoption

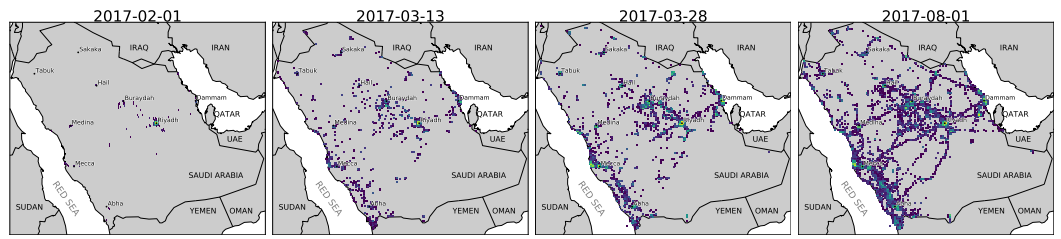
As for the temporal adoption, we will define three different phases that we will discuss in detail next.

C.5.1.1 Phase I: Early Inception (2016)

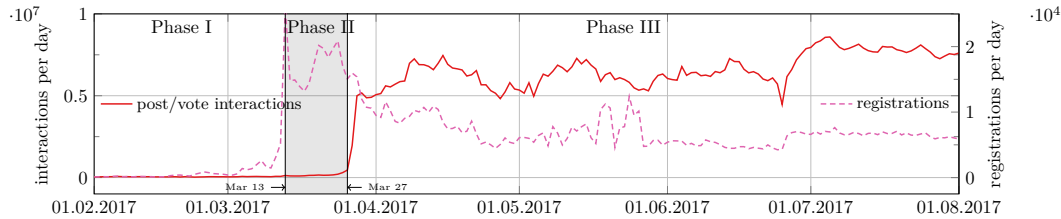
A first peak in usage and registrations can be traced back to August 3, 2016 (not shown). With negligible activity before in the KSA, the Jodel adoption grew by 140 and 170 users on two consecutive days. Afterwards, it experienced a small, but steadily increasing influx of new users. Due to the growing community, also the number of system interactions increased from 1k and 5k posts/replies on the first two days to more than 15k/day posts and replies throughout August 2016; the number of up- and downvotes evolved similarly. This early adoption coincides with an update of the similar YikYak application which introduces user handles and profiles, and thus abandoned anonymous posting capabilities (see [Kircher, 2016]). This finding *may* suggest that users switched to Jodel to keep-up posting anonymously. While this marks the birth of Jodel in the KSA, its widespread usage started months later.

C.5.1.2 Phase II: Sudden Growth in March

On March 13, 2017 the number of new user registrations and on March 27th, the number of messages posted to Jodel within the Kingdom of Saudi Arabia (KSA) increased almost 100-fold over the previous weeks, and continued to increase from there over the following weeks. The increasing app usage is highlighted in Figure [C.4b](#) showing the number of daily user interactions (y-axis) with the network by their type over time (x-axis). We omit interactions before February 2017 since there was only little usage within the KSA that is not



(a) **Qualitative geospatial development of user interactions in the KSA in 2017.** This figure shows snapshots in time (before the jump-start, at the beginning on March 13 and March 27 as well as Aug 1) across the country. The colored mesh depicts the number of system interactions log-normalized for each snapshot; lighter color describes higher activity. While the absolute amount of interactions in February is negligible, first users focus on the capital Riyadh.



(b) **Total user interactions in the KSA Feb to Aug 2017.** This figure describes the number of different user interactions with the system. The x-axis depicts the time, whereas the y-axis marks the amount (registrations right y-axis, other interactions left y-axis). New user registrations first peak on 13 March, whereas activity heavily explodes two weeks later. We are interested in who drives this jump-start and the subsequent user base development.

Figure C.4: Jodel usage adoption and development over time.

directly visible in the plot. The registrations suddenly peaked on March 13 at 28k new registrations and then decreased afterwards. The number of new registrations later settles at about 7.5k/day beginning in June. We define this sudden growth in both user registrations and the actual system usage in March as the beginning of the widespread adoption of Jodel within the KSA. We call this sudden adoption happening *jump-start*.

C.5.1.3 Signs of external triggers

This observation opens the question on what triggered the huge influx of new user registrations in March 2017. Since the design of the Jodel app inherently limits the ability of users to only communicate with others in close proximity, the large influx of new users at a country-wide basis was likely triggered *externally* rather than originating from internal growth. One would suspect that such a jump-start has its origin in either marketing or promotional activities—or by mentions of public figures. Knowing that the Jodel company did *not* launch any advertising in this region, the origin must be driven by users, advertising Jodel via external platforms. Since the Jodel user base is anonymous, we cannot provide ground truth information by interviewing early adopters on their motivation to start using Jodel. However, the sudden peak in March is correlated to increasing attention to the Jodel app on other online platforms. Examples include search activity for the Arabic term “Yudel” [Google, 2020].

To look for external triggers, we manually inspected the social media platforms Twitter and Instagram, given their popularity in Arabic speaking countries (see [Dennis et al., 2016]). This way, we identified 15 KSA-based influencers (i.e., social media users followed by a large number of users) who have shared *funny* content originally posted on Jodel on their social media accounts [jodel sa, 2018, iim7mdz, 2018, bduc_, 2018, 3w1_4, 2018, 5vmd, 2018, 1pi6i, 2018] within the time frame when the registrations started to peak. Figure C.5 is just one of many examples in which the user [iim7mdz, 2018] (694k followers on Twitter and 3.5M followers on Instagram as of June 2020) shared Jodel content within these two

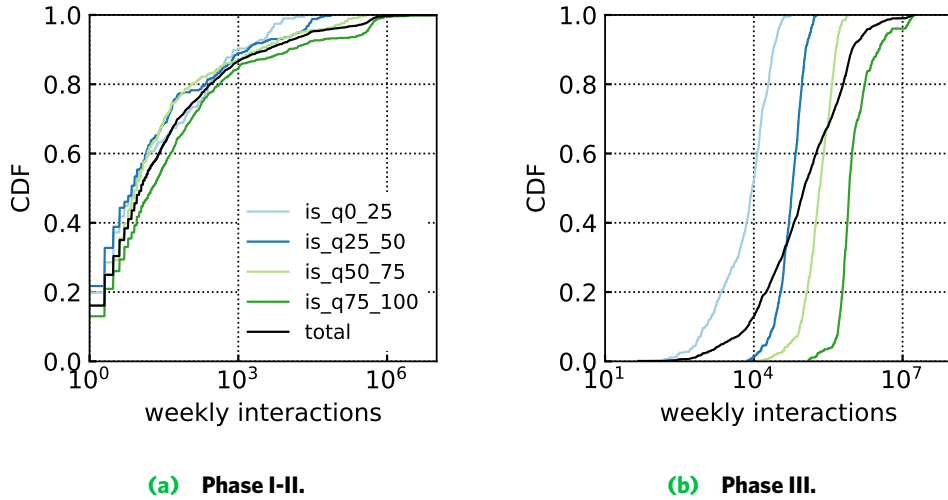


Figure C.6: Community-Quantile Interaction aggregates by week. We observe strong power-law distributions for community interactions. (a) The amount of weekly interactions remains qualitatively equal between later determined community sizes. (b) As per design, with stabilized community interactions, the weekly interactions tear apart in their quantiles.

a country—without any country-wide communication. Thus, differences in the adoption pattern between communities could be expected.

Community Interaction Volume. We will base our subsequent analyses per design on overall community interactions through partitioning them in quantiles. However, to better grasp the distribution of interactions between these groups, we present Cumulative Distribution Functions (CDFs) of weekly figures in two different time periods: *Left*) Phase I-II in Figure C.6a, and *Right*) Phase III in Figure C.6b. Both CDFs show the city quantile subsets and the total CDF in comparison. While the distributions are qualitatively similar, we do not observe major variations, i.e., though interactions follow a power-law across communities. As per design, the CDFs tear apart within Phase III.

Takeaway Jodel remained largely unknown within the KSA up to mid March in 2017. With high confidence, we suspect external Social Media having triggered a sudden growth in popularity afterwards.

C.6 Conclusions

In this section, we empirically characterize the adoption process of Jodel across Germany and Saudi Arabia. While the German user base has grown in an *organic* fashion, within the Kingdom of Saudi Arabia, Jodel experienced a sudden nation-scale influx of new users accompanied by corresponding interactions. We dissect various platform characteristics from the first post to saturation through the lens of an operator by using complete and ground truth data. The location-based nature of Jodel forming hundreds of *independent* communities throughout a complete country enables us to compare their adoption patterns. The major adoption phase is characterized by a sigmoid nation-wide user growth and a two-week delayed startup in user activity in *all* communities. That is, the adoption is characterized by a massive influx of users that occurred in all communities nationwide. We hypothesize that this adoption must have been triggered externally by other social media, such as Twitter or Instagram, and is not the result of organic growth or epidemic spread.

By comparing these communities w.r.t. interaction volume (size), we identify similarities, (power-law) scaling effects in community size and rare differences. However, we identify

scaling effects: larger communities attract more users to be active on a daily basis. Also, independent of community size, the observed amount of un- and popular content as well as the ratio of upvotes (happyratio) is similar across all city sizes. Social credit is granted within minutes in larger communities (reply to a post) while being orders of magnitude slower in smaller cities, scaling with size. We further identify that content voting popularity differs between the city subsets: users in larger communities are more likely to start new threads in comparison to smaller communities, although there already is a substantial amount of content available to them. While we find similarities between the community sizes in user lifetime and retention, regardless of community size, positive reactions correlate with a user's lifetime and her number of interactions. Yet, invariant to their size, all communities develop a stable daily active user base with more than 60% of the users keeping using the app until the end of our observation on average.

Chapter Summary

Within this chapter, we set out to exemplify the vast landscape of possible user adoption processes. We detailed onboarding processes to new and existing online services, while also observing changes in usage repurposing communication on well-established services. Thus, we have uncovered the *A)* user onboarding processes of applications in rich demand of whole societies for the individual and collective good at times of crisis, ★ *A) Corona-Warn-App scheduled for the good* , *B)* shifts in platform usage from to reallocating its purpose to a side-channel at nowadays hybrid warfare, ★ *B) Platforms as Side channels* , and *C1) and C2)* evolution of new social and entertainment platform, Jodel, being hyperlocal 📍 by design and thus being spatially limited to distinct communities, comparing two different country's landscape of local communities, ★ *C) Growth and Adoption of Jodel DE & SA* .

Though we started with a broad introduction from an application perspective, from this point on, we will focus on our rich data of the Jodel application as an example of a hyperlocal anonymous social messaging platform having identified that even the very same ingredients to a platform may result in vastly different outcomes.



USER INTERACTIONS

In a nutshell. All Online Social platform are defined by its user base and their interconnections and interactions. Building social ties, mining social credit, creating content, engaging discussions and emerging moderation processes through dis-/likes are the very essential functions. While Literature has enriched our knowledge about various platforms at various implementations and interaction/content focuses, the unique combination of Jodel being *Anonymous* and *Hyperlocal* remains largely in the dark.

While some works provide insights, we complement and enrich our understanding in the key design features with an in-depth data-driven structural analysis between hundreds of local communities across two spatially distinct countries: *Germany* and *Saudi Arabia*. After providing rich empirical insights uncovering heavy-tailed distribution allover, we show that **in-country** communities are **homogeneous**; we identify **cross-country shifts** within the distributions of interactions. We take a deep dive into vast dissimilarities adding a *temporal* dimension and discuss structural implications: The *German* user base enjoys *voting* more than creating contents, whereas the *Saudi* users prefer lively and lengthy *discussions*. We finish the chapter with a deep-dive **modeling** approach to the *Saudi* communities identifying various invariants and scaling laws across community size.

CONTENT

| | | |
|----------|--|-----------|
| .1 | Related Work | 80 |
| A | Structural Community Analysis | 81 |
| A.1 | Introduction | 81 |
| A.2 | Data Analysis | 81 |
| A.2.1 | Platform Content | 81 |
| A.2.2 | Community Perspective | 82 |
| A.2.3 | User Perspective | 86 |
| A.2.4 | Anonymous Friends | 87 |
| A.3 | Conclusion | 87 |
| B | Cross-Country Differences and Structural Implications | 89 |
| B.1 | Introduction | 89 |

| | | | |
|----------|-------|--|-----------|
| | B.1.1 | Research Questions | 89 |
| | B.1.2 | Approach | 90 |
| | B.1.3 | Results | 90 |
| B.2 | | Geographic Differences in Jodel Usage: DE vs. KSA | 90 |
| | B.2.1 | Overall per-user activity is country-independent | 90 |
| | B.2.2 | Difference: posting vs. voting | 91 |
| B.3 | | Structural Implications | 92 |
| | B.3.1 | Content Voting | 92 |
| | B.3.2 | Content Response Time and Volume | 94 |
| B.4 | | Conclusion | 95 |
| C | | Platform and User Centric Analysis - Spotighting SA | 97 |
| C.1 | | Introduction | 97 |
| | C.1.1 | Research Questions | 97 |
| | C.1.2 | Approach | 98 |
| | C.1.3 | Results | 98 |
| C.2 | | Platform Interactions | 98 |
| | C.2.1 | Partitioning the Communities by Rank | 98 |
| | C.2.2 | Interaction Dynamics | 98 |
| | C.2.3 | Platform Implications | 100 |
| | C.2.4 | Modeling Community-User Activity | 103 |
| C.3 | | A User-centric View | 107 |
| | C.3.1 | User Interactions | 107 |
| | C.3.2 | Anonymity - Absent Social Ties | 107 |
| | C.3.3 | Hyperlocality - User Communities | 108 |
| C.4 | | Conclusions | 108 |

Introduction

The last chapter has shown how (new) users adapt to platform and its usage, either from the very first interaction, or within well-established usage patterns. With Jodel, we showcased how even the very same platform ingredients to a messaging platform may lead to much different outcomes in user adoption. This chapter will now take a deeper look into the realm of user interactions, the essential and key driving elements, for any online community platform. The recent pandemic disruption has even accelerated the shift of our everyday's life into digital spaces, may it be at work or socially [Feldmann et al., 2021]. Note: We focus on message and platform contents in the next chapter [◆ 6: USER CONTENT](#).

All Social Media platforms revolve around its users and especially happening interactions. Such interaction may be if plentiful types, such as creating virtual friendships or following the stars. Besides structural interactions or bindings, people enjoy publishing content in various (combined) media types, engaging discussions, either under a profile successively mining social credit—driven by (dis-)likes and general content appreciation. Specific environments are not bound to pure enjoyment or actual friendship, but may emerge in interests, such as the various Stack networks—likewise moderated by (dis-)likes, social credit and gamification. Others, possibly simultaneously, disguise in anonymous confession boards, or possibly argue at total anonymity in a potentially toxic environment.

Current research contributes to the general [◆ User Interactions](#) across their respective platform and realm. Most studies either focus on analyzing social-media usage *worldwide* or by focusing on specific parts of the world, mostly English speaking. These works have enriched our understanding of social media. Yet it is unclear if or to what extent the unique design features of Jodel of Hyperlocality [📍](#) and Anonymity [🔒](#) influence user behavior and their interactions. While related work provides rich insights, we complement these perspectives with a data-driven approach comparing not only various independent communities, but also two spatially distinct countries having a different cultural background: Germany and the Kingdom of Saudi Arabia.

We provide rich structural insights to the Jodel platform across both countries in [★ A\) Structural Characterization](#) incorporating a first analysis dimension distinguishing between communities by size via interaction *volume*.

Structural overview comparison between the German and Saudi communities. We discuss various community insights, platform content, interactions, and user behavior. Further, we provide a rough picture that the anonymous settings does *not* particularly lead to interacting user clusters [✳️](#).

Focus prevalent significant differences in platform usage and showcasing platform implications. From found differences in the empirical characterization, we engage those findings in depth via a [★ B\) Cross-Country Comparison](#). After highlighting certain significant disparities in interaction distributions between DE & SA adding a *temporal* dimension to our analysis, we elaborate on structural implications.

Deepen understanding of Saudi community internals, closing a gap in research focusing a country of the Middle-East. As we believe that many worldwide regions are underrepresented in literature and especially due to lack of insights to such a rich dataset, we finish this chapter by providing an in-depth analysis and modeling of the Saudi user base and communities in [★ C\) Detailed Empirical Analysis KSA](#) over community size *rank*—in favor of a qualitative insight—and *time*.

.1 Related Work

The research community established a rich field of understanding human interaction within social media, yet not studying geographic differences in social media usage. As discussed earlier, empirical studies on social media focused on the birth and growth [Schiöberg et al., 2012, Mislove et al., 2008, Reyaee and Ahmed, 2015], social media usage in specific regions such as the Arab Gulf states [Reyaee and Ahmed, 2015] specifically focused to global usage [Leskovec and Horvitz, 2008], information propagation [Cha et al., 2009], specific platforms such as Facebook [Lewis et al., 2008, Nazir et al., 2008], YouTube [Brodersen et al., 2012], SnapChat [Vaterlaus et al., 2016], or Twitter [Kouloumpis et al., 2011, Bollen et al., 2011], or knowledge sharing ([Wang et al., 2013]). Such research tries to understand and identify social structures and influence [Kairam et al., 2012, Tang et al., 2009]. This way, they have shown that social networks usually create small-world networks ([Manku et al., 2004, Freeman, 2004]).

Modeling [Van Mieghem et al., 2011] and graph methods are common techniques to characterize platforms and analyze social ties [Gong et al., 2012, Bielenberg et al., 2012, Kumar et al., 2010, Magno et al., 2012].

Others measure effects of sampling [Catanese et al., 2011], or focus on geospatial distributions within global networks showcase that there exist inherent local biases [Mittos et al., 2020, Papasavva et al., 2020, Schiöberg et al., 2012]. This also holds true for Reddit; though not being reliant on geo-position, [Buntain and Golbeck, 2014] find that users tend to participate only in a single community (topical subreddit).

A recent body of research aims at understanding anonymous social networks. The desire for anonymity can result in throwaway accounts [Leavitt, 2015] and can also manifest in anonymous self disclosures [Birnholtz et al., 2015]. Anonymous content platforms have been detailed w.r.t. user behavior e.g., on 4chan [Bernstein et al., 2011, Papasavva et al., 2020]. It was empirically studied by [Wang et al., 2014], with a distinct focus on classifying the anonymity sensitivity of the posted content (see [Correa et al., 2015]). To the best of our knowledge only very few data-driven empirical insights to the very similar platform YikYak, in [McKenzie et al., 2015, Saveski et al., 2016, Saveski et al., 2016]

Summary. In all cases, we repeatedly find power-law distributions in content, followers, interactions, and users in the realm of online platforms. While Social Networks specifically implement social ties, others only provide opportunities to follow people—a unidirectional friendship—or are anonymous. Platforms introducing such social ties typically result in higher clustering coefficients of users within the networks: my friends likely also know each other; or other (abstract) concepts, such as content or location, may concentrate homogeneous user groups.

A Structural Community Analysis

A.1 Introduction

We engage into the realm of user interaction by providing a structural platform analysis. This section acts as a first empirical introduction to the dataset explicitly distinguishing between the German and Saudi Arabian communities that reveals basic distributions, but also already shows various insights. First, we focus on contents from a platform perspective, such as amounts of text and images, posts and replies, content lengths distributions, and distributions of available interactions types. Further, we detail daily and weekly usage patterns before we move on to a user perspective, which describes per user interaction distributions and the absence of social ties by showcasing distinct user-pair interactions.

Structure

♥ [A.2: Data Analysis](#)

📄 [A.2.1: Platform Content](#)

🌐 [A.2.2: Community Perspective](#)

😊 [A.2.3: User Perspective](#)

👤 [A.2.4: Anonymous Friends](#)

📁 [A.3: Conclusion](#)

A.2 Data Analysis

While we have already discussed heavy tailed community sizes and their interactions for DE and SA (cf. [♦ 4: USER ADOPTION C.1](#)), we provide a broader perspective into the various platform interaction types, entities, and their relation.

The following subsections will detail various details beginning with platform content types, the community user base sizes and interactions and temporal usage patterns; we further explore the user perspective w.r.t. interactions and gathered reactions. Lastly, we showcase how often users encounter each other within the anonymous setting.

A.2.1 Platform Content

First, we are interested in the type of content being posts on the platform as users may provide text or pictures. Further, the platform bases upon lively discussions within various discussion threads, which leads us to question to which extent new threads or replies to threads are submitted to the platform.

A.2.1.1 Text and Images

Within [Figure A.1](#), we show the cross-distributions between posts, replies and being a picture or not. Focusing on Germany first in [Figure A.1a](#), pictures, especially within threads (only 0.2%), are rarely posted totaling at only 2.3% of all contents. While 17.1% of all content represents threads, we find 4.8 times the replies.

In comparison to Germany, the users in the Kingdom of Saudi Arabia ([Figure A.1b](#)) tend to use more pictures within their communities—not only within threads (2.5%), but also within replies (1.0%). However, users are generally posting less threads (12.5%) than the DE counterparts.

| | | post | reply | | | post | reply | | | | | | |
|---------------|-------|----------------|---------------|------|-------|----------------|----------------|-----|-------|---------------|---------------|----------------|--|
| pic | 2.2% | 0.1% | Σ 2.3% | !pic | 14.9% | 81.9% | Σ 96.8% | pic | 2.5% | 1.0% | Σ 3.5% | | |
| | 14.9% | 81.9% | | | 10.0% | 86.5% | | | 10.0% | 86.5% | | Σ 96.5% | |
| | | Σ 17.1% | Σ 82% | | | Σ 12.5% | Σ 87.5% | | | | | | |
| (a) DE | | | | | | | | | | (b) SA | | | |

Figure A.1: Content Types. While the German users enjoy posting images more than the Saudi counterparts, they focus more on posting new content than replying and discussion within threads. The Saudi users also use pictures more often within discussions.

A.2.1.2 Text Content

As textual content is very predominant, we ask further how the given character limit of about 250 is used. That is, in Figure A.2, we present the message length and token distribution across replies and posts for the German and Saudi user base of a 10M random sample for each country. The violin depicts the Probability Distribution Function (PDF), while the horizontal lines denote quantiles (50% dashed, 25% and 75% densely dashed).

Text Lengths. We present message length distribution in Figure A.2a for DE & SA. Our very first observation is that message lengths are much longer in DE in comparison to SA. This difference is apparent within replies, which tend to be much shorter, but becomes very striking for posts. While the median DE thread has a length of about 100 characters, this value is much smaller for SA (45). The same holds true for replies where SA posts are much shorter than the German counterparts. We generally observe that thread starting posts are longer than their replies. Though it remains unclear to which extent these significant differences may be explained by details of the locally written language—German vs. Arabian.

Tokens Lengths. To provide further insights, we also show the amount of Tokens per message in Figure A.2b. Besides variances in very low character counts, the distributions is very similar to text the lengths.

① We provide more detail to **platform content** and provide a discussion about **the role of emoji** in section **◆ C: The Role of Emoji**.

A.2.2 Community Perspective

Next, we want to emphasize on the various independent local communities that emerge and showcase community discretization. First, we discuss amounts of interactions on a community basis, followed by showcasing macro-level time-dependent platform usage patterns.

A.2.2.1 User Base and Interactions

Thus, we distinguish between the available user interaction types for these communities. In Figure A.3, we provide CDFs of interaction types, but also amount of participating users, across all communities in Germany (DE, blue) and Saudi Arabia (SA, pink). The logarithmic

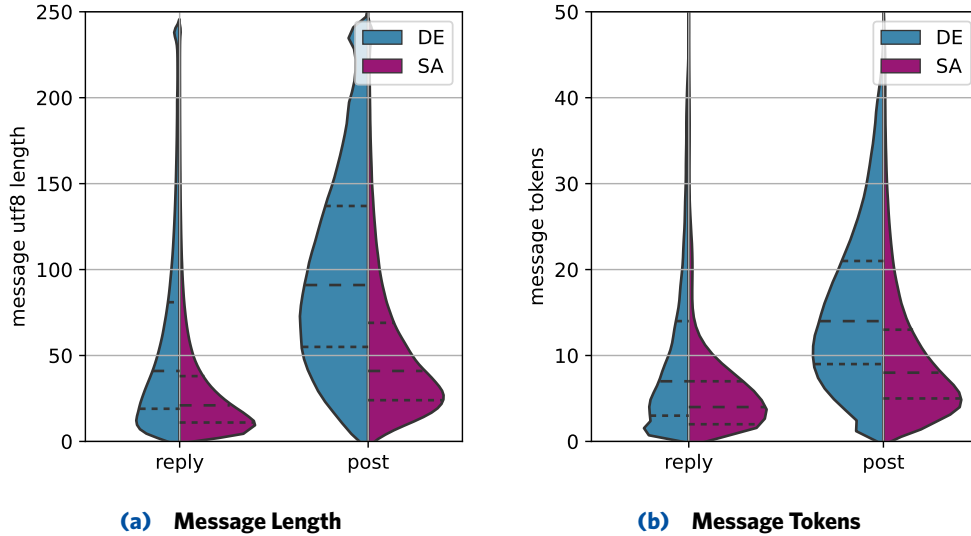


Figure A.2: Text Messages - Lengths and Tokens, DE & SA. Random Sample $n=10M$ per county. Both, message and token lengths distributions correlate. We generally observe longer thread posts in comparison to replies. German users tend to write and reply longer texts than the Saudi user base.

x-axis denotes the respective value within the distributions, whereas the y-axis denotes the Cumulative Distribution Function (CDF).

Limitations. As shown within the dataset description in Section [◆ D: Ground Truth Dataset and Corpus](#), observation timeframe for both countries differ significantly: While the German user base is growing since 2014, Jodel usage has kick started in SA only about half a year before the end of observation. Further, the amount of interactions per community vary heavily. While there are only few communities with a very lively user base, many others only occasionally has noticeable activity. We would like to remind the reader that the amount of discretized communities is about 6.8k within Germany, while only using about 100 discretized communities in Saudi Arabia stemming from different spatial popularity distributions within these countries.

Hence, the absolute community counts should not be compared between the different countries. However, we make considerable observations on a qualitative level: *i)* Throughout, we observe heavy-tailed distributions (note the logarithmic x-axis), and *ii)* The distributions allow for cross-country popularity comparisons.

ⓘ We provide further in-depth insights applying **temporal and spatial rasterization** as discussed in [◆ C.3.2: Partitioning the Communities by Volume and Time](#) within this chapter later.

Content Creation. For content creation in Germany, user have two options, posting a new thread (blue, solid) or engaging discussions (blue, dotted) within existing threads within her community. As seen earlier (Section [A.2.1.1](#)), across the whole range of communities, we observe that replying is more popular than creating new posts (the series is right-shifted in comparison). Likewise, we observe the same pattern for the Kingdom of Saudi Arabia user bases (posts: pink, solid; replies: pink, dotted), whereas the margin between both interactions is larger.

Voting. As for voting, the German users are very active in voting content down (blue, dashdotted), but mostly up (blue, dashed)—by a wide margin. Both interactions are more

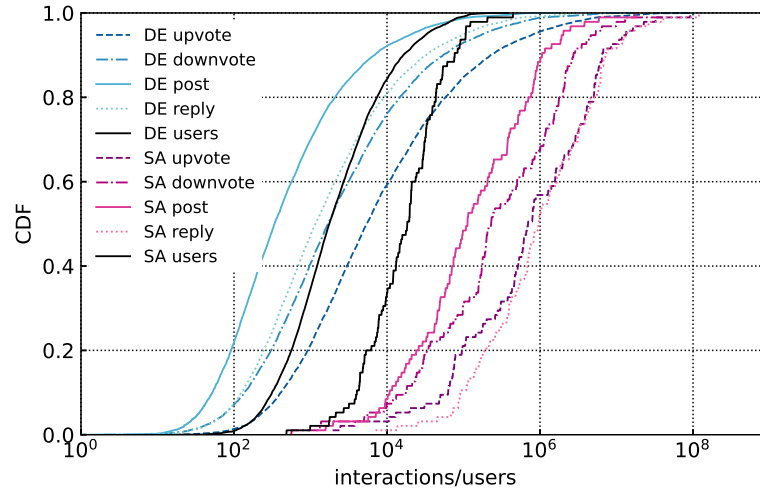


Figure A.3: Per Community Interactions and Users CDF, DE & SA. These CDFs resemble the total observed platform users and interactions per discretized community. While DE-SA Scales are not comparable due to different time periods, we are interested in qualitative differences. We observe heavy-tailed distribution all over at varying popularity and amounts per user.

popular than creating content. Likewise, the Saudi user bases also prefer upvoting (pink, dashed) to downvoting (pink, dashdotted). However, contrasting the relation between content creation and votes, the Saudi user bases share a different opinion: While creating new threads is the least popular platform interaction across all community sizes, replying to threads is by margins more popular in comparison to DE, placing it in the popularity regime of upvoting content especially within larger communities.

① We discuss effects of voting later from a **structural perspective** section [♦ A: Structural Community Analysis](#), and w.r.t. **distributed content moderation** in Section [♦ A: Distributed Content Moderation](#).

Users. As discussed earlier, the community discretization results in a heavy-tailed distributions for the interactions and amounts of content. We are next interested in linking the interactions to users, who drive the communities.

Therefore, we also plotted the user CDF for both countries (black, solid; DE left, SA right). Maybe expectedly, the amounts of users per community also describe heavy-tailed distributions for the German and Saudi communities. Nonetheless, we learn that the amounts of interactions qualitatively shows a correlation to amounts of participating users.

More interestingly, a comparison of the number of users to the various amounts of interactions reveals a notion of per user interactions already. For the DE communities, users on average all have upvoted multiple times, whereas other interactions are widely considered less popular. However, this comparison reveals increasing skews within the interaction power-law distributions w.r.t. amounts of users. I.e., the larger communities attract even more interactions compared to the smaller counterparts. That is, in about 80% of the DE communities, not all users have replied only once on average. Further, the top 10 communities account for about 20% of all platform interactions.

For the SA user bases, the skew within the user distribution appears slightly different compared to DE within the larger regimes favoring few very large communities in SA (i.e., the top 5 communities concentrate 50% of all interactions). Though we do not observe such a twist in the interaction dynamics, a skew within the interaction distributions to the amounts of platform participants prevails over the SA landscape.

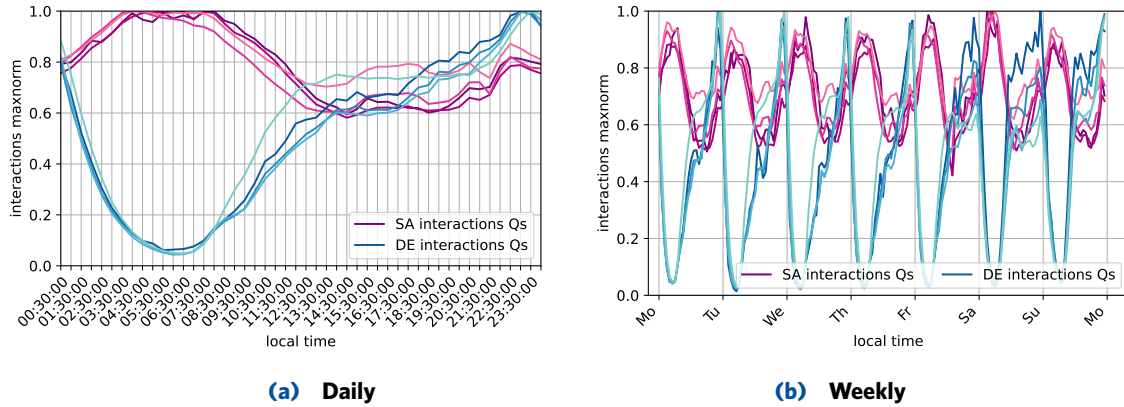


Figure A.4: Interaction Patterns - DE & SA. Aggregates across all observed data. The time zones are adjusted to local time. We distinguish between German (DE, blue) and Saudi (SA, pink) communities. Further, we subdivide the populations by their size quantile in volume: the darkest color denotes the lower q_0-25 quantile, whereas the brightest color represents the upper quantile q_{75-100} . *Left (a) Daily:* For DE, we observe a daily usage patterns similarly to general European internet traffic. The SA user base shows a very different usage pattern focussing on nighttime and a striking base level throughout the day. *Right (b) Weekly:* The SA user bases do not change in daily patterns across the week, while peaking Saturdays. While DE communities likewise follow the same pattern within the workweek, larger communities experience less usage on Friday and Saturday nights.

A.2.2.2 Daily/Weekly Activity Patterns

While he have learned about power-law distributions across the platform in terms of communities, but also for the amounts of users within these communities. The different skews within the CDFs between users to interactions also further hint into a heavy-tailed distributions for interactions per user as well.

However, still keeping the macro perspective on the communities and interactions, we are next interested in daily usage patterns. At least for DE, we would expect a typical day/night cycle as seen in internet traffic [Maier et al., 2009], where activity dips within nighttime around 4am in the morning, while beginning to rise again when people start working—increasing further throughout the day. The finale typically results in a steep increase to a peak time around 9-10pm and rapidly decreasing figures afterwards into the night again.

Thus, we next analyze typical day/night cycles aggregating the complete observation timeframe in Figure A.4. Whereas the x-axis denotes relative local time (time-zones are adjusted), the y-axis encodes the platform interactions normed by the maximum values ($y=1$). We subdivide this analysis into German (blue) and Saudi (pink) communities aggregated by community size quantiles in volume.

Note that our observations for the German user base are in line with related work on YikYak [Saveski et al., 2016].

Daily. As described for typical internet traffic within Europe, we take a deep dive into the daily usage patterns in Figure A.4a.

Focussing on DE first, the observed usage curve matches typical internet traffic [Maier et al., 2009] very well. Note however that measured observations at an IXP vantage point do not necessarily generalize well. There is a wide dip in usage throughout the night until morning spanning a gap of about 12 hours of interactions amounts being less than 60%. While the usage remains on a steady level between noon and afternoon, usage increases in the evening peaking at 11pm.

Interestingly, for SA, we observe a counter-pattern: The Saudi Communities are most active throughout nighttime, while sharing steady behavior from noon into the evening.

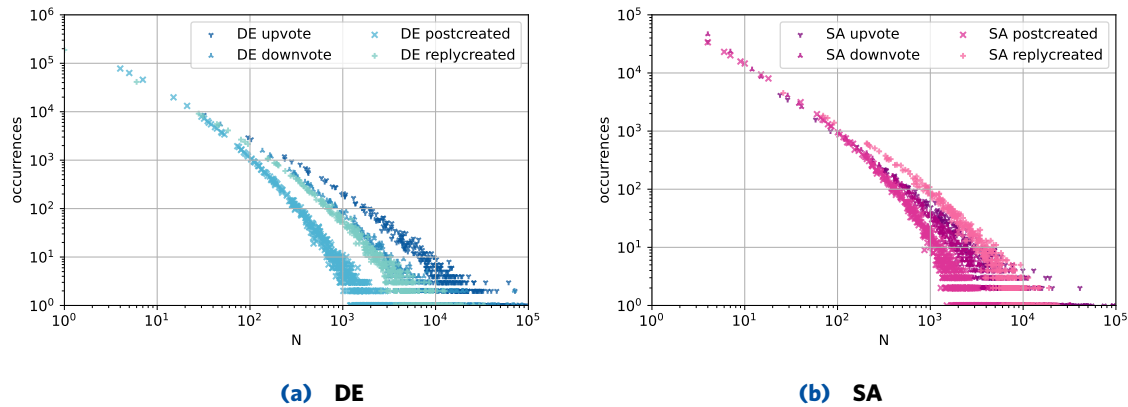


Figure A.5: Per User Casted Interactions PDF. Random sample $N=500$. We generally observe heavy-tailed distributions across all interactions types. Again, interaction preferences become apparent (a right-shift translates to being more popular).

Further, we do not observe a drop in usage as extensively seen for the German communities. Usage always remains on a very steady level above 60% of the observed maximum bin. Note however, that the larger communities of Q_{75-100} (bright pink) appear to be an outlier w.r.t. the other quantiles encountering an even higher interactions lower bound at about 70% of the maximum.

Weekly. Next, we zoom out to a weekly time period, which we present in Figure A.4b. For the Saudi communities, the daily pattern repeats across the whole week, while peaking on Saturday nights. While this is the same for the German communities along the workweek, they are quite less active Friday and Saturday night within the larger communities (brighter blues). Note that the lower DE quantiles (darker blues) do not follow the same dampening over the weekend.

A.2.3 User Perspective

After having enlightened the community perspective w.r.t. content and interactions, we next take a user centric view. That is, we analyze the distributions of platform interactions aggregated by users in Figure A.5, and further analyze received reactions to own content in Figure A.6 as Probability Distribution Functions (PDF) random sampled at $N=500$. Please note that these plots resemble a log-log axis of which the x-axis denotes the value of the respective metrics, while the y-axis denotes the amount of occurrences. We distinguish between German (DE, blue) and Saudi communities (SA, pink).

Interactions We provide details of the PDFs for the German communities in Figure A.5a, and the Saudi communities in Figure A.5b. Catching up on hints found earlier, we observe heavy-tailed distributions across the users in their participation regardless of interaction type. Again, we observe user preferences within the right-shift of the series trajectory.

Reactions Next, we want to shed light on the distribution of platform interactions w.r.t. how much attention users get. We define reactions as replies to a thread or gathering a vote on posted content (threads and replies), direction interactions. In Figure A.6a we show the German, and in Figure A.6b the Saudi user bases' distributions. Both distribution, replies and votes, again are heavy-tailed. That is the vast majority of users only receive few reactions—on average correlated to participation. Again, interaction preferences become apparent for the DE communities (a right-shift translates to being more popular). How-

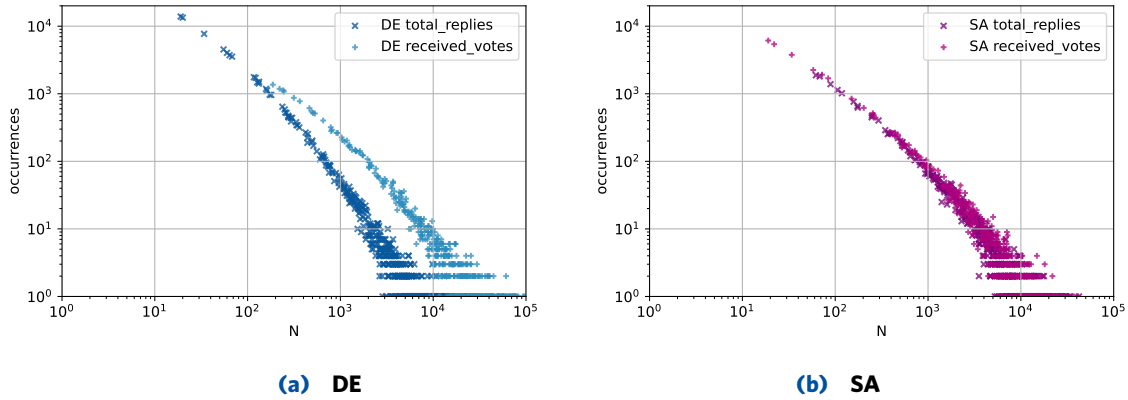


Figure A.6: Per User Received Reactions PDF. Random sample $N=500$. We again observe heavy-tailed distributions across all interactions types. And again, interaction preferences become apparent (a right-shift translates to being more popular). However, due to votes being less common compared to the DE user bases, the ratio in received interactions remains balanced between replies and votes.

ever, due to votes being less common compared to the DE user bases, the ratio in received interactions remains balanced between replies and votes.

A.2.4 Anonymous Friends

We finish our journey through the first characterization of the Jodel communities by determining how often users *see* each other again in the regime of an anonymous platform. More specifically, we ask how often two users interact in a thread/reply correlation, irrespective from whom creates a thread or reply.

We show a Complementary Cumulative Distribution Function (CCDF) of the ratio of users having encountered another user as defined in Figure A.7. We distinguish between German (blue) and Saudi (pink) communities. To focus on the impact of community size, we further subdivide our analysis into city quantiles by volume. Note that both country populations span different timeframes of general activity. Nonetheless, we spot differences across community sizes for both countries. That is, within the SA Q0-25 quantile, about 35% of the user-user interaction pairs remain a singleton, whereas about 17% of the pairs encounter each other at least two times. Users within the lower quantile communities (darker colors) experience more singleton encounters, whereas the probability of increased re-encounters increased with community activity.

A.3 Conclusion

Within this very first structural shallow dive into the driving aspects of the platform, we first focus on content types which is predominantly text. Lengths are rather short in replies for Germany (DE) and Saudi (SA) Communities likewise, whereas new threads tend to be longer in DE.

We identify power-law distributions for all platform interactions along the community discretization at varying popularity. The general usage-pattern over time follow a typical internet-traffic cycle in DE, whereas the SA users tend to be most active at nighttime. While the communities experience heavy-tailed distributions within interactions, we likewise identify heavy-tailed distributions for interactions per user as well—in terms of cast interactions, and received reactions to own content.

Finally, the anonymous setting does not particularly promote re-encounters of users. Specifically for larger communities with much more content, it becomes unlikely to ever interact with the same user multiple times again.

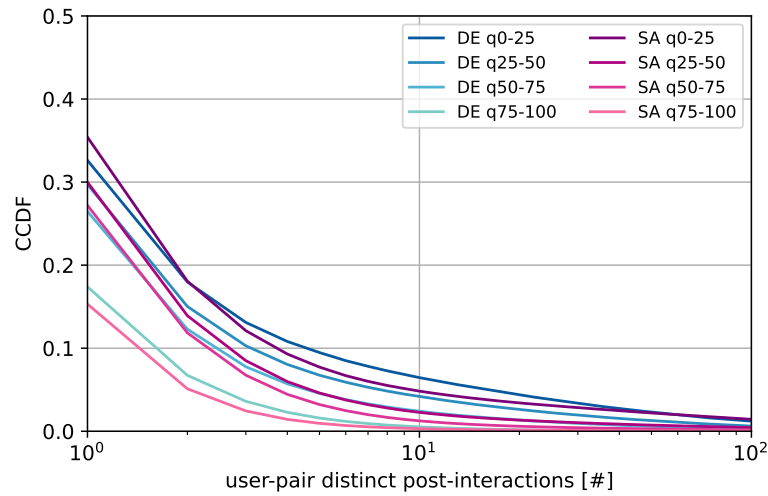


Figure A.7: User-Pair distinct Thread-Post Interactions, DE & SA, by Community Size. We compute all user pairs that have interacted through replying to a thread. The Cumulative Distribution Function shows that user re-encounters at all are rather rare across all community sizes; to a larger extent in SA. Smaller community sizes increase this probability.

Aligning with Related Work, Social Media platforms (e.g., in terms of interactions, or friendships/followers) results in power-law distributions all over, the rich get richer. Furthermore, we observe for Jodel, that due to anonymity and thus absent social ties, user re-encounters become very rare. This may lead to ephemerality as observed on other platforms as well [Bernstein et al., 2011], while the communities are in fact creating identities as seen on other platforms [Heston and Birnholtz, 2016a, Guta and Karolak, 2015, Zhang et al., 2017, Gaudette et al., 2021]. Memes [Zannettou et al., 2017, Zannettou et al., 2018, Chen, 2012] as well are an integral part of the Jodel platform (as e.g., has been shown in specific emoji usages [◆ C: The Role of Emoji](#)).

B Cross-Country Differences and Structural Implications

In this section, we empirically analyze two examples of a Western (DE) versus Middle-East (SA) Online Social Messaging App. By focusing on the system interactions over time in comparison, we identify inherent differences in user engagement. We take a deep dive and shed light onto differences in user attention shifts and showcase their structural implications to the user experience. Our main findings show that in comparison to the German counterparts, the Saudi communities prefer creating content in longer conversations, while voting more conservative.

B.1 Introduction

Given the the quite different Jodel adoption processes (cf. [◆ 4: USER ADOPTION](#)), and profound differences in interactions as showcased before, we next want to provide a specific view in cross-platform differences and their implications to the platform and its user base.

Research provided a general understanding through the empirical and qualitative analyses of a number of different networks. Examples include structural measurements of classic online social networks [Mislove et al., 2007, Nazir et al., 2008, Schiöberg et al., 2012, Kairam et al., 2012] as well as more specialized variants such as microblogging [Bollen et al., 2011], picture sharing [Vaterlaus et al., 2016, Cha et al., 2009], or knowledge sharing [Wang et al., 2013]. Works in this field analyzed the networks' *structure*, mostly using graph-theory approaches. This way, they showed that social networks usually creates small-world networks [Manku et al., 2004, Freeman, 2004]. The influence of cultural or geographic backgrounds on usage largely remain unknown.

Most studies either focus on analyzing social-media usage *worldwide* or by focusing on specific parts of the world, mostly English speaking. These works have enriched our understanding of social media. Yet it is unclear if or to what extent the obtained measures differ between different geographic regions. Cultural differences are known to exist that drive human behavior in social networks, e.g., the degree of connectivity [LaRose et al., 2014] or how marketers use social media to impact purchase decisions [Goodrich and de Mooij, 2014]. Yet, little is known on how geographic or cultural backgrounds may impact the way users interact with a social media platform in terms of the generated traffic; that is, content creation and content voting.

Structure [JH5]

[📍 B.2: Geographic Differences in Jodel Usage: DE vs. KSA](#)

[❤️ B.3: Structural Implications](#)

[🗨️ B.3.1: Content Voting](#)

[📊 B.3.2: Content Response Time and Volume](#)

[📄 B.4: Conclusion](#)

B.1.1 Research Questions

In this work, we take the rare chance to analyze ground truth information provided by a social network operator to compare interactions with a social media platform of a Western (Germany) and a Middle-Eastern (Saudi Arabia) country. We selected both countries since they represent the largest user-bases of the Jodel platform, while simultaneously representing a largely different (cultural) background. Our data sets capture the entirety of all social media interactions in both countries since the very first post. This way, we can, for the first time, shed light on whether geographic or culture specific differences exist between both countries w.r.t. how the user-bases generate and vote content.

B.1.2 Approach

The studied social media platform Jodel is location-based and anonymous. Most importantly, the feature of Jodel to form independent local communities enables us to compare in-country and between country effects and thereby to clearly identify country specific usage differences. Further, it does not display any form of user profiles or other user-related information that would introduce visible social credit; users solely interact framed into their physical proximity and based on their topic preferences. This results in a *pure* form of communication that is reduced to content, since any form of influence by user profiles such as social status is absent. This makes Jodel an ideal platform to study differences in content creation and voting, i.e., the entirety of active interactions with this social network. We shed light on fundamentally different user behavior and engagement patterns within such anonymous spaces having received less attention as of today, across the Kingdom of Saudi Arabia and Germany; and showcase structural implications.

B.1.3 Results

Our contributions are as follows.

- While not a primary focus of our work, we empirically show the very different adoption processes of a new social media platform in both countries.
- We show that, invariant to time and community size, users in Saudi-Arabia (Middle-East) behave fundamentally different to the German counterparts (Western country). They prefer creating content, but vote slightly less than the German users. This highlights, for the first time, that country-level differences in the usage culture in social media exist that create drastic differences in user behavior of the very same social media platform.
- We exemplify the implications of shown user engagement to the social media platform. E.g., we show that the availability of more content in Saudi Arabia naturally decreases the available votes per post, which can have serious impact on, e.g., distributed community moderation techniques.

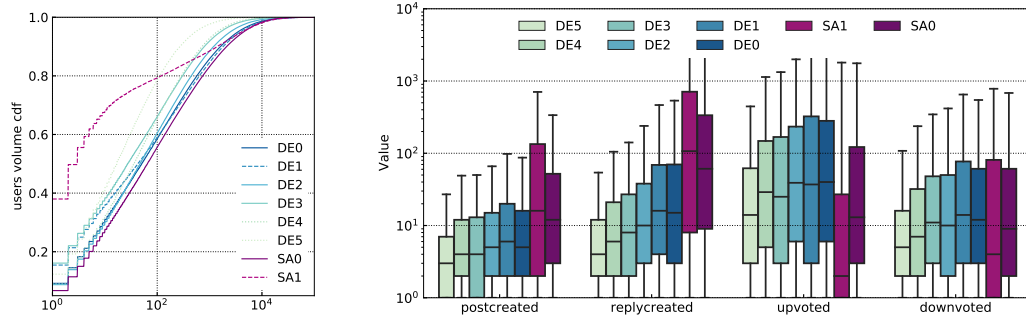
B.2 Geographic Differences in Jodel Usage: DE vs. KSA

Is there a systematic difference between Jodel users in Germany and in the KSA in the way they use the social media platform? In other words, do culture or country specific usage behaviors exist that uniquely define traffic profiles of the very same platform in each country? While social media usage has widely been studied, the question of in-platform variation and behavioral differences is still open. In this section, we set out shedding light on this aspect by comparing the Jodel usage in two countries with a different social and cultural background.

We study the question of (cultural) differences w.r.t. user behavior in Jodel usage by investigating differences in active user *interactions* with Jodel, i.e., posting and voting, accounting for all possible active system interactions. We base this evaluation on two factors: partitioning by *i*) time and *ii*) interaction type, which we discuss next.

B.2.1 Overall per-user activity is country-independent

First, we subdivide all interactions into independent half-year periods as described before (DE0, . . . , DE5, SA0, SA1). These independent periods enable to compare the behavior of the networks at different times to account for differences in the adoption of Jodel in DE and SA. We then aggregate the interactions to each user and present the resulting interactions per



(a) Total per user interactions per time period. Per user interaction volume is very similar over time.

(b) Per user interactions by time period (DE0, ...) and interaction type. Over time, we observe increasing engagement. There is a substantial difference in content creation and upvote activities between SA and DE.

Figure B.1: Differences in per-user interactions by time and country. *a)* While the individual users behave similar across the board w.r.t. interaction volume, *b)* attention shifts towards content creation for the SA user groups, whereas downvoting is less popular in comparison to DE.

user CDF in Figure B.1a. We show each half-year period as separate CDF (series) for each of the two countries. Invariant to time, we observe quite similar heavy tailed distributions; that is, most users are not very active, such that e.g., 60% of all users each have up to only 100 interactions. The distribution for SA1 deviates from the pattern since it captures a timeframe before being popular (considerably fewer data points). In general, we observe that irrespective of time and country, users follow a similar usage behavior—also in absolute terms (not shown).

B.2.2 Difference: posting vs. voting

Next, we further partition the data by the type of interaction in addition to the time slices used before. That is, we show distributions of interactions per user subdivided into the voting interactions (upvoted, downvoted) and content creation interactions (postcreated, replycreated) as a box plot in Figure B.1b. Note the logarithmic y-axis. Further, the whiskers denote the 5%/95% percentile.

German users tend to increase their engagement over time at increasing platform activity regardless of the interaction type. While upvoting is the most prominent type of interaction for the German users, voting content down and replying to content are roughly equally less prominent. The SA users prefer content creation, especially replying, whereas upvoting happens less frequently.

Remarkably, all time periods within a country are determined by similar behavior. In other words, posting content is the dominant type of interaction in the KSA, while it is voting in Germany. This represents a clear difference in platform usage that can be observed between these two countries (also regardless of community size, not shown).

The ratio of up to overall votes remains positive at a happyratio (upvotes to total votes) of 83% for DE and 71% for SA. The figures for the SA1 partition need to be taken with a grain of salt due to only few users; however, the engagement spread is higher compared to the latest timeframe SA0.

Takeaway Invariant to time and community size, the SA users (Middle-East) behave fundamentally different to the DE counterparts (Western country). They heavily prefer creating content, but vote slightly less than the German users. This highlights, for the first time, that cultural patterns in social media user behavior exist that create drastic shifts in how a very same social media platform is used in each country. This finding may be considered

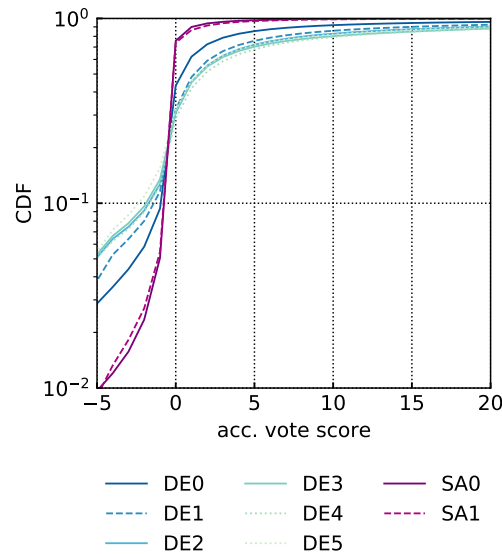


Figure B.2: Platform accumulate vote scores and vote interactions per post. Larger communities experience a stronger heavy tail in vote scores. Accumulate vote scores CDF over time.

even more interesting, given Jodel being an anonymous platform that enables a very pure form of communication; it entirely focuses on posted content in absence of any user profile.

B.3 Structural Implications

With the identified fundamental attention shift in user behavior between SA and DE users in content creation and voting, we now aim at studying resulting implications on the platform. According to the operator, the communities in both countries are considered to be well-functioning. That is, regardless of implications arising from different usage profiles, participants are enjoying spending time on the platform to the most part; e.g., by creating content, voting, or just lurking.

B.3.1 Content Voting

B.3.1.1 Accumulated votes overview

Votes on posted content have two roles in Jodel: *i*) they show content appreciation to others (e.g., enable users to sort by popularity) and *ii*) enable distributed user based content moderation that removes content with negative vote scores (cf. chapter [2: JODEL](#)). Factors influencing vote distributions can thus have structural implications on the platform. To study differences in voting behavior, we first take a look into accumulate vote score ($\#upvotes - \#downvotes$) distributions. There are two given bounds for posts gathering votes: 1) Posts beyond a negative threshold are no more displayed on the platform. 2) There is no conscious upper bound given by the system, yet posts are only temporarily displayed within the various app feeds (see [2](#)) and therefore, the time for interacting with them is inherently limited. Given these constraints, all communities naturally enjoy a rather positive mood. To put an emphasis on the temporal dimension, we show the CDF of vote scores to posts over time (DE0, ...) in Figure B.2. In the earlier times of DE with less activity, more posts were able to gather more votes as illustrated by the DE3..5 series resembling a broader distribution. There is a slight decrease of accumulated vote scores throughout time, hence interactions per post, for the DE communities—also observed within the KSA. The distributions become more and more long-tailed over time and appear scale-shifted. Noteworthy, a split into community sizes confirms this finding: larger communities may

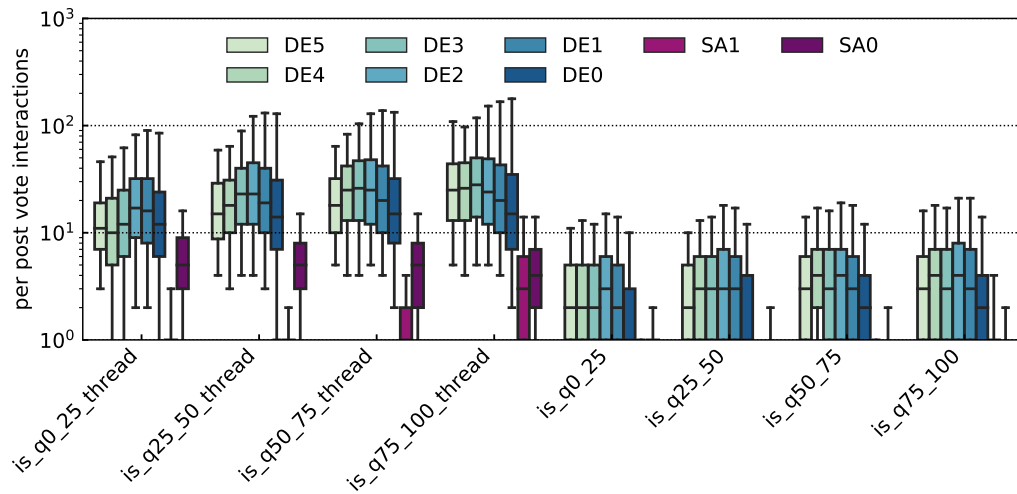
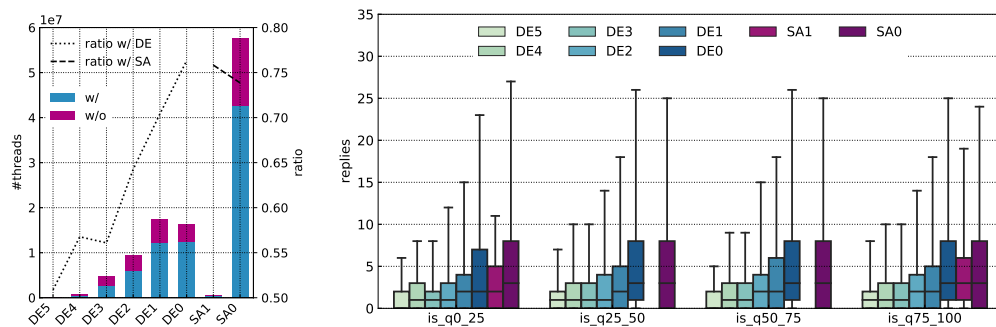


Figure B.3: Platform accumulate vote scores and vote interactions per post. Due to their exposure, threads usually collect more votes than replies buried within them. Vote interaction distributions by time period (DE0, ...) split by community size (q0_25,...) and being a thread start (thread) or is a reply buried within a thread.



(a) Thread w/o replies over time and response ratio. Most engagement (left, Fig. B.4a), communities experience longer threads can attract replies. **(b) Thread length over time by community size.** With increasing time and response ratio. Most engagement (left, Fig. B.4a), communities experience longer threads can attract replies. on average invariant to community size.

Figure B.4: Thread engagement with responses - how many and how much. a) At an increasing ratio over time, within recent time periods, most threads attract replies. b) With growing amounts of interactions over time, the average thread length increases to a similar level in DE and SA.

reach far higher absolute scores, but the distributions become more skewed correlated to the observed interaction volume (not shown).

Due to SA users producing much more content, the feeds displayed in the app also get renewed completely at a very high pace. Thus, SA posts compete harder for time to collect possible votes in comparison to their DE cousins; the feeds also promote observed long-tails. What implications does this shift have on experienced vote distributions?

B.3.1.2 Votes per Post.

To better understand the voting interactions and the observed skew in accumulated vote scores, we next normalize observed figures to a per-post basis. The box plots in Figure B.3 show various per post vote interaction distributions across time (DE0, ...), and community size (q0_25, ...), while further distinguishing between threads (thread) and replies.

We find the long tail of high vote scores in the rather long 95% percentile whiskers on the log scaled y-axis. Invariant to time and community size, the median German user enjoys voting on threads with median levels around 10 to 30 votes gathered by each post

throughout time. Naturally given by the communication structure and app design, content buried within threads is much less appreciated; they accumulate only two to three votes in DE. As discussed before, the SA posts stand in stark contrast at three to four votes within the main Q0 timeframe. I.e., opposed to German users, the average participant within the Kingdom of Saudi Arabia cannot expect to receive any vote on her content—especially and naturally not on replies.

Takeaway We show that the availability of more posted content in Saudi Arabia decreases the available votes per post, which can influence community moderation techniques that depend on voting.

B.3.2 Content Response Time and Volume

While voting or liking is a vital part of a social network, it can only exist because of posted content and replies. We thus next study geographic properties that influence the response time and volume.

In Figure B.4a, we show the amounts of posts with and without replies (bars) and the ratio (lines) across time for both countries. The German communities increase their response cover over time, while it instantly is equally high for the SA communities at about 90%. I.e., 9 out of 10 users in both countries can expect getting at least a single reply on a thread.

B.3.2.1 Response volume

As most users receive a reply, does the total achieved thread length correlate with community size, and how does this interplay with the distribution shift in content creation? We answer this question with distributions given in Fig. B.4b as a box plot, which depicts the thread length gathered per post across time (DE0, ...) and community sizes (is_q0_25, ...).

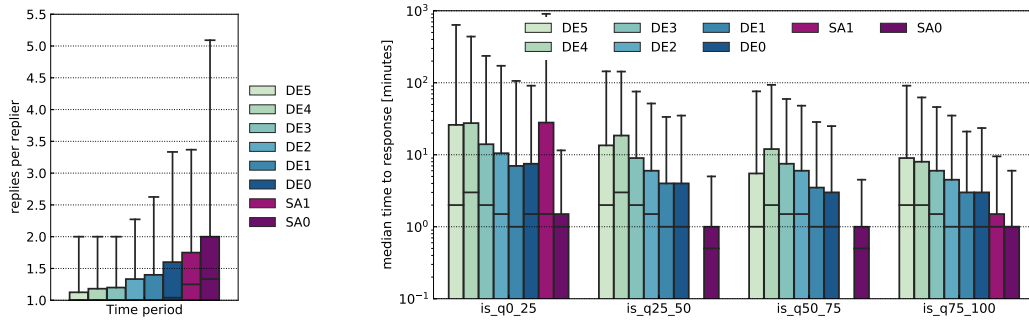
First, the 95% percentile whiskers indicate a long-tailed distribution in the length of threads, which we confirm (now shown). Second, the amount of replies is invariant to community size as the distributions are very similar; however, there still exists a huge spread from the 75% to the 95% percentile (whiskers) due to the long-tailed distribution. Second, we observe an increasing trend over time. This increasing engagement is also apparent when looking deeper into the interactions (cf. section C.3.2.2: [Slicing by Community Interaction Volume](#) and alike split by community size—not shown).

Takeaway Most posts in both countries get a reply; even at larger volumes for SA, the thread lengths are similar to DE.

B.3.2.2 Conversations

Having established an understanding of the amount of replies most users experience, we get into more structural detail. We define *conversationness* as the ratio between replies per replier as a proxy for conversations—where lower ratios naturally depict a heterogeneous set of repliers, while higher ratios indicate fewer participants forming a back and forth conversation.

We present the distributions of this ratio over time for both countries in Figure B.5a as box plots; the 95% whisker indicates long-tailed distributions, which we can confirm (not shown). Over time and with increasing network activity, all German communities increase up to about 1.6 replies per thread participant within the 75% quantile in DE0. This indicates a shift from rather random single comments becoming less popular in favor of interacting with each other. Given the high preference on creating content and vividly replying, this



(a) Conversationness - Replies per thread participant are long-tailed and substantially larger within SA.

(b) Thread engagement speed. The median timeframe between consecutive answers within a thread decreases over time in the meanwhile growing DE environment. Likewise, larger communities experience faster responses---in SA widely immediately.

Figure B.5: Thread engagement with responses - who and when. *a)* The observed overall activity increase in DE results in longer conversations over time, the SA users talk substantially longer. *b)* The average time until receiving a response reduces to only few minutes in both countries, SA still takes the lead.

trend is particularly apparent and reinforced in the SA communities at rates of up to two replies for the 75% quantile of SA0.

Takeaway Employing a conversationness metric, we identify SA users to be more conversational compared to DE.

B.3.2.3 Response time

We have seen that most threads receive at least some replies. While the counts in responses may matter quantitatively, we also want to shed light on the time-dependent dynamics of the reply interactions. Fig. B.5b shows a box plot of the distributions of the time between consecutive responses within a thread split by time (DE0, ...) and community size (q0_25, ...); note the log y-scale. Unfortunately, our dataset does not allow for this evaluation on vote interactions (cf. [♦ D: Ground Truth Dataset and Corpus](#)).

From this evaluation, we gain two major insights: 1) throughout time with increasing activity and engagement, the German communities establish shorter response times down to only minutes. Having reached a sustainable community size, the response times no longer drop. 2) The SA communities instantly drop response times substantially below the German counterpart to only a single minute within most threads. Note: High response times within small communities in SA1 are due to small amounts of data; Missing series indicate no present data.

Takeaway In comparison to DE, the SA communities are more vividly responsive as the average response times are considerably lower.

B.4 Conclusion

In this section, we show that the usage behavior of users in Germany (DE) fundamentally differs from users in Saudi Arabia (KSA) in the anonymous and location based Jodel network. This study is enabled by the feature of Jodel to form independent local communities allowing us to compare in-country and between-country effects and thereby to clearly identify country specific usage differences. We empirically characterize usage behavior based on ground truth user interactions data provided by the operator. While we can rule out marketing effects by the operator, our findings motivate future work that study root causes.

We find that, independent of time and community size, KSA users prefer content creation (posting and responding), while German users tend to interact slightly more passively (voting). Other than this shift towards content, due to the users in both regions else behaving identical on a per-user measure, we find rather identical community engagement. However, due to much more content being available within SA per user, posts compete harder in gathering votes than the German counterparts, which can have implications for vote-based content moderation schemes. Further, the average number of replies also does not increase in comparison; Still, reply times are much smaller due to higher activity. The average Saudi user tends towards having longer conversations. Overall, we identify time- and geographic-invariant differences between DE and SA user engagement as the latter substantially focus on creating content, giving a slight lead in voting to the Germans. This provides a new interaction-based perspective on geographic difference of social media usage that have not yet been studied. We hope to raise awareness for the rich and colorful space of online social networks and their various user groups—each having its very own peculiarities, even given same-same boundaries.

C Platform and User Centric Analysis - Spotighting SA

Social media is subject to constant growth and evolution. As shown early, user behavior between our two countries differ significantly. Due to sparseness in literatures, this section empirically sheds lights on and characterizes the landscape of emerging independent Jodel communities in Saudi Arabia. Unlike established social media, the studied network Jodel is anonymous and location-based to form hundreds of independent communities country-wide whose adoption pattern we compare. We take a detailed and full view from the operators perspective on the temporal and geographical dimension on the evolution of these different communities---from their very first the first months of establishment to saturation. This way, we make the early adoption of a new type of social media visible, a process that is often invisible due to the lack of data covering the first days of a new network.

C.1 Introduction

As we have seen, the very same platform ingredients may still lead to much different user adoption processes and platform usage, i.e., shown for DE & SA. Due to observed surprising kick-start user adoption, and related work specifically lacking contributions focusing geographical regions, such as the Middle East, we decided to investigate Jodel within the Kingdom of Saudi Arabia in detail next.

This section presents the first empirical characterization of the nation-scale usage of an anonymous, location-based messaging app in the KSA; as of today, the platform is still in vivid use within the region. Further, given the cultural particularities of Saudi Arabia, the local and anonymous nature of the app, we set out to characterize interactions and application usage patterns in this emergent setting.

Our observation period includes the time from the first registered user in March 2015 to the country-wide establishment in August 2017. We focus, however, on the time from the first significant app interactions within the KSA in Aug 2016 until the beginning of Aug 2017. We provide in-depth empirical findings and model of various user interaction metrics, structured into two broader sections: A platform view, *i*) providing insights on community volume aggregates, and a user view *ii*) providing insights on application user aggregates.

Structure [JH8]

- ☺ C.2: Platform Interactions
 - 🕒 C.2.1: Partitioning the Communities by Rank
 - 📊 C.2.2: Interaction Dynamics
 - 👉 C.2.3: Platform Implications
 - 🌐 C.2.4: Modeling Community-User Activity
- ☺ C.3: A User-centric View
 - 👤 C.2.4.1: Registrations
 - ✍️ C.2.4.2: Interactions
 - 🌟 C.2.4.3: Voting
- 📄 C.4: Conclusions

C.1.1 Research Questions

The aim of this section is to *empirically characterize* and *model* the user behavior of new social media w.r.t. various community- and user-driven metrics. Jodel is a well-suited network

to study this question given the fact that its location-based nature—in which no country-wide communication is possible—enables us to compare the behavior of hundreds of *independent* communities country-wide. The establishment of these communities raises the general question how they adopt evolve over time and especially with size. That is, we are first interested in community aggregates by interaction volume and interaction dynamics thereof. Second, we want to shed light on user-centric behavior from joining the platform to specific interaction distributions.

C.1.2 Approach

Our study is driven by the comparison of a plethora of independent communities country-wide by focusing on *i)* a community driven analysis to understand interlinked interactions, and *ii)* a user's perspective of participation and platform experience. By comparing these communities w.r.t. different properties, we provide models where applicable and thus largely identify similarities, scaling effects and differences at times. This way, we can empirically characterize and model the application usage of Jodel on a country wide scale.

C.1.3 Results

We characterize interaction distributions across the communities influencing the major app design decision displaying various content feeds—posting vs. replying and upvoting vs. downvoting. Extending on this, we derive further metrics defining community experience. Further, we show that the user influx and per user interactions can be modeled with a power law very well. We analyze user behavior w.r.t. lifetime and retention between different communities, which is very similar. However, daily user activity and churn scales with community size, whereas we observe differences in the interactions of users dropping out of the platform.

C.2 Platform Interactions

The primary function of the Jodel app and social media in general revolves around user interactions. In the case of Jodel, framed into location, this is restricted to anonymous content sharing, communication and dis- or liking contents. We next discuss the dynamics of these platform interactions. That is, we focus on the communities within the Kingdom of Saudi Arabia, specifically with the country's background of the sudden adoption and growth; further, the example of SA depicts a unique niche of a user base that has often been neglected by research so far.

C.2.1 Partitioning the Communities by Rank

In previous analyses, we have applied a community partitioning by interaction volume, focussing active usage of the platform. I.e., the quantiles are heavily skewed w.r.t. number of communities due to observed heavy-tailed distributions. To shift the focus more into the various ranges, we employ a *ranked partition* by their interaction volume: q0-25 represents the < 25% number of cities with the lowest number of interactions while q75-100 represents the top 25% number of cities by interactions.

C.2.2 Interaction Dynamics

Since the Jodel app design provides three different content feeds (*most recent*, *most discussed*, and *loudest*), we are interested in relations between interaction amounts—compared to our partitioning driver community-size. Thus, we next set the stage and provide an overview of how the total platform interactions types, e.g., posting or upvoting, interlink to each other.

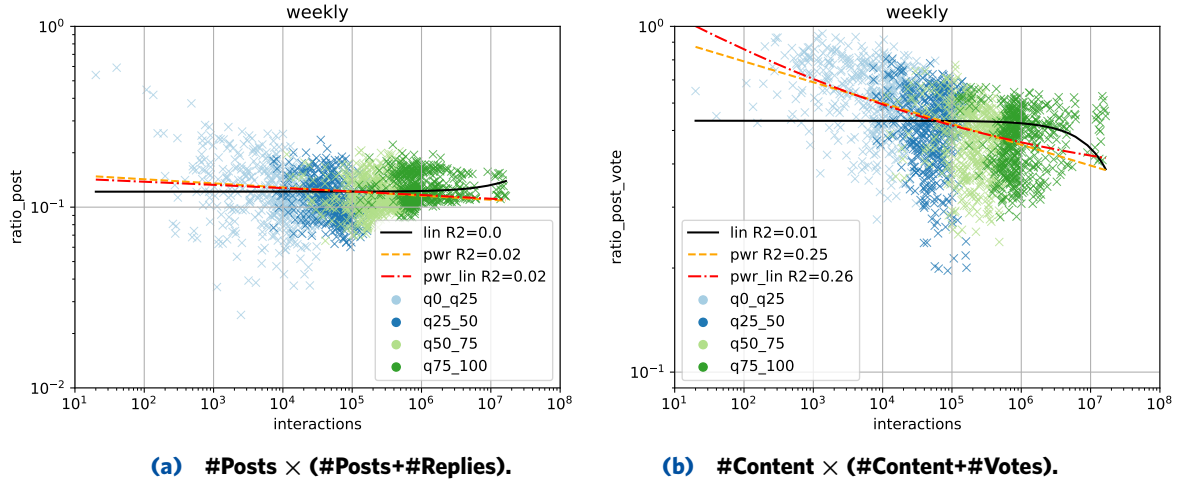


Figure C.1: User interaction ratios by city size. (a) This plot depicts #posts to total content by city size. Starting new threads is less popular in general with a slight downwards trend with city size on average. (b) This plot depicts the ratio of content creation to votes to by city size. With increasing community sizes, the amount of created content to votes converges towards equal popularity.

In Figure C.1 we provide three different scatter plots of weekly averages over interactions per community (x-axis), colored by their quantile. We added the following fitting curves and their fitting R2 score to ease interpretation, which will also be used throughout this section in subsequent evaluations:

$$\begin{array}{llll}
 \text{linear} & \text{lin} = & a & + & b & x \\
 \text{power law} & \text{pwr} = & & & b & x^c \\
 \text{shifted power-law} & \text{pwr_lin} = & a & + & b & x^c
 \end{array}$$

C.2.2.1 Creating Content

First, we measure the postratio (y-axis) denoting the amount of threads (posts) compared to the total amount of content (posts and thread replies) in Figure C.1a, that is a postratio converging towards 1 is post dominated, whereas reply dominated converging to 0 exemplified next. Together with conversationness (B: Cross-Country Differences and Structural Implications), we argue that postratio is a vital indicator in favor of communication (n to m) instead of predominantly shouting content out (1 to n). Due to rather steady-state community behavior as discussed in \blacklozenge C.5: Adoption of Jodel in the KSA, as per design, defined quantiles (color) fit nicely the weekly amount of interactions by community in distinct regimes, i.e., there is an apparent gradient between data points across quantiles. For largest communities, postratios vary from 0.1 to 0.2, i.e., only 1 in 10 or 2 in 20 total content contributions resembles a new thread, whereas, with fewer interactions, values gradually vary stronger, e.g., from 0.06 to 0.23 for q25-50, and even more so for q0-25. Nonetheless, we identify a very homogeneous distribution of threads and replies across all independent communities at a postratio \pm std of 0.122 ± 0.038 on average, i.e., there are about 12 threads in 100 content contributions. The imbalance between threads and considerably more replies heavily influences user experience w.r.t. content perception due to replies being one of three in-app content feeds, *most discussed*, as described in Section A.

C.2.2.2 Content Frequency

On weekly averages, we observe more variation for smaller communities. Nonetheless, *i*) the postratio (posts/total content) is mostly identical across all communities, and *ii*) the con-

tent to total interaction ratio is within arguably similar regimes with a slight downwards trend with community size, the amount of new posts per timeframe qualitatively follows the interaction distributions as shown in Figure [◆ 4: USER ADOPTION C.6b](#). Thus, the resulting frequency of new posts within the local *most recent* app-feed heavily increases with community activity, that can be modeled very well with a shifted power-law with a R2 score of 0.98. Note that these averages do not account for a day/night cycle: rendering the update frequency higher at night. To put this into perspective within Phase III: In the largest community Riyadh, we find about 85 new posts per minute. The second-largest community Jeddah only experiences 29 new threads per minute. However, the most recent feed for Q0-25 communities only gets updated every five minutes.

C.2.2.3 Content and Voting

While content creation forms the basis of the independent communities, only distributed voting brings them to life—making content disappear and favoring mainstream contributions. To bring both interaction types in relation, we scatter plot the ratio of content to votes (y-axis) in Figure [C.1b](#), that is values converging towards 1 (0) are post (vote) post dominated. Again, as per design, we observe a color gradient along the x-axis as expected. Though on average, content dominates total interactions on average at a ratio±std of 0.53 ± 0.13 , from our curve fittings, a linearly shifted power law describes a ratio downwards trend best, i.e., we interpret provided figures as a shift towards higher amounts in replies compared to threads with increasing interactions. Albeit applying a low-pass filter by averaging, we find further confidence of this trend within quantile averages±std: {q0-25: 0.64 ± 0.14 , q25-50: 0.52 ± 0.12 , q50-75: 0.48 ± 0.10 , q75-100: 0.48 ± 0.09 }. Overall, while being very noisy, we observe that both interaction types are equally popular across communities on average rendering any other platform interaction being either contributing content or voting. However, there is a subtle gradient towards a post-domination regime for smaller communities. Although this ratio is not inflicted to any feed, we argue that it provides a first key insight of community capacity for the applied distributed moderation scheme.

C.2.3 Platform Implications

Having set the stage in providing first insights to overall content- and vote interactions, we next keep our community perspective and discuss key structural implications. Due to the network living off of new started threads (posts) and especially discussion within these threads, we next focus on thread response times and conversationness that measures discussion participant homogeneity. Further, as the platform relies on user voting for content steering, we also investigate voting consensus measures.

C.2.3.1 Community Response Time

For measuring response times, as a simple proxy, we restrict our evaluation to the response time of the very first reply to a thread. Note that due our dataset using post timestamps for votes as well, vote interactions do not allow for this evaluation.

In Figure [C.2](#), we show a scatter plot time until a first response to a post occurs over community interactions. While the log scaled x-axis shows denotes the weekly interactions, the log scaled y-axis denotes the time to a first answer measured in minutes. The shading represents the city quantiles; Lines denote applied curve-fittings.

While we observe largely noisy distributions in Phase I, the time to a first answer starts peaking significantly above 7 days on average in Phase II (largely q75-100 data points with fewer interactions below 1k interactions/week). With the uptake in total system interactions, this value decreases for all community sizes. Nonetheless, we overall observe huge

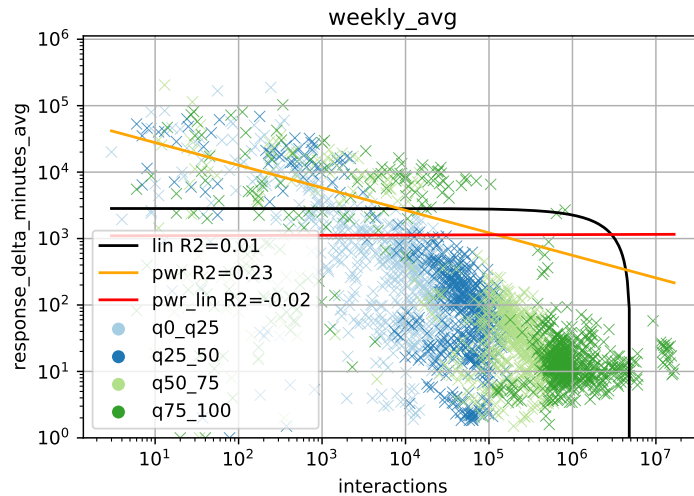


Figure C.2: Interactions × Reply Times The average time until a post receives a response average per city on a weekly aggregate. The distributions are heavily scattered, while we observe a natural trend towards lower response times with increasing interactions.

variations in response times across the board. Larger communities gradually maintain a 100-fold lower response time compared to their smaller counterparts. This may be a primary driver for attracting more participation—the rich get richer, though larger communities widely also correlate to population figures. We model this distribution with a shifted power-law at R2 scores of 0.23. Note that our community-approximation may overestimate larger communities (if intercity distances are larger than the app radius); given this bias, content frequencies arguably still remain quite high.

C.2.3.2 Conversationness

We measure conversationness as the number of discussion participants in a thread to the total amount of answers. That is, values converging towards 0 represents only few participants, whereas values converging towards 1 denote a very wide audience participating within a thread. We focusing on daily or weekly averages across all communities, we find huge variations in the conversationness. On average, conversationness remains equal across community sizes at values of 0.66 with minor standard variation of up to 0.07 for both, daily and weekly aggregates.

To provide deeper insights into the structure of average conversations on the platform, in Figure C.3 we show a scatter plot of conversationness to average thread length. Again, we do not discover any difference between community sizes. However, there is a natural trend to lower conversationness values with increasing thread length, i.e., longer discussions likely are held between various participants. Whereas shorter thread lengths up to five replies experience values above 0.65, we observe a rather linear decrease to only 0.55 for 15 replies. Though a simple linear curve fits quite well, from our references, power-law (pwr) approximates this distribution best with an R2 score of 0.81 and better accounts for the heavy tail at shorter thread lengths.

C.2.3.3 Community Voting Popularity

Finally, we analyze the influence of user metrics on votings by taking the network perspective again. To study the proportion of votes cast to content, we show the sum of posts and replies to the sum of votes and posts/replies per city as scatter plot in Figure C.1b. A ratio of 0.5 indicates both activities, voting and creating content, to be equally popular. Ratios

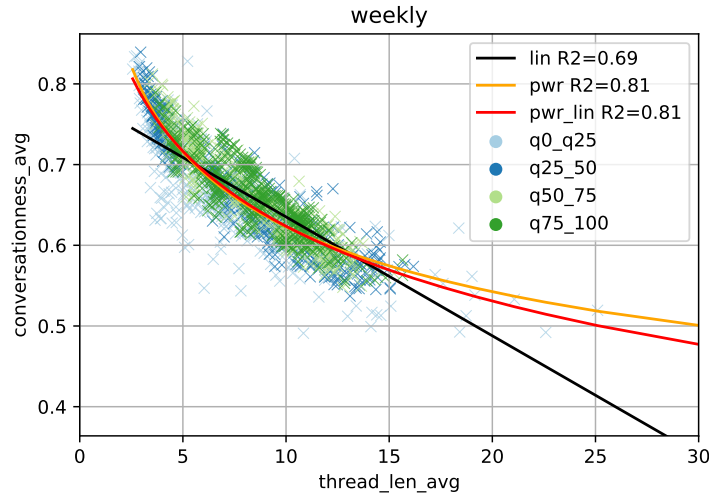


Figure C.3: Thread Len × Conversationness. The average conversationness across communities on a weekly aggregate. Larger thread length correspond to a decreased conversationness indicating that discussions are of mixed nature in number of participants.

> 0.5 highlight that creating content is the more popular activity.

We observe the ratio to converge to ≈ 0.5 , in particular for larger communities. There are, however, subtle differences. For 51% of the especially smaller cities, content creation is the more popular activity, expressed by a ratio of ≥ 0.5 . In opposition, larger communities tend to prefer voting on content. However, the two biggest cities responsible for overall activity converge to equal popularity at a ratios of 0.5083 and 0.5074, respectively. We find similar results if we only concentrate on Phase III starting in April (not shown).

This result may be surprising as content voting in online platforms may be considered as even more popular—however, this does not hold true for Jodel in the KSA.

Be taking another look into happyratios w.r.t. community size, we show a detailed scatter plot in Figure C.4. While the x-axis denotes the amount of interactions per SA community, the y-axis describes the happyratio. We generally observe that happyratios largely revolve around the average with only few outliers; however, please note the log y-axis. Our curve fitting does not resemble meaningful results, yet it becomes very apparent that the ratio between up- and downvotes does not seem to be correlated to community size in terms of interactions.

Findings Users in larger communities are more likely to start new threads in comparison to smaller communities, although there already is a wide range of content available to them.

C.2.3.4 Community Vote Consensus

Finally, we want to provide deeper insights into the community voting behavior as is represents a vital factor for content appreciation and distributed moderation. Previously in Section ??, we learned that overall community interactions are almost equally shared between creating content and voting. Further, due to their exposure, especially threads are very likely to be upvoted; thus, cumulative scores are largely equal or above 0. Yet, actual cumulative scores follow a power law being heavy tailed across posts (not shown). That is many posts may only receive few votes—if any; only few will receive exceptional scores, promoted by the app's *loudest* feed.

Given our observations of community voting behavior, we are further interested in homogeneity. To what extent do users agree on dis- and liking contents, therefore do we find controversies in steering the communities? As a measurement for vote consensus of a post,

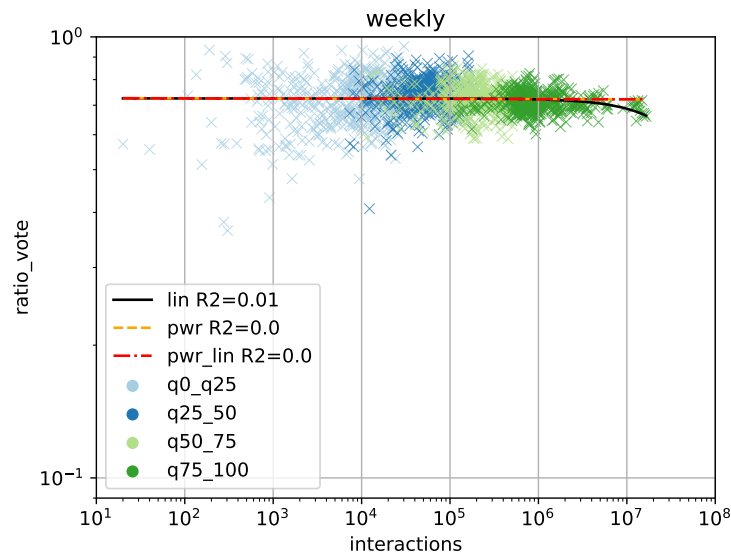


Figure C.4: Happyratio (#Up- vs. #Votes), received. This plot shows #upvotes to total votes by city size. Overall, positive votes are dominating. While being noisy for smaller cities, bigger communities tend to be more critical in their votes.

depending on which interactions are more dominant, we map the downvotes to overall votes to -1.0 and upvotes to overall votes to $0..1$. I.e., values converging towards -1 (1) denote dominating downvotes (upvotes); whereas both figures are equally cast around values of 0 . We provide a scatter plot of per city weekly consensus averages across communities in Figure C.5. Note that we filtered out posts without votes. The x-axis describes the amount of weekly community interactions, whereas the y-axis denotes the vote-consensus score as described; color represents community quantile.

As expected, due to the predominance of upvotes, consensus scores are widely on the positive side between 0.0 and 0.6 . I.e., about 70% (90%) upvotes may represent a consensus of 0.3 (0.6). While we observe heavier skews in weekly averages for $q25-50$, there is an overall trend to increasing consensus scores in larger communities—rendering them more homogeneous. For reference, the distribution fits a shifted power-law with R^2 scores of 0.23 .

Findings *i)* Communities that have reached a certain critical user base show a lively behavior in terms of response times to posts—on average. Within larger communities, people experience a reply to their post within 10 to 20 minutes on average, while smaller communities reply orders of magnitude slower. A light correlation remarks a natural *scaling effect* towards lower response times with increasing community activity. *ii)* *Invariant* of community size, longer discussions naturally attract more participants. *iii)* This is also reflected in vote consensus scores on average largely being within the same positive regime—*invariant* to community size. We find less variation in consensus values for larger communities that also tend to be more homogeneous.

C.2.4 Modeling Community-User Activity

To highlight the rapid growth of Jodel within the KSA, we next analyze and model driving and resulting factors over time proxied through the amount of registrations and interactions per community member. We re-use discussed partitions from before (cf. [♦ C.2.1: Partitioning the Communities by Rank](#)).

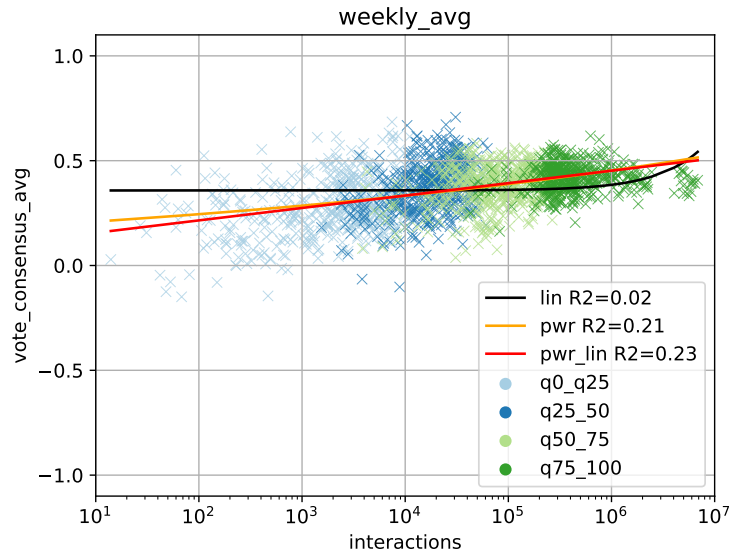


Figure C.5: Avg Vote Consensus received per city aggregated by week. While the average vote consensus varies for smaller communities, it converges towards values of 0.5, i.e., 50% more upvotes. A power law approximates this distribution at R2 scores of 0.23.

C.2.4.1 Registrations

In Figure C.6a, we show the average number of new users per day (solid lines) and their standard deviation (shaded background) for groups of cities ranked by their number of system interactions (activity). The figure shows four groups as percentiles according to community interaction volume. While the distribution varies heavily specially within Phase I & II as indicated by the standard deviation, interestingly, all communities—irrespective of their size—show the same adoption pattern. However, at steady state in Phase III, the registrations are *scaled* by activity—an unexpected finding given the independence of the communities. That is, the 3-phase adoption occurred at all communities simultaneously scaled by activity. Here, the most active cities are characterized by the largest influx of new users, while the least active ones have the lowest influx. The observation that all communities simultaneously followed the same adoption pattern supports our hypothesis that the adoption is triggered by external stimuli to simultaneously reach users nationwide. Notably, the smaller communities experienced a larger influx of new users in phase II than larger communities indicating that more users were motivated to start using Jodel relative to their overall size. Thus, the sudden adoption in March 2017 (Phase II) is characterized by a country-wide adoption of Jodel where each community experienced a substantial influx of new users. Afterwards, in Phase III, each community experiences a rather constant influx of new users that only differs in absolute numbers relative to the size of the community.

To provide more evidence of this scaling effect in Phase III, we present cumulated registrations across weekly community interactions as a proxy for activity in Figure C.6b. This scatter plot's x-axis denotes weekly interactions per community, while the y-axis accounts for registrations. Data point colors denote a city's quantile. We observe a correlation between the number of newly registered users and the overall activity of each community. That is, with increasing community activity, the communities also enjoy more and more new users onboarding. The logarithmic representation may be deceiving in variability; in absolute terms, the standard deviation is almost as high as the mean registrations, regardless of aggregation time period (daily, weekly). I.e., the std normalized by mean results in a peak ratio of 1.97 for the q75-100 quantile, whereas we find values of 0.76 to 0.97 for smaller communities on average. Out of our fitted curves, all model the observed correlation at R2

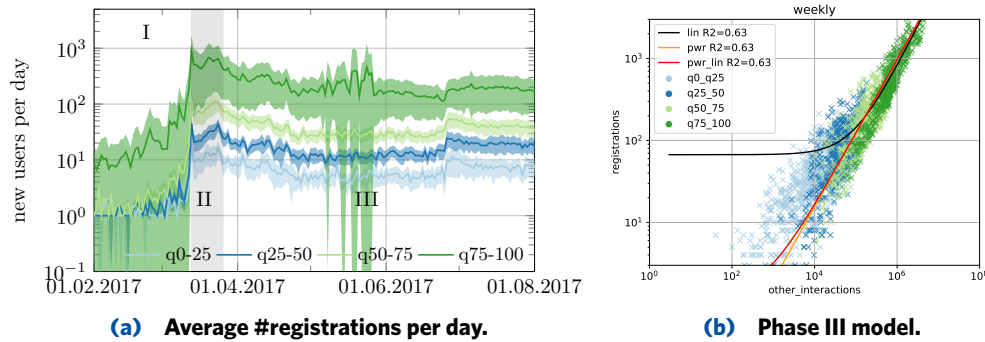


Figure C.6: Registration over time by city subgroup; modeling registrations to weekly city-interactions. (a) This figure describes average \pm std #new user registrations per day per city quantile. (b) This scatterplot denotes the correlation between new user registrations and other city interaction on weekly aggregates. A power-law fitting yields R2 scores of 0.63, indicated by the linear distribution on the log-log plane.

scores of 0.63. However, qualitatively, we would prefer power-law as it models the few interactions regime better.

C.2.4.2 Interactions

Before we focus the user perspective in the upcoming subsection, we already shift towards measures about users. As we observe a steady influx of new users in especially within steady state Phase III, we wonder whether the amount of platform interactions per user follows the same pattern as the registrations—answering whether only more users of larger communities are responsible for more contributions, or if there are self-reinforcing effects at play as well—the rich get richer.

That is, we start with the amount of community interactions over time in Figure C.7a. The x-axis show time, whereas the y-axis denotes the number of interactions per user. We plot the average amount (solid lines) and the corresponding standard deviation (shaded background) for all community quantiles. As observed for registrations, we largely find a qualitatively equal picture. As seen before in [C.5: Adoption of Jodel in the KSA](#), except for few larger communities, there is only little activity in Phase I & II. Similar to registrations, the volume of interactions over time remains stable and the quantiles equally tear apart with reasonable margins.

For better understanding this scaling effect between community size and per user interactions, we next provide a scatter plot showing community averages over time across both dimensions in Figure C.7b. The x-axis represents the amount of weekly interactions per community; the color denotes the community quantile. On the y-axis, we show the amount of interactions per user. Note that we discuss the (power-law) per user interaction distributions in more detail later (Section C.3.1). We again observe a clear correlation between both dimensions with only little variation, which is very similar to our findings in registrations. This distribution can be modeled exceptionally well with all our fitting approaches with R2 scores of 0.98; however, the lower regime is qualitatively represented best with a shifted power-law.

Findings *i+ii*) Geographically, *invariant* to communities size, observed registrations and interactions follow the same qualitative behavior. Surprisingly, the massive influx of new users in March (phase II) occurred in simultaneously nationwide; all communities first peak in registrations and interaction peaks follow two weeks later. This supports the earlier hypothesis that the adoption in phase II is likely triggered externally. Amounts of registrations and per user activity follows a power-law over the total community activity, a *scaling effect*.

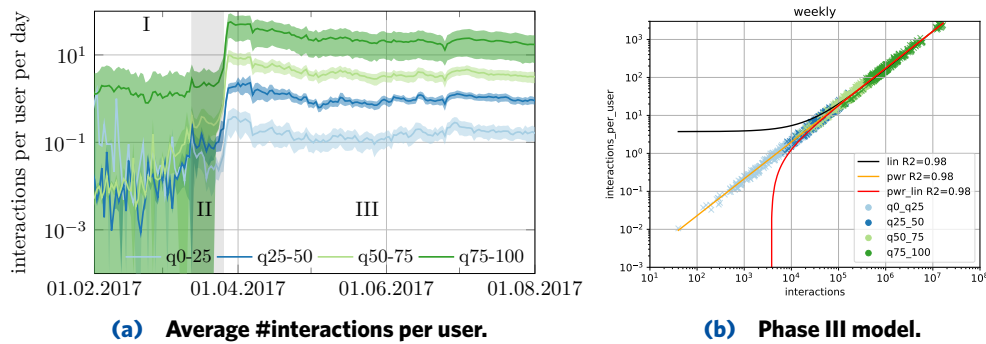


Figure C.7: Interactions per user over time by city subgroup; modeling interactions per user to weekly city-interactions. (a) This figure describes average \pm std interactions per user per day per city quantile. (b) This scatterplot denotes the correlation between per user interactions and total city interaction on weekly aggregates. A power-law fitting yields R2 scores of 0.98, indicated by the strong linear distribution on the log-log plane.

C.2.4.3 Voting

User Popular vs. Unpopular Content. First, we are interested in popular and unpopular content and the main drivers; a user-centric view. We define *unpopular* content by using the application's initial threshold for posts not being displayed anymore at a cumulative vote score of -5 and below. Further, we define *popular* content by any vote score equal to or above a particular day's all-posts' 90% percentile vote score individual to each city to keep our measure invariant to the community size (note that other thresholds lead to similar results).

Figure C.8a depicts the amount of popular posts per user across our subsets of cities normed by the amount of posts in each city at a given point in time. While the x-axis describes the temporal dimension, the log-scaled y-axis denotes the amount of popular posts per user. The shaded areas denote the standard deviation within each city subset. We observe quite noisy measures within Phase I due to the low usage. The pattern stabilizes in Phase II with the high influx of new users. At the same time, the amount of popular posts per user decreases drastically, which then remains on that level in Phase III. Within this phase, we observe a stable ratio of popular posts that does not vary over time. Qualitatively, this development does not differ across community size, whereas the smaller communities tend to be slightly more positive—still, the results are very similar across all community sets.

Next, we show the amount of unpopular content likewise in Figure C.8b normed by the amount of posts in each city at a given point in time. Again, while the x-axis describes the temporal dimension, the log-scaled y-axis denotes the amount of unpopular posts per user. The shaded areas denote the standard deviation within each city subset. While the standard deviation is higher than for the popular posts, the absolute amount of posts we consider unpopular is higher than for the popular posts, but inherently relies on our choice of our applied popular-threshold (top 90% percentile). As seen for popular content, we observe a rather noisy pattern in Phase I, which is also released by Phase II. Here, a gradual decrease in unpopular posts occurs with time across all community sizes until June 2017, rising slightly again afterwards—a similarity.

Findings The amount of popular and unpopular content per user reaches an equilibrium quickly after the jump-start of the communities. There are no significant qualitative differences across regions or city sizes. Less than one out of ten posts is considered unpopular.

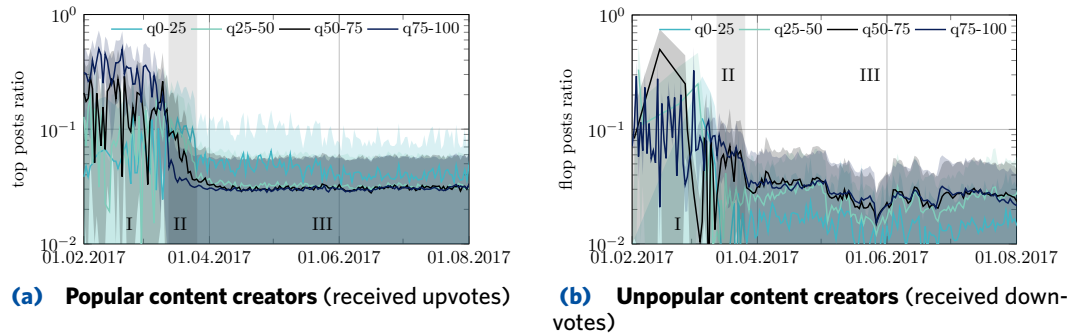


Figure C.8: Un- and popular content per user over time. After a noisy start in Phase I and II, the amount of top and flop posts per user both tune in to a steady level in Phase III.

C.3 A User-centric View

After having investigated the adoption pattern and provided an empirical overview of interlinked community interactions. However, up to this point, we are missing the important complementary user perspective. Thus, we next set out to characterize user behavior w.r.t. different communities in detail, and discuss metrics capturing the app's key design features.

C.3.1 User Interactions

First, we want to clear the stage by showing per user platform interactions as a cumulative distribution function (CDF) in Figure C.9a. While the logarithmic x-axis denotes the amount of total user interactions, the y-axis represents the accumulative fraction of users. We plotted the CDF for each community quantile. About 10% of all users only opened the app (registration event) and never actively participated. We further find that depending on community size, 50% within larger to 70% within smaller communities, the users only interacted up to 100 times with the platform. The CDF indicates a power-law distribution, which a linear shifted power-law curve fits very well with R2 scores of 0.98—however, the heavy tail experiences a drop-off not being modeled well (now shown).

Findings *Invariant* to community size, we find the wide majority of users browsing rather casually. Only few power users contribute overbalanced.

C.3.2 Anonymity - Absent Social Ties

Social Networks have been shown over and over again to form small-world connection graphs, i.e., a user probably knows at least some friends of friends while keeping an overall small graph diameter. Such social graph depend on social structure and ultimately some kind of user profile reflecting social credit. This however is not possible in a completely anonymous environment like Jodel.

To show that communication on Jodel is very ephemeral, we determined how often users encounter each other by replying on one's thread. We show our results as a complementary cumulative distribution function (CCDF) across encounters distinguished by community quantile in Figure C.9b. The logarithmic x-axis denotes encounters, i.e., how often users re-interact with each other within any other thread. On average, about 83% (10%) of all encounters remain unique (re-encountered once*). For more encounters, we observe an exponential decrease in occurrences, such that two (three) encounters happen across in 10% (3%) of all encounters. However, due to their volume, q75-100 values introduces a skew towards higher values dat 84% (10%) unique* encounters. We observe a shift of fewer encounters with increasing community activity. That is, in smaller q0-25 communities, only 62% (17%) encounters remain unique*.

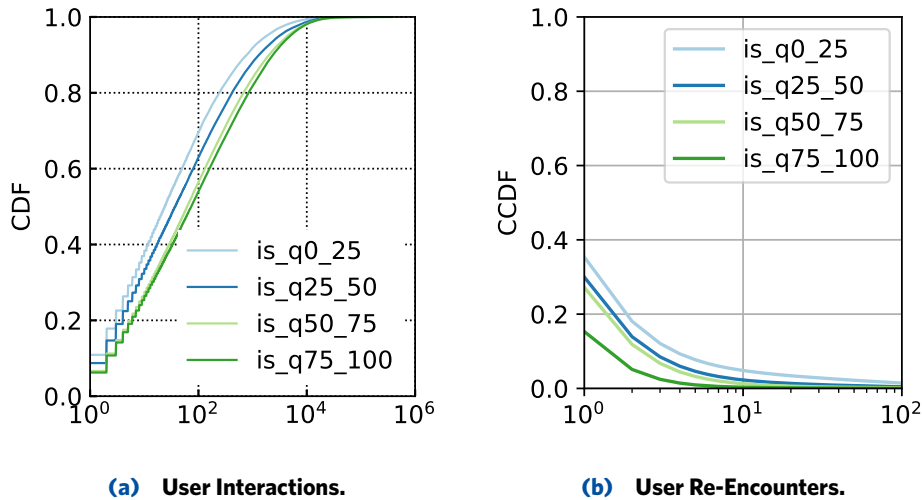


Figure C.9: User Interactions and Anonymity in a nutshell. (a) Interactions per user are power-law distributed. Most users only rarely participate, fewer others engage the platform more often. (b) Most user encounters across threads are unique; thus re-seeing another user is unlikely with exponentially decreasing probability, multi-encounters become less likely in larger communities.

Findings We find a natural *scaling effect* in increasing community activity leading to users being less likely to interact with the same person ever again—rendering platform communication rather ephemeral.

C.3.3 Hyperlocality - User Communities

While we have previously elaborated on anonymity as a central app design feature, we now want to shed light on the other property of hyperlocality. That is, we explore the amount of communities a user interacts with, and further evaluate to which extent users focus their content on their favorite community.

We show cumulative distribution functions (CDFs) for the amount of communities a user has ever interacted with in Figure C.10a, distinguished by the user's home community quantile. While we do not observe considerable differences across community sizes, most users stick to a single community. The amount of communities per user rapidly decreases, such that about 75% of all users participate in up to two communities.

To provide a deeper insight to which extent users distribute their activity across communities, we additionally measure the fraction of a user's home community (Top1) as a CDF in Figure C.10b. As seen before, depending on community size, 40-60% of all users participate only in a single community. However, users having participated in more than one community still focus on their home community, i.e., about 20% of users distribute less than 25% across others. Further, we observe that users in larger communities tend to share more interactions on their home community; this might be a result of reported hints to use fake-GPS for joining Jodel in larger communities being believed to be *better*.

Takeaway Most users participate only in a single community. Those with multiple communities still focus on their home-community.

C.4 Conclusions

In this section, we empirically characterize the nation-scale of Jodel within Saudi Arabia. The location-based nature of Jodel forming hundreds of *independent* communities throughout a complete country enables us to compare their user behavior.

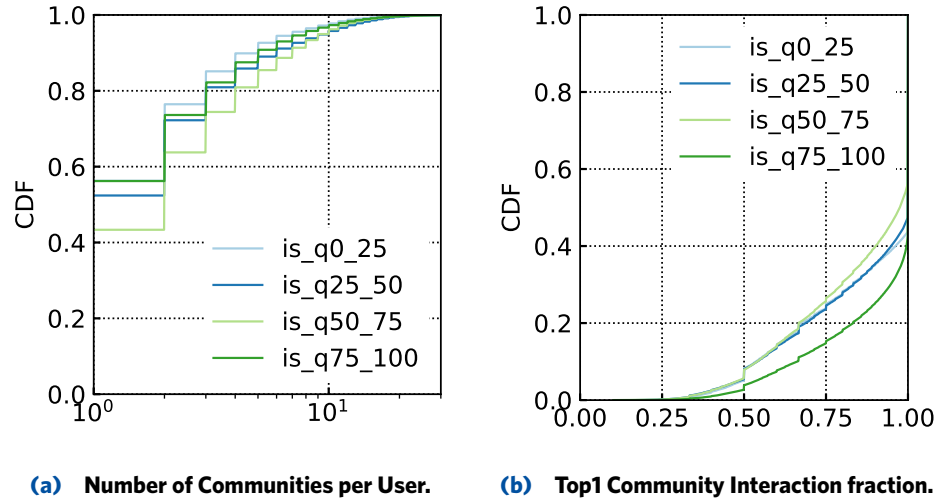


Figure C.10: Hyperlocality in a nutshell. (a) Most users only ever participate in a single community with a steep decrease. Observations remain similar across community sizes. (b) Users having participated in multiple communities still focus on a single home community. Users of larger communities tend to share more interactions on their home community.

By comparing these communities w.r.t. interaction volume (size), we identify similarities, (power-law) scaling effects in community size and rare differences. However, we identify scaling effects: larger communities attract more users to be active on a daily basis. Also, independent of community size, the observed amount of un- and popular content as well as the ratio of upvotes (happyratio) is similar across all city sizes. Social credit is granted within minutes in larger communities (reply to a post) while being orders of magnitude slower in smaller cities, scaling with size. We further identify that content voting popularity differs between the city subsets: users in larger communities are more likely to start new threads in comparison to smaller communities, although there already is a substantial amount of content available to them. While we find similarities between the community sizes in user lifetime and retention, regardless of community size, positive reactions correlate with a user's lifetime and her number of interactions. Yet, invariant to their size, all communities develop a stable daily active user base with more than 60% of the users keeping using the app until the end of our observation on average.

In future work, it would be interesting to derive whether these adoption effects are controllable, i.e., can they be applied to a new country?

Chapter Summary

We discussed a [★ A\) Structural Characterization](#) of the Jodel platform across German and Saudi communities, dissecting the plethora of emergent communities by interaction volume. We subsequently detail striking differences identified within the structural and empirical characterization in depth via a [★ B\) Cross-Country Comparison](#). Specifically given the outstanding adoption process as showcased in [C.5](#), and significant differences in usage patterns compared to DE, we chose to spotlight community internals w.r.t. size and time for the Jodel SA landscape, i.e., provide a [★ C\) Detailed Empirical Analysis KSA](#). We believe that many worldwide regions are underrepresented in today's literature.

This closes our chapter about user interactions. While the dataset enables countless future investigations and experiment, we believe that our data-driven ground truth empirical findings will provide meaningful insights to understanding our day-to-day communication tools.



USER CONTENT

In a nutshell. Besides adoption patterns and user interactions, we are next interested in actual user contents. We explore **Information Diffusion** across Jodel's German landscape of spatially distinct hyperlocal 📍 communities *leveraging hashtags*. By employing various common *spatial* metrics w.r.t. hashtag propagation and extending them by a *temporal* dimension, we identify different **classes of hashtags**. We analyze community interactions and identify the *larger communities as propagation hubs*, that allow for local diffusion into spatially local smaller communities. Further, we showcase a machine learning algorithm that is capable of classifying hashtags very well.

Given the particularities of the adoption pattern in Jodel in Saudi Arabia and its regional underrepresentation in research, we set out to determine **actual contents** of Jodel communities in SA. According to Amnesty International, there are political tensions w.r.t. freedom of speech; though e.g., the labor market has been opened for women, people are often still facing a rather conservative society; thus, platform anonymity 🛡️ may play a bigger role in this instance. Anonymous platforms have shown to promote open communication even about taboo topics, e.g., within confession boards [Birnholtz et al., 2015], or Jodel [Larsen, 2020]. We develop a classification scheme distinguishing between **intents** (*why*) and **topics** (*what*) of messages. A series of **crowdsourcing campaigns** reveals that the Saudi users love talking about **People & Relations**, *themselves*; they seek (*local*) *information and for contacts*. On the contrary to the hypothesis w.r.t. anonymity, though we also identify taboo topics, the platform has *not* become extremist, toxic, or political---as *boredom and entertainment* also has been identified as the most common reasons to use Jodel in a survey study with German-speaking users.

Lastly, we take a deep dive into very vivid **Emoji usage on Jodel**, and how to leverage this emoji-richness for Natural Language Processing (NLP) tasks. **Empirical insights** for the Saudi and German user base show that emoji usage focuses on *Smileys & Emotion*; in particular, the usage of specific emoji is heavy-tailed; they are often used to enrich and connotate text messages. We train a classical **Word-Emoji embeddings** on our Social Media dataset and showcase qualitatively and quantitatively that the model **captures semantic associations well**: Emoji to Text, Text to Emoji, Emoji to Emoji. However, we also show that due to the inherent incapability for also capturing contexts word such embeddings, the model faces potential problems. Nonetheless, in the advent of *explainable Artificial Intelligence*, we advance on classical word embeddings and add **interpretability**. By applying an adjusted

transformation via the *POLAR-framework*, we leverage **semantic word and emoji differentials** that create a measurable vector space. Our extensive crowdsourcing evaluation campaign provides evidence that the POLAR approach identifies best suited differentials and provides scale measures well **in line with human judgement**. We specifically show that the expressive space added by emoji can also enrich interpretability.

CONTENT

| | | |
|----------|---|------------|
| A | Information Spreading along Hashtags | 116 |
| A.1 | Introduction | 116 |
| A.1.1 | Research Questions | 117 |
| A.1.2 | Approach | 117 |
| A.1.3 | Results | 117 |
| A.2 | Related Work | 117 |
| A.3 | Dataset Description and Statistics | 118 |
| A.4 | Jodel DE Hashtag Usage and Spread | 119 |
| A.4.1 | Overall Hashtag Use | 119 |
| A.4.2 | Spatial Properties of Jodel Hashtags | 119 |
| A.4.3 | Temporal Properties of Jodel Hashtags | 123 |
| A.4.4 | Spatial vs. temporal dimensions | 124 |
| A.4.5 | Influence and Similarity of Cities | 125 |
| A.5 | Jodel DE Hashtag Classification | 126 |
| A.5.1 | Hashtag Content Categories | 127 |
| A.5.2 | Features | 127 |
| A.5.3 | Classifiers and Results | 127 |
| A.6 | Conclusions | 128 |
| B | Anonymous Messaging Contents | 130 |
| B.1 | Introduction | 130 |
| B.1.1 | Research Questions | 131 |
| B.1.2 | Approach | 131 |
| B.1.3 | Results | 131 |
| B.2 | Related Work | 131 |
| B.3 | Content Classification Schema | 132 |
| B.3.1 | Design objectives and development | 132 |
| B.3.2 | Iterative schema development | 133 |
| B.3.3 | Implementation | 133 |
| B.4 | Classification Campaign | 134 |
| B.4.1 | Study Design | 134 |
| B.4.2 | Campaign and Schema Quality Evaluation | 135 |
| B.5 | What Jodel Users talk about in SA | 136 |
| B.5.1 | Countrywide Perspective on Jodel Content | 136 |
| B.5.2 | Intents \mathcal{I} | 137 |
| B.5.3 | Topics Θ | 137 |
| B.5.4 | Intents $\mathcal{I} \times \Theta$ Topics | 137 |
| B.5.5 | City-Level Perspective on Jodel Content | 138 |
| B.5.6 | Reactions upon Content by Jodel Users | 140 |
| B.5.7 | How to Not Scale Out | 142 |

| | | |
|----------|--|------------|
| B.6 | Jodel SA Hashtags | 142 |
| B.6.1 | Qualitative Hashtag Classification | 142 |
| B.6.2 | Findings | 143 |
| B.6.3 | Selected Taboo Picks | 143 |
| B.6.4 | The Story of Dating and Happy Marriages. | 143 |
| B.6.5 | Findings | 144 |
| B.7 | Conclusions | 144 |
| C | The Role of Emoji | 145 |
| C.1 | Social Media Emoji Usage | 147 |
| C.1.1 | Emoji Usage by Type (Unicode Group) | 147 |
| C.1.2 | Emoji Popularity | 147 |
| C.1.3 | Emoji per Post | 148 |
| C.1.4 | Emoji Text Position | 149 |
| C.1.5 | Emoji Skin Tones | 149 |
| C.2 | Making Sense out of Emoji | 152 |
| C.2.1 | Introduction | 152 |
| C.2.2 | The Jodel Emoji Embedding Dataset (JEED1488) [SD1] | 153 |
| C.2.3 | Related Work | 153 |
| C.2.4 | Word-Emoji Embeddings | 155 |
| C.2.5 | Emoji2Emoji Associations | 155 |
| C.2.6 | Emoji2Text Associations | 158 |
| C.2.7 | Text2Emoji Associations | 160 |
| C.2.8 | Future Work | 164 |
| C.2.9 | Conclusions | 164 |
| C.3 | Interpreting Emoji | 165 |
| C.3.1 | Introduction | 165 |
| C.3.2 | Related Work | 167 |
| C.3.3 | Creating Interpretable Embeddings | 168 |
| C.3.4 | Embedding and Polarization | 170 |
| C.3.5 | Human Evaluation | 171 |
| C.3.6 | Conclusions | 175 |

Introduction

Up to this point, we have discussed adoption pattern and general user interactions on a meta perspective in terms of posting, replying, for voting upon such items. We next are interested in content characterization and its implications to Jodel as a platform, given its design properties.

However, a large body of research across a plethora of different platforms provided deep insights into various aspects of **◆ User Content** as discussed in **◆ D: Social Network Analysis**. By staying on a meta perspective, research has leveraged the coupling of content to social structures forming graphs. This approach allows of (epidemically) information spreading and diffusion modeling through social networks [Dow et al., 2013, Cannarella and Spechler, 2014, Yan et al., 2013, Woo and Chen, 2016, Matsubara et al., 2017, Kamath et al., 2013], and other online platforms, such as Twitter [Romero et al.,], or Youtube [Brodersen et al., 2012, Xu et al.,]. Researchers have even applied cross-platform tracking for investigating information spreading across the web w.r.t., e.g., memes [Zannettou et al., 2017].

Content diffusion through the independent Jodel communities. While most analyzed platforms a globally accessible, research suggests that still spatial local clusters exist among platform users, information spreading characteristics induced by hyperlocality **📍** as found in Jodel remains unknown. That is, we explore **★ A) Information Spreading** represented by hashtags across the Jodel platform in Germany. For easing comparison, we employ metrics as used for Twitter in [Kamath et al., 2013], also a microblogging platform, but globally accessible and non-anonymous, or Youtube [Brodersen et al., 2012]. By extending these metrics with a temporal dimension, we provide deep insights into how local communities are linked: Smaller cities around the heavy user bases tend to be highly influenced. We identify various hashtag types across the spatial and temporal dimension, provide examples, and showcase a classification approach with machine learning.

In-depth message content analysis - intents (why) and topics (what). Digging deeper, we are further interested in actual messaging contents. While related work has revealed rich insights on various platforms that a similar to Jodel to at least some extent, other elaborated on why users might opt for anonymous platforms in qualitative studies [Kang et al., 2016]. Given the sudden adoption process in Saudi Arabia (see **C**) and political circumstances and Jodel possibly empowering its users to escape from a conservative environment enjoying freedom of speech in the veil of anonymity **🔒**. In **★ B) Message Contents**, we provide our results of an in-depth human crowdsourced content evaluation. Inspired by [Kang et al., 2016] and [Paul et al., 2011, Correa et al., 2015], we developed a rich classification scheme that distinguishes between message *intents* (why) and *topics* (what) performing at substantial expert coder agreement. By performing a series of classification campaigns, we provide insights into the communication that is happening within Saudi Arabia w.r.t. what people talk about, and why they might do so.

Broadening our perspective, leveraging day-to-day social media emoji usage. Finally, we would like to highlight the popular topic of **★ C) The Role of Emoji** which has acquired the driver's seat within nowadays casual day-to-day communication abbreviating objects, edibles, weather, animals, persons, professions, sports, symbols, and lots more. However, their emergence and most important function might arguably be portraying certain emotions as they have shown to provide valuable information for, e.g., sentiment analysis [Berengueres and Castro, 2017]. Further, they enable transporting salient cues in writer's (self-)identification or interpretation [Barbieri and Camacho-Collados, 2018, Robertson et al., 2018, Robertson et al., 2021b].

We Set out providing empirical community insights to emoji usage within the German and Saudi user bases in [★ C.1\) Social Media Emoji Usage](#) . Further, neural embeddings as being used for text have shown to unlock improved downstream tasks by adding emoji [Hu et al., 2017a, Felbo et al., 2017c]. We provide qualitative and quantitative insights into a Word-Emoji embedding trained on our Jodel dataset within Germany applying Word2Vec [Mikolov et al., 2013a] in [★ C.2\) Making Sense out of Emoji](#) . Resulting embeddings reflect semantic associations well as we identify countless reasonable associations; while uncertainties arises for words having multiple meanings, or may have created a specific platform understanding. Lastly, we leverage semantic differentials specifically incorporating emoji to transform a Word-Emoji embedding into an interpretable counterpart in the advent of explainable AI in [★ C.3\) Interpreting Emoji](#) . While creating explainable embeddings are usually learned end-to-end [Subramanian et al., 2018a] relying on sparsity [Panigrahi et al., 2019], we opted for a different approach leveraging semantic differentials. By adjusting the POLAR [Mathew et al., 2020] framework, we employ rather simple linear projections for adding interpretability to an input embedding space. By conducting various crowdsourcing campaigns and test setups, we provide evidence that this approach performs well in line with human judgement. It selects suitable dimensions and assigns reasonable scale measures on selected differential scales. I.e., emoji can improve interpretability of POLAR embeddings, specifically in interpreting emoji themselves.

A Information Spreading along Hashtags












In this section, we study the usage and propagation of hashtags on Jodel, that is entirely anonymous, but more importantly hyperlocal, i.e., only showing nearby posted content. The platform thereby forms hundreds of local communities and opens the question of how information propagates within and between these communities. We tackle this question by applying established metrics for Twitter hashtags to our ground-truth data set of Jodel posts within Germany that spans three years. We find the usage of hashtags in Jodel to differ from Twitter; despite embracing local communication in its design, Jodel hashtags are mostly used country-wide.

A.1 Introduction

Social media has become a popular and ubiquitous tool for consuming and sharing digital content (e.g., textual or multimedia). This sharing leads to information propagation and spreading across users and even across different networks [Zannettou et al., 2017]. Understanding this propagation has thus motivated research studies to investigate the dynamics of information adoption, spreading, and (complex) contagion of information [Kamath et al., 2013, Cannarella and Spechler, 2014, Yan et al., 2013, Woo and Chen, 2016, Ferrara et al., Sanli and Lambiotte, 2015, Romero et al., Dow et al., 2013], e.g., in the form of memes. A widely studied platform in this regard is the microblogging service Twitter that enables users to reach a global audience and for which sampled post data is available via APIs. Analyzing the post contents' (e.g., included memes) is, however, a very challenging application of natural language processing. Since users often self-classify their posts by adding hashtags to ease retrieval, analyzing hashtags is a promising proxy measure for analyzing memes or post contents. This has resulted in metrics to analyze hashtags and thereby valuable insights into their spreading behavior [Kamath et al., 2013].

Unlike Twitter and other classical social media platforms, Jodel *i)* does not have user profiles rendering user to user communication anonymous, and more importantly *ii)* displays content only in the proximity of the user's location, thereby forming local independent communities. Despite the emerging use of such platforms, little is known on how their key design differences impact information propagation compared to global platforms.

Structure [JH9]

-  [A.2: Related Work](#)
-  [A.3: Dataset Description and Statistics](#)
-  [A.4: Jodel DE Hashtag Usage and Spread](#)
 -  [A.4.1: Overall Hashtag Use](#)
 -  [A.4.2: Spatial Properties of Jodel Hashtags](#)
 -  [A.4.3: Temporal Properties of Jodel Hashtags](#)
 -  [A.4.4: Spatial vs. temporal dimensions](#)
 -  [A.4.5: Influence and Similarity of Cities](#)
-  [A.5: Jodel DE Hashtag Classification](#)
 -  [A.5.1: Hashtag Content Categories](#)
 -  [A.5.2: Features](#)
 -  [A.5.3: Classifiers and Results](#)
-  [A.6: Conclusions](#)

A.1.1 Research Questions

In this section, we present the first study on information spreading in such an emerging platform by investigating the hashtag propagation in Jodel as a prominent application in this space. We take a detailed look on hashtag propagation through the lens of a platform operator by having the unique opportunity to analyze data provided by Jodel for messages posted in Germany from September 2014 to August 2017. This longitudinal data set enables us to study how this key design pattern of forming local communities by only displaying content to nearby users influences the hashtag usage and compares to the global counterpart Twitter.

A.1.2 Approach

Our study is based on using established metrics designed to capture the spatial focus and spread of Twitter hashtags [Kamath et al., 2013] to Jodel, as seen for Youtube Videos [Brodersen et al., 2012]. We show that these metrics can be applied to the temporal dimension to cover the spread of hashtags in time, enabled by our longitudinal observation period. We further study similarities in hashtag usage between cities and their spacial impact—finding that larger cities/communities influence the smaller ones. The correlation of spatial and temporal metrics reveal that hashtags can be grouped into four different hashtag classes distinguished by their spatial and temporal extent. In the last step we show that these groups are distinguishable by machine learning models, informed by manual labeling of 450 most frequently used hashtags.

A.1.3 Results

We find that Jodel’s popular hashtags are used country-wide, whereas less popular hashtags tend to be more local. We show that classical metrics capturing the spatial propagation can be applied to the temporal domain. By applying these metrics, we see that popular hashtags are used over the long-run, while less popular hashtags tend to be more short-lived. We show that the used hashtags can be grouped into four classes by their spatial and temporal extent. We further show that these four groups can be learned by statistical models with high accuracy, based on comparing five different classifiers (k-nearest neighbour, regression trees, naive bayes, LDA, ZeroR). Thus, statistical methods can distinguish between different meme types found in Jodel.

A.2 Related Work

We identify three main areas within related research: *i)* general meme spread modelling, *ii)* the use case of microblogging, e.g., Twitter, and *iii)* others; which we will discuss next.

Spreading and contagion models. A classical approach to study information diffusion is applying spreading models. Epidemic models have been applied to memes, where a meme can *infect* people by coming in contact with it (SIR models)-possibly extended with mechanics for *recovery* (SIRS models), e.g., in [Yan et al., 2013, Woo and Chen, 2016]. Although these approaches model the growth of hashtag popularity well, most fail to map the typical power-law decay [Matsubara et al., 2017]. Their application to hashtags is further limited by requiring an infection time, i.e., when a user learns about a hashtag. Passive information consumption such as reading is typically not included in most social network data.

Twitter. The study of hashtag usage and diffusion mostly targets Twitter given its popular use of hashtags and ability to geotag posts. Although Twitter has no boundaries regarding distance (i.e., unlike Jodel), cities closer to each other share more hashtags, supported

| Metric | #Entries | Description |
|--------------|------------|---|
| Hashtag Uses | 41,038,733 | # of hashtags occurrences |
| Hashtags | 13,110,573 | # of different hashtags |
| | 11,092,360 | # of hashtags used only once |
| Messages | 26,955,008 | # of messages that contained hashtags |
| Users | 1,240,404 | # of users posting contents with hashtags |
| Locations | 6,830 | # of different posting locations/cities |

Table A.1: Hashtag Dataset Statistics. The data ranges from the application start in late 09/2014 up to beginning of 08/2017.

by an analysis of the Twitter trending topics in [Ferrara et al.,]. The authors find three clusters of hashtag similarity across the biggest cities in the US and speculate that the spread is related to airports. To study non-stationary time series of hashtag popularity, [Sanlı and Lambiotte, 2015] applies a statistical measure originally used for neuron spike trains to hashtags. It is capable of giving information on how regularly hashtags are used. They find that low to mediocre popular Twitter hashtags are on average rather bursty, while extremely popular ones are posted more regularly. The influence of content (e.g., politics, music, or sports) on the hashtag adoption is studied in [Romero et al.,]. The authors find that especially political hashtags are more likely to be adopted by a user after repeated exposure to it than hashtags of other topics.

To capture the spatio-temporal dynamics of Twitter hashtags, *focus*, *entropy*, and *spread* were proposed as metrics [Kamath et al., 2013]. By applying these metrics to Twitter, the authors find hashtags to be a global phenomenon but the distance between locations to constraint their adoption. We will use these metrics to study Jodel and we extend them with a temporal dimension within our analysis. To study the how cities impact each other regarding hashtag adoption, [Kamath et al., 2013] also proposed a spatial impact metric to capture the similarity of hashtag uses in two cities—a metric that we will adopt likewise. They show that the biggest influencers were big cities with large user bases.

Other platforms. Besides Twitter, few studies consider other platforms. The sharing cascades in Facebook are studied in [Dow et al., 2013]. Similar cascades are found by studying how the blogosphere and the news media influence each other [Leskovec et al.,]. Memes do not have to be in the form of images or text, but can also be videos—as such, e.g., [Xu et al.,] studies the diffusion of memes on Youtube, or Whisper [Cao et al., 2012].

Other works focused on the influence of events in terms of the spreading behavior. E.g., [Becker et al., , Kotsakos et al.,] used statistical classifiers on contextual features to distinguish between memes and events. Researchers have also tried to detect events, e.g., by analyzing the Twitter stream [Li et al., , Weng and Lee, 2011] and inferring where an event happens [Walther and Kaisser,]. There were also efforts to detect earthquakes and estimating the epicenter in realtime [Sakaki et al.,]. Also, user positions can be at least vaguely estimated as shown in [Chandra et al.,].

We complement these works by studying the hashtag usage and diffusion on Jodel. Its property to only display posted content to nearby users differentiates Jodel from other studied social networks that disseminate content globally (e.g., Twitter or Facebook). It thus *might*—and as we will see: *will*—feature a fundamentally different spreading behavior.

A.3 Dataset Description and Statistics

Hashtags DE. We have extracted hashtags from the message contents by applying a regular expression matching a ‘#’ followed by any amount of alphanumeric characters (including German umlauts and Eszett), dots, dashes or underscores. This resulted in a total amount of about 41 M hashtag uses within 26 M different messages and 13 M different

hashtags. These messages were created by 1.2 M users having posted in about 7 k different locations.

Within the set of hashtags, we observe that 11.1 M are only used once. This leaves about 2 M hashtags that have been used multiple times, i.e., ≥ 2 , and therefore are suited for our hashtag propagation analysis at all. After manual sample screening, the predominant reason for this huge amount of hashtags occurring only once is that on Jodel, they are often used as a unique stylistic feature, support content, or are misspelled reuses—in contrast to a self-categorization that might be expected.

A.4 Jodel DE Hashtag Usage and Spread

In this section, we analyze the spread and propagation of content in Jodel by using hashtags as a proxy measure. That is, we leverage the user's ability to tag posts with hashtags to relate to topics, add categories or metadata to posts. Although hashtags are sometimes used as a rather stylistic feature (e.g., by using numbers as hashtags to link multiple character limited posts together), more popular ones overall reasonably capture topics and memes in the posts.

We will see that some hashtags are specific to the Jodel platform and very local possibly due to its location-based design. Beginning our analysis in this Section with a study of hashtag popularity, we follow this up with their spatial and temporal spreading extent. We lastly study the hashtag usage in different cities and how they influence the hashtag adoption.

A.4.1 Overall Hashtag Use

Our data set consists of 27 M posts with hashtags. We overall find 41 M occurrences of 13 M unique hashtags of which only 2 M are used multiple times (cf. Table A.1).

A.4.1.1 Popularity

We begin by studying the hashtag popularity. Figure A.1a shows the distribution of a hashtag's occurrence (x-axis) vs. the corresponding amount of unique hashtags in our dataset (y-axis) on a log-log scale. We observe that the vast majority of hashtags are only used few times. The distribution is heavy-tailed and of similar shape, as observed in Twitter [Kamath et al., 2013].

A.4.1.2 Location distribution

We next study how many hashtags (y-axis) are used in how many locations (x-axis) in Figure A.1b. We see that not only the occurrences per hashtags is heavy-tailed, but also their geographic spread. These results are also very similar to Twitter [Kamath et al., 2013].

A.4.1.3 Findings

We find most hashtags are being used only very few times. The hashtag usage follows a heavy-tailed distribution, which also holds true for the number of different locations in which they occur. That is, only a few hashtags are heavily popular and used in many locations—others to a lesser extent, or not.

A.4.2 Spatial Properties of Jodel Hashtags

We next study spatial properties of Jodel hashtags, e.g., if a certain hashtag only occurs in a local community or over which geographic distance the usage of a countrywide hashtag

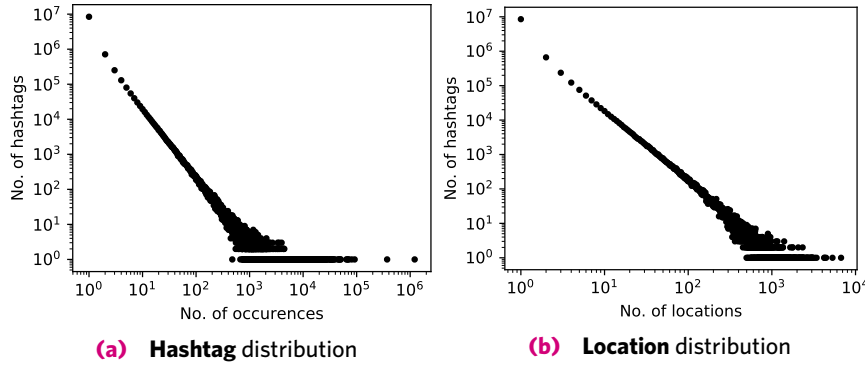


Figure A.1: Power-Law Distributions: a) the hashtag distribution w.r.t occurrences and the corresponding amount, b) the location distribution w.r.t occurrences for a hashtag and the corresponding amount---both distributions are heavy-tailed.

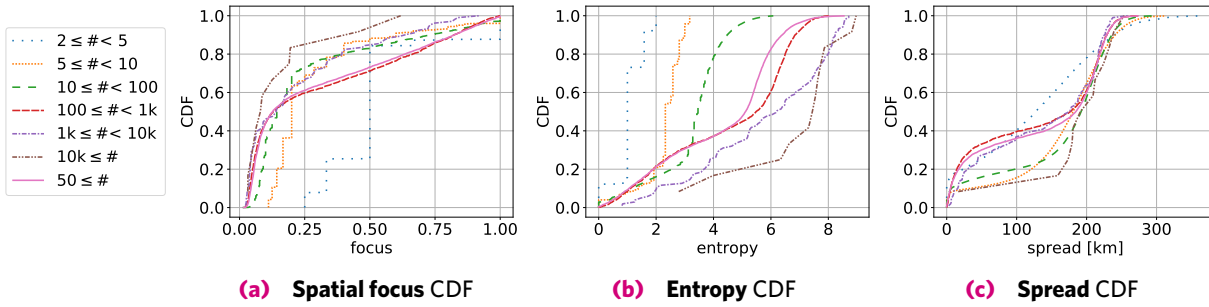


Figure A.2: Spatial hashtag metrics: focus, entropy and spread (left to right). All figures follow the partitioning by hashtag occurrences shown left. a) The more popular a hashtag is, the more unfocused it gets. Most hashtags show a very low focus, while very unpopular hashtags naturally tend to be more focused. b) Likewise, less used hashtags naturally can only be used in few locations, while more popular hashtags are used in many locations. c) Some hashtags on average span geographically only up to ± 50 km, whereas most are used all over the country.

is spread. To capture these spatial properties, we use three hashtag metrics originally proposed for Twitter: *focus*, *entropy*, and *spread* [Kamath et al., 2013]. These metrics enable us to judge if content diffusion in Jodel actually is—due to its design—indeed *more local* than a comparable microblogging platform without geographical communities, like, e.g., Twitter.

Data filtering. We restrict our set of hashtags by only considering hashtags that occurred first in 2016 or later. This way, we focus on a time in which the app has an established user base in Germany.

A.4.2.1 Focus

The focus metric captures how locally or globally (i.e., in our case countrywide) focused the use of a hashtag is [Kamath et al., 2013]. To achieve this, the set of hashtags and the set of locations are defined as H and L , respectively, of which for a given hashtag $h \in H$ and location $l \in L$, O_l^h is the set of occurrences of h in l . Then, the likelihood via this proxy measure of observing a hashtag h in a location l is defined as:

$$P_l^h = \frac{|O_l^h|}{\sum_{m \in L} |O_m^h|}$$

The focus location of a hashtag is defined as the location with most occurrences of that hashtag and further provides a fraction of the occurrences in the focus location compared to the number of overall occurrences. It is defined as $F^h = \max_{l \in L} P_l^h$. Then, the focus for

hashtag h is defined as a tuple of the focus location $l_f = F^h$ and its the likelihood via this proxy measure $P_{l_f}^h$. Hashtags only popular in a few cities will have a higher focus, whereas globally popular hashtags will have a lower one. A limitation of the focus metric is that it provides information only about one single location, but nothing about the distribution.

We show the focus distribution of hashtags in Figure A.2a, where a series represents a CDF for a set of hashtags partitioned by their occurrence. As the hashtags are subject to popularity, i.e., usage frequency, these partitions define different log-based groups within out dataset (cf. Figure A.1a). Our observation is that the focus distribution is skewed towards low focus values regardless of hashtag occurrences. That is, 60% of all hashtags that occur ≥ 5 times have a focus of ≤ 0.25 . This means that from all occurrences of such a hashtag, only 25 % occur in its most popular city, whereas the remaining 75 % of the hashtag occurrences is in other cities. Therefore, the focus distribution indicates that the usage of most hashtags is not focused on a single city but is rather spread over multiple cities. Further, the observed skew within the distributions towards low focus values differs from hashtag usage observations in Twitter in which the hashtags' focus was uniformly distributed [Kamath et al., 2013]. The prevalence of low focus values is unexpected and interesting; the design of the App to only display nearby posts could have caused a skew towards high focus values, in which the usage of most hashtags would be more concentrated. This, however, is not the case.

A.4.2.2 Entropy

Borrowed from Shannon's theory, the entropy captures in how many locations a hashtag is used [Kamath et al., 2013]. For a hashtag h , it is defined as:

$$E^h = - \sum_{l \in L} P_l^h \log_2 P_l^h$$

With the applied logarithmic basis of two, this value rounded up according to Euler defines the minimum number of bits required to represent the amount of a hashtag's locations it has spread to. The higher the diffusion of a hashtag, the higher its entropy; i.e., the entropy defines the number of locations a hashtag occurred in by the power of two. For more often used hashtags, both entropy and focus are resistant to small changes in the data (e.g., single occurrences in another ten locations).

Similar to the focus, we show the entropy distribution as CDFs for hashtags likewise partitioned by occurrences in Figure A.2b. We observe that only a negligible number of hashtags is used in a single city (entropy 0). Looking into the different partitions, we identify that less popular hashtags clearly tend to a smaller entropy. However, for the more popular hashtags having at least 50 occurrences, more than 60 % of the hashtag occurrences are in ≥ 16 cities (entropy 4). As already indicated by the focus distribution, the usage of most hashtags is thus not concentrated to a single city only but spread over multiple cities. In summary, the hashtag usage shows a trend to higher entropy values with an increased number of occurrences; the more popular a hashtag is, there more it is spread across different cities, which supports our findings for the focus.

A.4.2.3 Spread

To obtain information about the geographical expansion, we can use the spread metric defined as the mean distance of the geographic midpoint of the set of hashtag occurrences [Kamath et al., 2013]:

$$S^h = \frac{1}{|O^h|} \sum_{o \in O^h} D(o, G(O^h))$$

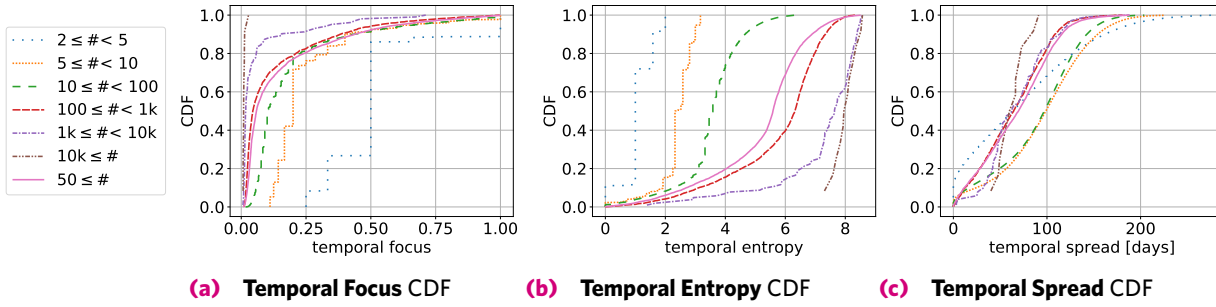


Figure A.3: Temporal hashtag metrics: temporal focus, temporal entropy, and temporal spread (left to right) are temporal adoption to the spatial counterparts (cf. Section ??). All figures follow the partitioning by hashtag occurrences shown left. *a)* The temporal focus decreases with hashtag popularity, i.e., they become used over longer time periods. *b)* This finding is supported by the temporal entropy showing that more popular hashtags are more widespread across multiple dates. *c)* The temporal spread indicates a possible distinction between a smaller set of short-lived hashtags and a large set of long-lived hashtags.

where D is the distance in kilometers and G is the weighted geographic midpoint. As on our scale (Germany), the spherical shape of the Earth is only of minor importance, we use the weighted average latitude and longitude as the midpoint. A spread of 50 km thus means that the average usage of a hashtag occurs within ± 50 km.

We show the spread distribution again as CDFs of partitions by occurrences in Figure A.2c. The distributions reveal that there are three groups of hashtags: *i)* Only rarely used hashtags (≤ 5 occurrences) show a rather linear spread, *ii)* More frequently used hashtags (5..100 occurrences) show a slight bimodal distribution as they either have a small spread up to 50 km, or most of them show a rather big spread > 150 km. The same holds true for hashtags that are heavily used. *iii)* Hashtags that are used often, but do not belong to the heavy tail, strengthen the bimodal observation as about 40 % only have an up to 50 km, whereas most others are spread wider.

We note that higher spreads are likely the value a Germany-wide hashtag may achieve. While there is no (known) comparable analysis for Twitter or similar platforms, we conclude that the lower-spread hashtags are most probably an implication of Jodel's nature building location-based communities. I.e., there are hashtags that are used in a geographically restricted area at small distances.

A.4.2.4 Findings

We observe that most hashtags in Jodel are used rather countrywide, i.e., their usage does not concentrate on single cities and spreads over larger geographic distances. This is unexpected since the design of Jodel to form local geographic communities could also result in a more geographically focused usage of hashtags. However, while most hashtags are used rather globally, up to 40 % have a local spread of ± 50 km and thus are a potential consequence of Jodels' design.

Twitter Comparison. A direct comparison to [Kamath et al., 2013] can be made within our series of hashtags at least having 50 occurrences (pink solid lines). While the *focus* CDF for Twitter hashtags is rather linear with the exception of 20 % having focus 1, the focus on Jodel is distributed in an opposite fashion. That is, 60 % of Jodel hashtags (≥ 50 occurrences) tend to be non-focused below a value of 0.25, but are likewise equally distributed above—having almost no hashtags with focus 1. As for the *entropy*, most hashtags on Twitter are used very locally, which can only be observed for least popular hashtags on Jodel—many more popular hashtags are used across the country. Similarly, the *spread* on Twitter is either local for few hashtags, but then increases linearly, which is identical for the least

and heavily popular hashtags on Jodel—others show a pronounced bimodal distribution between local and countrywide scope.

A.4.3 Temporal Properties of Jodel Hashtags

We are next interested in studying how hashtags develop over time (e.g., gain in popularity). This is possible given our longitudinal data set. Therefore, we adopted focus, entropy, and spread for our temporal analysis. Instead of locations as in our spatial analysis, we use the creation time of a hashtag's post (grouped to days for focus and entropy) for each hashtag occurrence. The grouping to days makes sense due to limited content presence within the usually highly dynamic Jodel feeds for larger communities.

A.4.3.1 Temporal Focus

We show the temporal focus distribution as CDFs partitioned by hashtag occurrences in Figure A.3a. Recall that the temporal focus now defines the probability of a hashtag to be used on its most popular day, i.e., a temporal focus of 1 indicates that a hashtag is exclusively used on a single day whereas a focus of near 0 would suggest a spread over the entire observation period. We observe that about 80 % hashtags have a low temporal focus ≤ 0.25 , suggesting that their lifetime is not focused on a single point in time. The more popular they become, the temporal focus decreases, i.e., they remain popular over time. However, least popular hashtags tend to a higher temporal focus in comparison. In summary, there are almost no hashtags focused to a single day. For those that are being used only a few times, this implicates random re-use that is probably not correlated, whereas popular hashtags are used throughout the observation period.

A.4.3.2 Temporal Entropy

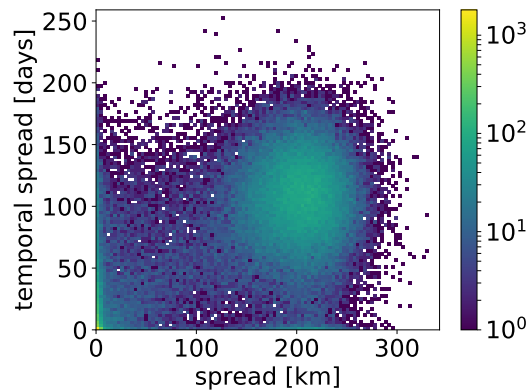
The temporal entropy defines the number of days on which a hashtag is used. We show its distribution as CDFs partitioned by hashtags occurrences in Figure A.3b. We observe that only a negligible amount of hashtags are used on exactly one day (entropy 0). Except for the only rarely used hashtags, more than 90 % occurrences have an entropy above 2, i.e., they were used on more than 4 (2^2) days. Further, the higher the occurrences (popularity) of a hashtag, the higher the entropy. This indicates that popular hashtags are used for longer time periods.

A.4.3.3 Temporal Spread

The temporal spread defines the average time period in days in which a hashtag is used. For example, a temporal spread of 50 days means that the average usage period of a hashtag is ± 50 days (past and future) from the temporal weighted midpoint. We show the distribution of the temporal spread as CDF again partitioned by hashtag's occurrences in Figure A.3c. We observe that the temporal spread is distributed equal (linear CDF) across all partitions. However, the activity period is again influenced by the popularity of a hashtag; the more popular a hashtag is, the higher is the temporal spread. The presented series that only include hashtags with very few uses depict a large set of hashtags with a temporal spread of more than 100 days—the significant skew towards a larger spread strengthens our belief that such hashtags occur independently from each other (cf. temporal focus).

A.4.3.4 Findings

Popular hashtags in Jodel are seldomly a flash in the pan but are mostly used over extended time periods. In particular, the more popular a hashtag is, the longer and frequent its usage



(a) Spatio-Temporal Heatmap. Combined spatial (x-) and temporal (y-axis) spread shown as a heatmap of hashtags occurrences (z-axis). We observe clusters: *i*) countrywide long-lived hashtags, *ii*) countrywide short-lived hashtags, and *iii*) Both, local short- and long-lived hashtags.

| | | | |
|----------|-------------|-----------------|-------------|
| temporal | long-lived | local phenomena | other memes |
| | short-lived | local event | events |
| | | local | global |
| | | spatial | |

(b) Identified Hashtag Classes on the Jodel platform according to the spatial and temporal spread metric.

Figure A.4: Correlation between Spatial and Temporal Spread. *a*) describes the number of hashtags and their spread properties in the restricted dataset. *b*) shows our derived classes of hashtags according to the spread metrics.

period becomes, whereas less popular ones rather occur independently from each other. This is interesting since the Jodel app provides—unlike Twitter—only limited functionality to search for hashtags as hashtags may only be clicked when seen in a post, i.e., for a purposeful re-use it must be known.

A.4.4 Spatial vs. temporal dimensions

Having analyzed the spatial and temporal dimensions in isolation, we are now interested in how they correlate. For example, hashtags that occur in one geographic area have a low spatial spread, but can be active over a short or longer timespan as indicated by the temporal spread. Therefore, we focus on correlating the spatial and temporal spread and omit other metrics since they provide a similar picture. Figure A.4a shows the spatial spread on the x-axis and the temporal spread on the y-axis of all hashtags having at least 30 occurrences since 2016. The hashtags can roughly be clustered into four groups as shown in Figure A.4b. *i*) A temporal spread of 100 days and a spatial spread of 250 km (long-lived and countrywide). We would expect countrywide hashtags that are statements and also memes in this group, as both kinds are often spread out on the landscape and rather long-lived. *ii*) Located around a spatial spread of 250 km, but the temporal spread is only a few days (short-lived and global). Hashtags in this group are, for example, about countrywide events. Also, some memes that are short-lived could be in that group. *iii*) Spread around 0 to 30 km and temporal spread of 0 to 70 days (long-lived and local). Here, we would expect hashtags about phenomena that are particularly local due to the community structure of Jodel. *iv*) Short-lived and local hashtags. This group can involve for example local events. We will base our content classification of hashtags in Section ?? on these identified groups.

Findings. The correlation of spatial and temporal spread clusters the hashtags into four groups, identified by long-lived vs. short-lived and countrywide/global vs. local spread.

That is, there are some long-lived and short-lived countrywide hashtags, while we also identify long- and short-lived local hashtag occurrences.

A.4.5 Influence and Similarity of Cities

We have seen that some hashtags occur rather locally, which is an essential aspect of the Jodel application. We have also seen that many hashtags spread through many Jodel communities. Therefore, we next want to examine how much communities influence each other in the sense of causing other cities to adopt a hashtag. We are particularly interested in which cities source and popularize trends before others adopt them.

A.4.5.1 Spatial impact

To get insights of on cities' impact on another, we use the spatial impact metric from [Kamath et al., 2013]. The hashtag specific spatial impact $I_{A \rightarrow B}^h$ of two cities A and B and a hashtag h is defined as a score in the range $[-1, 1]$. A score of 1 means that either all occurrences of that hashtag in city A happened before all occurrences in B , or that there are no occurrences of that hashtag in B at all. The same applies in the reverse case scoring -1 . Values around 0 indicate that both cities adopted the hashtag roughly at the same time. In short, this measure describes which city adopted a hashtag earlier, and therefore *may* have influenced the other city. The spatial impact $I_{A \rightarrow B}$ is then defined as the average hashtag's spatial impact *for all hashtags* that occur in at least one of the cities.

As an example, we compare the cities Aachen, Hamm, and Overath with the 500 most popular cities. For each of the three cities, we show the spatial impact on every of the 500 most popular cities as a histogram in Figure A.5. We chose Aachen as the birthplace of the Jodel network with a large technical university and 250 k inhabitants, Hamm as a medium-sized city without university and 180 k inhabitants, and Overath as a smaller city with 27 k inhabitants. The histograms x-axis denotes the spatial impact, while the y-axis covers the number of other cities in comparison. From the given examples, we observe that Aachen is the most influencing city within this comparison (and also on the whole platform Jodel—*not shown*), with most of its scores being between 0.5 and 1. Hamm is both influenced by cities as well as influencing other cities, whereas Overath is heavily influenced by most other cities (probably also due to a low population and therefore fewer users). By also qualitatively looking into other cities spatial impact histogram, we can only conclude that cities with a higher population impact cities with a lower population. This finding that large cities influence smaller ones is in line with observations on Twitter [Kamath et al., 2013].

We remark that the spatial impact metric does not normalize by community size and thus comparing communities of unequal size can provide an advantage in this metric to the larger city. Even if the hashtags in the big city never spread to any other city, it would still impact a small city using this measure. Nevertheless, this still supports the findings also shown for Twitter that larger cities usually have a higher impact.

A.4.5.2 Hashtag similarity

We previously have seen that cities impact each other. To understand the communities hashtags better in comparison, we use the *hashtag similarity* [Kamath et al., 2013] measure of two locations A and B as $\text{sim}(A, B) = |H_A^{50} \cap H_B^{50}|/50$, where H_L^{50} defines the 50 most popular hashtags in location L .

For each location, we calculated the hashtag similarity to all others. Figure A.6 shows the results for Aachen, Munich, and Overath in averages for groups of 100 locations. While the x-axis describes the distance to other cities, the y-axis denotes the similarity score. For Aachen and Overath, we observe that closer locations are on average more similar

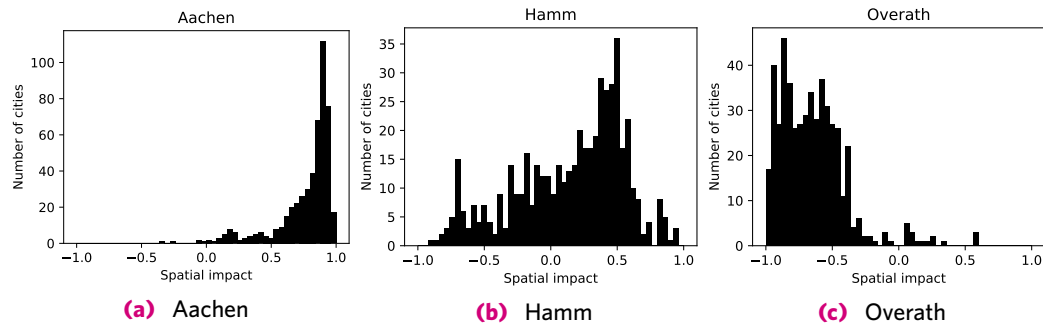


Figure A.5: Spatial Impact Histograms from Aachen, Hamm, and Overath to the top 500 locations in the complete dataset. Aachen heavily influences most other cities, Overath is mostly influenced by other cities, and Hamm is both influenced by several cities and influencing other cities.

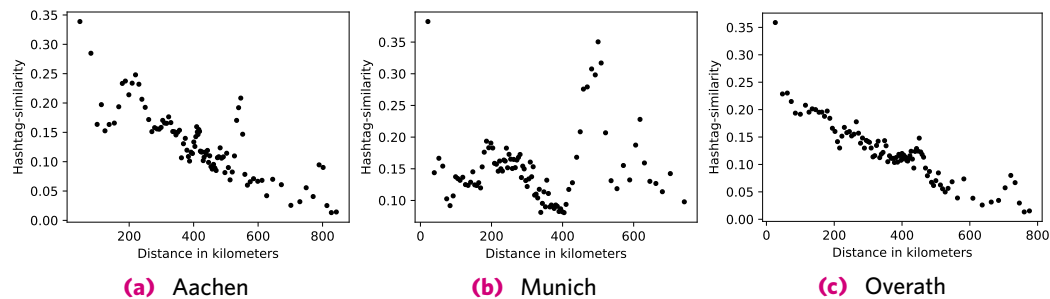


Figure A.6: Hashtag similarity of Aachen, Munich, and Overath to cities in a certain distance. Cities closer to each other tend to share more hashtags. However, big cities are similar to each other no matter the distance. Averages of groups of 100 locations.

than locations farther away. However, there are several peaks of which the biggest ones represents Berlin¹. It seems apparent that big cities are connected to each other and share hashtags no matter the distance, which is supported by the example of Munich. Yet, small cities like Overath are less affected. [Ferrara et al.,] showed similar results for Twitter: W.r.t hashtags, big cities are more similar to each other than to closer, smaller cities.

We verified that this also applies for Jodel considering all hashtags of both cities. The relation we see for Overath of closer cities having more hashtags in common has likewise been shown for Twitter [Kamath et al., 2013]. Our hypothesis is that on Jodel, hashtags *travel* long distances between big cities and then spread across smaller cities within the local neighborhood.

A.4.5.3 Findings

While the hashtag similarity metric does not directly reflect individual user's contribution to hashtag spreading, it still provides insights into the dis-/similar hashtag usage of communities. Big cities share more popular hashtags and are therefore generally more similar to each other, whereas smaller cities gradually share their most popular hashtags with their local neighborhood. In combination with the spatial influence, this supports our conclusion that hashtags likely spread via the bigger cities into such local neighborhoods.

A.5 Jodel DE Hashtag Classification

Within our analysis of hashtags, we have observed that the hashtags can be clustered into different groups (cf. Figure A.4a & A.4b). We know from literature that there are corresponding types of hashtags on e.g., Twitter. That is, [Kamath et al., 2013] distinguishes

¹Within our dataset, Berlin is split into districts and therefore present multiple times.

between local interest hashtags, regional and event-driven hashtags, and other worldwide memes. We were wondering if and in which way Jodel’s locality actually catalyzes other—very local—or prohibits global hashtags. For answering this questions, we create a statistical classifier for determining the hashtag type in three steps: *i*) defining suitable hashtag classes in line with our observations so far, *ii*) manual hashtag classification for providing an answer on a content level, and *iii*) training and validation of statistical models.

A.5.1 Hashtag Content Categories

Leveraging hints from Section [◆ B: Cross-Country Differences and Structural Implications](#), manual inspection and expert domain knowledge, we first iteratively defined and verified four different meme classes as follows:

- **Local events:** Often trends originating from a single post (e.g., a funny story) that gained attention in the local community. It is typically very local and short-lived.
- **Local phenomena:** Trend usually related to local persons or buildings. It is typically very local and long-lived.
- **Events:** Short-lived or recurring trend usually related to a real-world happening of larger interest.
- **Other memes:** Memes not included in Jodelstories or Local phenomena.

We labeled the most 450 popular hashtags that had their first occurrence after 1st January 2016 to filter out most of the generic statements. Besides, this makes the classes more balanced, as local trends are much more prominent in this restricted dataset. Due to missing context information or non-fitting classes, we could not classify 49 hashtags. The majority (64 %) of the remaining 401 hashtags were labeled *other meme*, whereas *local phenomenon* (82) represents the second biggest class, Events (35) and Local Event (29) being relatively equal in size.

Having learned that we indeed find trends in terms of hashtags that w.r.t our previous metrics and the manual classification reflect the locality of the Jodel application, we next try to establish the classification methods for them. Thus, we define features that we will use including the presented and analyzed metrics plus some additional temporal and text-based ones in the following section.

A.5.2 Features

Our aim is to create a statistical classifier for determining the hashtag type. For our classification approach, we used the features listed in Table [A.2](#). This list includes all spatial and temporal metrics that have been discussed before. Besides simple features like hashtag and comment counts, we further added temporal metrics of *peak increase* being defined as the number of posts in seven days prior to the peak divided by the number of posts on the peak day—and *peak decline* alike, but after the peak. These features, therefore, describe how suddenly a trend occurred and disappeared.

A.5.3 Classifiers and Results

A.5.3.1 Classifiers

We have applied different statistical methods to our classification problem: k-nearest neighbors, Classification and Regression Trees, Naive Bayes, Logistic Regression, LDA and ZeroR as a baseline. We used 10-fold cross-validation on our manually classified hashtag dataset to verify the results of each classifier. All classifiers outperform the baseline ZeroR-classifier.

| Feature | Definition |
|------------------|---|
| Focus | The focus of the hashtag. |
| Entropy | The entropy of the hashtag. |
| Spread | The spread of the hashtag. |
| Temporal focus | The amount of Jodels posted on the peak day of the hashtag divided by the total number of uses. |
| Temporal entropy | Similar to spatial entropy where different days are considered. Gives a number for the “randomness” of the distribution. |
| Temporal spread | Similar to spatial spread of the avg distance [days] from the weighted midpoint of all occurrences of the hashtag. |
| Local variation | The local variation of the hashtag. A measure for the regularity of the hashtag’s usage. |
| Hashtags | Average number of hashtags per Jodel. |
| Comments | Average number of comments per Jodel. |
| User diversity | Number of unique users of the hashtag divided by its total use. |
| Exclamations | Fraction of Jodels that contain an exclamation mark. |
| Questions | Fraction of Jodels that contain a question mark. |
| Peak increase | Compares post volume of seven days before the peak with the height of the peak. Is a measure for how “sudden” the peak occurred. A low value indicates a sudden increase in popularity. |
| Peak decline | Seven days after the peak divided by the height of the peak. Describes how fast interest declined after the peak day. A low value means the interest disappeared suddenly. |

Table A.2: Features used for Classification.

| Hashtag Class | Precision | Recall | F1-Score |
|-----------------|-----------|--------|----------|
| Event | 0.66 | 0.80 | 0.70 |
| Local event | 0.79 | 0.72 | 0.74 |
| Local phenomena | 0.87 | 0.95 | 0.91 |
| Other memes | 0.97 | 0.93 | 0.94 |

Table A.3: LDA Classifier Results, Precision, recall and f1-score. Averages of 10 runs with different dataset-splits.

While all approaches perform well (detailed results omitted), LDA resulted in a good compromise of the smallest average \pm standard deviation. Therefore, we only present the results of the LDA classifier in Table A.3. We observe that *events* have the lowest precision value with 0.66. However, this is still a good result as less than 10% of the hashtags are events. The other results are good as well, especially the local phenomena and memes with high F1 scores.

In this classification, both the spatial and the temporal features provided most benefit as removing them caused in both cases a considerable drop in accuracy of at least 0.1, whereas user diversity had only a very minor influence.

A.5.3.2 Findings

We have shown that we can predict the class of a hashtag by using its spatial and temporal properties. In conclusion, this confirms our theory that the Jodel platform actually has specific local short-lived and long-lived hashtags that differ to countrywide generic memes and events. While we may extend the classification scheme with more features and could apply advanced machine learning techniques, such as neural networks, this is a first step towards automatically classifying certain countrywide/global and in opposition local trends on Jodel—either being short- or long-lived according to our defined classes.

A.6 Conclusions

Within this section, we studied the hashtag propagation through the lens of a platform operator by having the unique opportunity to analyze data from Germany (2014 to 2017) provided by Jodel. With this longitudinal data set, we studied the key design pattern of being location-based and its influence on hashtag usage and spreading in comparison to the

global counterpart Twitter. We applied established metrics designed to capture the spatial focus and spread of Youtube Videos [Brodersen et al., 2012], or Twitter hashtags [Kamath et al., 2013] to Jodel and extend them with a temporal dimension covering the diffusion of hashtags in time. While we find significant qualitative differences to Twitter of hashtags generally being less focused on Jodel and thus having a higher entropy, the spatial spread also deviates from Twitter. Yet, we find evidence for local hashtags that are a potential result of Jodel’s design.

Further, we identify similarities in hashtag usage between nearby and larger cities and present case studies of their spacial impact supporting this finding. By correlating spatial and temporal metrics, we identify four different hashtag classes distinguished by their spatial and temporal extent. Informed by manual labeling of 450 most frequently used hashtags, we created an automatic classification scheme using machine learning models with great success.

As for future work, while we focused on the empirical birds-eye view on the hashtag usage, it will be interesting trying to apply epidemic modeling approaches. Further, individual user behavior and possible groups w.r.t. their spreading influence will provide deeper insights—especially in the sense of Jodel’s design choice of being location-based.

B Anonymous Messaging Contents

In this section, we study what users talk about in a plethora of independent hyperlocal and anonymous online communities in a single country: Saudi Arabia (KSA). We base this perspective on performing a content classification of the Jodel network in the KSA. To do so, we first contribute a content classification schema that assesses both the intent (why) and the topic (what) of posts. We use the schema to label 15k randomly sampled posts and further classify the top 1k hashtags. We observe a rich set of benign (yet at times controversial in conservative regimes) intents and topics that dominantly address information requests, entertainment, or dating/flirting. By comparing two large cities (Riyadh and Jeddah), we further show that hyperlocality leads to shifts in topic popularity between local communities. By evaluating votes (content appreciation) and replies (reactions), we show that the *communities* react differently to different topics; e.g., entertaining posts are much appreciated through votes, receiving the least replies, while beliefs & politics receive similarly few replies but are controversially voted.

B.1 Introduction

Anonymity on internet platforms is often controversially discussed between *i*) enabling freedom of speech and *ii*) enabling toxic environments [Papasavva et al., 2020]. Prior work studied the spectrum of discussed topics on anonymous to non-anonymous platforms showing that users have preferences which posts should be anonymous and which should not [Correa et al., 2015]. Prime platform examples for the latter are anonymous, but also forming location-based communities, e.g., as researched empirically on Whisper [Wang et al., 2014], or analyzed w.r.t. sensitive contents and the impact of anonymity [Correa et al., 2015]. The anonymous location-based app YikYak can be characterized with entertaining and informational contents leveraging self-supervised learning [Black et al., 2016], while others find evidence for flirting and dating [Wu et al., 2017] via crowdsourcing. Qualitative studies identify a broad range of motivations for anonymous posts, e.g., social isolation, social venting, requesting and granting emotional support, identity, while eliminating fear of rejection, to name a few [Vaterlaus, 2017].

This opens the question if anonymity yields to a richer content spectrum, especially in more conservative regimes. In the case of Saudi Arabia, [Guta and Karolak, 2015] report on interviews with KSA women about boundaries and new freedoms, granted through the Internet—rendering anonymous platforms specifically interesting. Alsanea portrays the Saudi life in the 2007’s novel *Girls of Riyadh* [Alsanea, 2007] through the eyes of four young girls. Nonetheless, society is continually changing, and has changed, e.g., women’s right to vote in 2015, or the KSA was about to lift the women driving ban in 2017²

Orthogonal to anonymity, recent online messaging platforms embrace hyperlocality, i.e., they display posted content only to spatially local users. It is an open question whether this property implies shifts in discussed contents. One platform that combines both properties—anonymity and hyperlocality—is Jodel. The app only displays content posted within the users’ proximity—unlike Twitter and other platforms, no communication with remote users is possible. The platform became popular in the KSA in 2017.

Structure [JH4]

 [B.2: Related Work](#)

 [B.3: Content Classification Schema](#)

 [B.4: Classification Campaign](#)

²https://en.wikipedia.org/wiki/Women_to_drive_movement

♡ [B.5: What Jodel Users talk about in SA](#)

[B.6: Jodel SA Hashtags](#)

📄 [B.7: Conclusions](#)

B.1.1 Research Questions

Given the different cultural background in the KSA, we are interested in (RQ1) what are discussed contents on the Jodel platform in the KSA and how is it perceived and reacted upon by the communities. We thereby study effects of Jodel’s key design features of *anonymity* and *hyperlocality*. What are the Jodel KSA users talking about?—*Behind the veil*. Subsequently, we raise the question (RQ2): How can we design a suitable crowdsourcing annotation schema to assess Jodel content? Last, (RQ3) how can we classify hashtags—as proxy measure for post content.

B.1.2 Approach

We take the rare chance to analyze ground truth information provided by the social network operator to study a random sample of Jodel content posted within the KSA. We enable content classification using a new content annotation schema that assesses the *intent* (why) and the *topics* (what) of a post. We apply the scheme to 15k randomly sampled posts that are annotated by expert native-speaking classifiers. By splitting the data set by city, we study local content biases between two major cities in the KSA. Leveraging empirical data, e.g., vote scores, or #replies, we study how users appreciate discussed content-classes.

In a last step, we extract and classify hashtags w.r.t. sensitivity.

B.1.3 Results

We contribute a content classification schema to classify social media posts by their intent (why) and topic (what). We apply the schema to ground truth data from the hyperlocal and anonymous Jodel network in the KSA, using native-speaking expert classifiers. We successfully put this schema to work for classifying Jodel content (e.g., high rater agreement). In contrast to what might be expected in anonymity, we observe a rich set of benign (yet at times controversial in conservative regimes) but non-toxic topics. We further show local topical biases within the two largest cities. Differences exist also in how content triggers responses and how users appreciate content through voting. We develop a second schema to classify the top 1k hashtags.

B.2 Related Work

While a large body of work aimed at understanding Online Social Networks, two key design features of new types of networks have received little attention so far: anonymity and hyperlocality.

Anonymous platforms are known for their ephemerally and toxicity, e.g., 4chan [Papavasava et al., 2020]. We contrast this perspective by showing that this is not generally the case; while Jodel is an anonymous platform, the posted content is largely non-toxic. Location-basedness has been analyzed on e.g., flickr [Cha et al., 2009], or Twitter [Yin et al., 2011]. The latter was used to model information diffusion [Kamath et al., 2013], also conducted on Jodel itself [JH9]. Yet, Twitter and Flickr enable global communication—not being possible on Jodel—embracing local communication.

Jodel’s niche of combined anonymity and hyperlocality was analyzed on e.g., Whisper [Wang et al., 2014, Correa et al., 2015], and very similar app YikYak of which we highlight a few examples. A wide range in methodology can be found in empirical [Saveski

| Abbreviation | Class Information | #Annot | pdf |
|---------------------------------|---------------------------------------|---|-----------|
| $\mathcal{I} = \text{Intent s}$ | | #Coders=2; $\alpha_M^{\mathcal{I}} = 0.74$, <i>substantial</i> | |
| EntObserv | Entertaining Observation | 122 | 0.019 |
| DistRelComp | Distress Release & Complain. ab. Self | 580 | 0.091 |
| GenEntert | General Entertainment | 618 | 0.097 |
| Info | Information Sharing | 638 | 0.100 |
| SocVentComp | Social Venting & Complain. ab. Others | 644 | 0.101 |
| Other | Other | 691 | 0.108 |
| Self | Self Expression | 927 | 0.146 |
| Seek | Seeking Interaction/Information | 2,149 | 0.337 |
| $ \mathcal{I} = 8$ | | Σ | 6,191 1.0 |
| $\Theta = \text{Topic s}$ | | #Coders=2; $\alpha_M^{\Theta} = 0.64$, <i>substantial</i> | |
| IllegalViolence | Illegal & Violence | 155 | 0.018 |
| AnimalsNature | Animals & Nature | 228 | 0.026 |
| FitnessHealth | Fitness & Health | 267 | 0.031 |
| FoodDrink | Food & Drink | 358 | 0.042 |
| ProdService | Products & Services | 376 | 0.044 |
| EduWork | Education & Work | 479 | 0.056 |
| FashionBeauty | Fashion & Beauty | 506 | 0.059 |
| SocialMedia | Social Media | 608 | 0.071 |
| BeliefsPol | Beliefs & Politics | 707 | 0.082 |
| EntertCulture | Entertainment & Culture | 896 | 0.104 |
| Other | Other | 983 | 0.114 |
| Self | Self (Personal) | 1,092 | 0.127 |
| PeopleRelation | People & Relationships | 1,960 | 0.228 |
| $ \Theta = 13$ | | Σ | 8,615 1.0 |

Table B.1: Annotation Schema and Crowdsourcing Results. We code *intents* (\mathcal{I}) catching the individual incentives, and *topics* (Θ) representing discussed contents. Figures are given in the amount of jodels, number of annotations, and the resulting probability. The overall coder agreement by MASI distance is substantial, $\alpha_M^{\mathcal{I}} = 0.64$ and $\alpha_M^{\Theta} = 0.74$.

et al., 2016], but also qualitative studies [Lee et al., 2017]. To quantify discussed topics, works leveraged (self)-supervised models [Black et al., 2016] finding platform content to be rather ephemeral in an impersonal environment, which is confirmed via survey study in [Vaterlaus, 2017, Lee et al., 2017]. Nonetheless, another work revealed “Personal Admission”, “Observation” or “Information/Advice” resembling popular content [Heston and Birnholtz, 2016a]. Others conducted crowdsourcing at large scale [Wu et al., 2017] finding “Dating & Sex” or “Local Life, Weather & Announcements” being topmost discussed topics.

We conclude that research on anonymous hyperlocal platforms has matured over past years, yet existing studies ironically neglect their key feature—they focus on the Western/US region only. Within the context of the Jodel particularities within the KSA, we provide answers as to what drives individual user behavior and what are discussed topics; enabled by our methodological approach to a generic crowdsourcing annotation schema.

B.3 Content Classification Schema

In this section, we contribute a crowdsourcing schema that enables to classify content posted in social media platforms. Our schema assesses two key aspects: *i)* why a user posted, i.e., what is the purpose or intent(s) \mathcal{I} of a post. *ii)* What topics Θ are presented in a post. For each the intents \mathcal{I} and the topics Θ , multiple labels can be attributed by human annotators to a single post.

B.3.1 Design objectives and development

We iteratively developed and refined the presented schema over multiple months. Our objective was to arrive at a *minimal set of categories* easing the classification task. The categories should have little to no semantic overlap to make classes easily distinguishable; for both easing annotation and better interpretability of results. Categories must further be

sufficiently expressive for the content posted on Jodel, i.e., the amount of posts annotated with “Other” should be minimal. The design of the categories naturally involves a trade-off between being very specific (many categories) and ease of use (few categories).

B.3.1.1 Intent \mathcal{I} of a post

In the first category, we assess *why* users post in social media, i.e., the user’s driving intent of a post as interpreted by the human annotator. In our schema, we use eight possible intents that we base upon a prior work’s [Kang et al., 2016] taxonomy, derived from semi-structured interviews with social media users. Table B.1 shows the list of intents \mathcal{I} , e.g., if a user is sharing information or is seeking for information. The selected intents can be assessed by human classifiers solely by reading the posted textual content. As posts may have multiple intents, we allow multi-labels per post.

B.3.1.2 Topic Θ of a post

In the second category, we assess *what* topic a post is about. Our initial set of categories bases on prior work on content classification of the Whisper network [Paul et al., 2011, Correa et al., 2015], which we iteratively refine and adapt to content shared via Jodel KSA. We show the list of topics Θ in Table B.1. We also opted for multi-labelling.

B.3.2 Iterative schema development

We based the initial version of the schema on prior work ([Kang et al., 2016] for the intents and [Paul et al., 2011, Correa et al., 2015, Wu et al., 2017] for the topics), that we have iteratively refined and adapted in multiple classification campaigns, each based on a small random samples of Jodel posts. Qualitative coder feedback was in line mentioned works, we do not find any specifically toxic environment. An empirical view shows that from the overall KSA’s content we find only $p \approx 1.6\%$ of outvoted (disliked) or otherwise blocked posts and replies. Though this figure is higher for in-app prominently displayed posts with $p \approx 7.3\%$, only a minor fraction of these posts has been blocked by escalated moderation and flagging with $p \approx 2.3\%$. This indicates that applied moderation works; especially due to distributed moderation being very interesting by itself, we leave this topic for future work. That is, we excluded any focused toxicity class as it also arguably might not fit well into topics nor intents.

In each campaign, we identified categories being used seldom, missing, or being semantically ambivalent, i.e., they resulted in strong disagreement among annotators. After each classification run, we discussed disagreement and other challenges with our annotators, ultimately leading to an improved version of the schema. We present our final schema and put it to work enabling us to classify Jodel posts with substantial coder agreement.

B.3.3 Implementation

We realized the crowdsourcing system from scratch as a web application in PHP, shown in Figure B.1. The system is self-hosted and enables us classifying sensitive content that cannot be made available to external services or users (e.g., via common crowdsourcing platforms such as Amazon Mechanical Turk). We can define and monitor annotation campaigns: Describe which posts should be classified by how many of the available annotators.

ماقالتي صحيني ودارمه الصبح ونام متأخر اصحبه ولا؟ ومو متأكد
صالح ولا

Which types are present in the following Jodel?

Multiple Selection is possible

Complaint Confession Consultation Invitation Positive Story Quote Random Wish

Which topics are present in the message of the post?

Multiple selections are possible














| | | | |
|---|--|--|--|
| Entertainment & Culture  <input type="radio"/> | Education & Work  <input type="radio"/> | People & Relationships  <input type="radio"/> | Fitness & Health  <input type="radio"/> |
| Products & Services  <input type="radio"/> | Animals & Nature  <input type="radio"/> | Fashion & Beauty  <input type="radio"/> | Food & Drink  <input type="radio"/> |
| Beliefs & Politics  <input type="radio"/> | Self (Personal)  <input type="radio"/> | Social Media  <input type="radio"/> | Illegal & Violence  <input type="radio"/> |
| Other  <input type="radio"/> | | | |

Figure B.1: Crowdsourcing Classification System. Our coders are presented a post to be read. Then, they answer two subsequent questions: *i*) What is the intent, and *ii*) What is the topic of this post?---Allowing multiple labels.

B.4 Classification Campaign

B.4.1 Study Design

We apply our content classification schema (Section ??) to annotate Jodel posts in the KSA. To protect the users' privacy and in compliance with the Jodel ToS, we cannot share the posts on external platforms such as Amazon Mechanical Turk. We thus run all campaigns on internal and protected machines that are only accessible to expert classifiers that we invite and associate with our research group. The human annotators are experts that *i*) are Arabic native speakers and *ii*) are familiar with the dialect spoken in the KSA (e.g., by originating or having lived in the region). Within our schema development, we realized that the annotators' origin (e.g., Egypt) *can* challenge the understanding of local KSA dialects, and can thus lead to disagreement between annotators. Therefore, we selected future annotators by removing language boundaries and ensuring more consistent annotations. Using expert classifiers reduces the number of classifiers needed; prior work showed that using non-expert classifiers requires a factor of 4 more classifiers [Snow et al., 2008]. Also, since all of our classifiers are known to us and trusted, we do not need to employ control questions to detect cheating attempts as in crowdsourcing on public platforms.

For coherent classification results, we focus on the content of the starting post, i.e., not classifying complete threads nor replies, due to findings within the schema development phase. We experimented with presenting more contextual information to our annotators by including the entire discussion thread (i.e., original post *and* its replies). Particularly longer discussion threads tend to shift from one topic to others and are thus challenging to label coherently.

For the schema development, we employed four expert classifiers aged 20-30 years with a 1:1 male:female ratio. Three out of four classifiers have prior experience with the KSA

dialect (e.g., from having lived in Saudi Arabia or Oman). Within development and schema optimization, we performed about 7,700 classifications across various setups with multiple classifiers into a feedback loop of discussing ambiguity, disagreement, ambivalence, and other experiences—each resulting in a new schema version. Later we settled with the final schema that includes topics and intents.

B.4.2 Campaign and Schema Quality Evaluation

To study the content of Jodel posts in SA with our fixed final schema, we employed two classifiers (aged 20-30 years, male and female, from Syria and Iraq), who iteratively performed five subsequent classification campaigns (Table B.2). All campaigns use sampled post data from *i)* the entire KSA, *ii)* Jeddah, and *iii)* the capital Riyadh. They first completed a training period to qualitatively familiarize with Jodel contents and our schema. Since the agreement for all campaigns is high, we opted for using all campaigns for evaluation.

Next, we evaluate the quality of the campaign and the schema.

B.4.2.1 Qualitative Coder Feedback

From analyzed Jodels, both annotators believe that dominantly teenagers and young people use the platform. Since the Jodel network is anonymous, we lack any demographic info to validate this claim. The classifiers further noted that a number of posts focus on finding partners for online games, especially *ludo star*, a mobile app version of the board game *Don't Get Angry* [Wikipedia contributors, 2022] (identified as *seeking interaction* in the later results).

B.4.2.2 Coder Agreement

We measure our expert classifier interrater-agreement with Krippendorff's alpha [Krippendorff, 2012]. A standard approach that provides several benefits: *i)* it behaves well with any number of classifiers, *ii)* it is capable of handling missing data, i.e., single classifications, *iii)* it adjusts for sample sizes, and *iv)* it may be used for various types of data—nominal in our case. Due to our *multi-label* approach, we further need to use a suitable distance metric that compares sets of labels. We present our agreement results in Table B.2 using various distance metrics: Binary, Jaccard, and MASI [Passonneau, 2004].

There is no clear up-/downwards trend in agreement across the campaigns; thus, it remains unclear whether annotators accustom better to the classification scheme. When analyzing intents and topics separately, we note that intents \mathcal{I} generally suffer less from non-agreement. The topic Θ classification has led to less agreement consistently. Later iterations yield better results for both, intents and topics. Our results should be viewed with care as [Krippendorff, 2012] suggests not to use data sets with alpha values below 0.667 for any non-tentative conclusion, yet our achieved agreement is still well above chance.

B.4.2.3 Multi-Labels

We analyzed the amount of classified posts with multiple intents \mathcal{I} or topics Θ . While \mathcal{I} intents almost accidentally only were assigned a multi-label twice, we found 1,917 multi-labels across our coders for Θ topics ($p=\{2: 0.81, 3: 0.17, 4: 0.02\}$). Observed multi-labels are usually not specifically in line between the classifiers as can be seen in lower α_B scores across the board; lower α_M values in comparison to α_J confirms this finding as the MASI distance adds a distinct bias according to subset-similarities.

Thus, we also investigated on observed confusion between multi-label Θ topic annotations and annotators. We find a strong diagonal as expected due to substantial agreement.

| No | #Posts | #Coders>1 | \mathcal{I}/Θ | α_B | α_J | α_M | Agreement _M |
|-----|--------|-----------|----------------------|------------|------------|------------|------------------------|
| 1 | 733 | 0 | | - | - | - | |
| 2 | 1,999 | 0 | | - | - | - | |
| 3 | 400 | 398 | \mathcal{I} | 0.66 | 0.66 | 0.66 | substantial |
| | | | Θ | 0.44 | 0.57 | 0.52 | moderate |
| 4 | 1,000 | 993 | \mathcal{I} | 0.69 | 0.69 | 0.69 | substantial |
| | | | Θ | 0.45 | 0.60 | 0.55 | moderate |
| 5 | 400 | 398 | \mathcal{I} | 0.70 | 0.70 | 0.70 | substantial |
| | | | Θ | 0.57 | 0.62 | 0.61 | substantial |
| all | 4,532 | 1,789 | \mathcal{I} | 0.74 | 0.68 | 0.74 | substantial |
| | | | Θ | 0.57 | 0.68 | 0.64 | substantial |

Table B.2: Classification agreement for multiple iterations on thread starting posts; For intents \mathcal{I} and topics Θ we show classifier agreement by α_B Binary, α_J Jaccard, and α_M MASI distance. There is a substantial agreement between our two coders. Coders agree better on intents (\mathcal{I}) than topics (Θ).

However, we identify the axis along *People & Relationships* as most ambiguous. Other single Θ -hotspots are worth a look: Some may raise self-explainable confusion: e.g., Θ -Self \times *FitnessHealth*, *EduWork*, *FashionBeauty*, or *EntertCulture* \times *AnimalsNature*.

B.4.2.4 Overall Confusion

In Figure B.3a and Figure B.3b, we provide the complete picture of confusion within our annotation schema for intents and topics. We de-biased the join operation by introducing a natural weighting factor of $n \times m^{-1}$ as it would otherwise favor multi-label annotations. Note the log color scale.

For intents \mathcal{I} , in Figure B.3a, we observe a strong correlation across the diagonal as expected from substantial annotator agreement. However, several confusion hotspots remain, some being self-explanatory: E.g., *GenEntert* \times *EntObserv*, or *DistRelComp* \times *SocVentComp*; yet we must take note of the rest.

As for topics Θ , in Figure B.3b, we observe similar patterns to the multi-label confusion. Though we observe substantial agreement on the strong diagonal, we again see evidence for ambiguity along *PeoplRelation*, against *FitnessHealth*, and naturally along Θ -Other. Noteworthy, we also identify several hotspots along Θ -Self.

B.4.2.5 Takeaway

We have presented our schema of intent and topics at work within a series of subsequent crowdsourcing campaigns for Jodel KSA. Within about 15k annotations (intent-#annot=6,191, topic-#annot=8,615), we find substantial coder agreement ($\alpha_M^{\mathcal{I}} = 0.75$, $\alpha_M^{\Theta} = 0.64$) across #posts=1,789 (#Coders=2) and conclude that our proposed schema has sufficient quality for further evaluation.

B.5 What Jodel Users talk about in SA

In this section, we study the Jodel post contents (i.e., *topics* and their *intents*) in the KSA that result from our classification campaign.

B.5.1 Countrywide Perspective on Jodel Content

We begin with analyzing the content classification for the country-wide overall annotations, before we study differences between two cities. First, we discuss Table B.1 showing the popularity of topics Θ and intents \mathcal{I} by annotation counts. Since topics and intents are intertwined, we also show the combination of $\mathcal{I} \times \Theta$ as a heatmap in Figure B.2. We complement this heatmap by discussing topic distributions across intents next (not shown).

B.5.2 Intents \mathcal{I}

The dominant intents are (see Table B.1): *I-Seek* ($\Sigma 34\%$) and *Self* ($\Sigma 15\%$), followed by *SocVent & DistRel* ($\Sigma 19\%$), *Info* ($\Sigma 10\%$), and *Entertainment* ($\Sigma 12\%$). We only observe little disagreement between classifiers, explained by a possible ambiguity within $\mathcal{I}\text{-Entobserv} \times \text{GenEntert}$, or apparent confusions along *Self*.

B.5.3 Topics Θ

Albeit slightly weaker annotator agreement, the discussed topics largely revolve around $\Theta\text{-PeopleRelation}$ accounting for $\Sigma 25\%$ annotations, which is also our most confused category. We find other popular themes in $\Theta\text{-Self}$ ($\Sigma 13\%$), *Other* ($\Sigma 13\%$), followed by *EntertCulture* ($\Sigma 10\%$), and *BeliefsPol* ($\Sigma 8\%$). *IllegalViolence* ($\Sigma 2\%$), *AnimnalsNature* ($\Sigma 3\%$), and *FitnessHealth* ($\Sigma 3\%$) are least popular.

B.5.4 Intents $\mathcal{I} \times \Theta$ Topics

We identify specific hotspots of interests by combining $\mathcal{I} \times \Theta$ as a heatmap in Figure B.2. Jodel is mostly being used out of the intent of *I-Seeking Information & Interaction* ($\Sigma 34\%$) for $\Theta\text{-PeopleRelation}$ ($p[\Theta|\mathcal{I}]=23\%$) and *EntertCulture* ($p[\Theta|\mathcal{I}]=12\%$), closely followed by others. Likewise, finding *I-Self* ($\Sigma 15\%$) Expression across the board, users again focus on $\Theta\text{-PeopleRelation}$ ($p[\Theta|\mathcal{I}]=23\%$) and *Self* Expression ($p[\Theta|\mathcal{I}]=14\%$). Out of $\mathcal{I}\text{-GenEntert}$ ($\Sigma 10\%$), we want to highlight $\Theta\text{-BeliefPol}$ ($p[\Theta|\mathcal{I}]=57\%$). Whereas $\mathcal{I}\text{-SocVentComp}$ ($\Sigma 10\%$) almost naturally goes along with the topic $\Theta\text{-PeopleRelation}$ ($p[\Theta|\mathcal{I}]=41\%$), $\mathcal{I}\text{-DistRelComp}$ ($\Sigma 9\%$) aligns with $\Theta\text{-Self}$ ($p[\Theta|\mathcal{I}]=31\%$) and *PeopleRelation* ($p[\Theta|\mathcal{I}]=22\%$).

B.5.4.1 Anonymity

From our analysis, it comes apparent that most content posted on Jodel indeed is related to users' intent for *I-Seeking Information & Interaction*, and *Self* Expression accounting for 50% of all annotations with strong trends towards the topics $\Theta\text{-PeopleRelation}$, *EntertCulture*, and *Self* totaling for $p[\Theta|\mathcal{I}]=37\%$, 16%, 13%, $\Sigma 66\%$) within these intents. Further, another $\Sigma 19\%$ of posts are driven by $\mathcal{I}\text{-SocVentComp}$ and $\mathcal{I}\text{-DistRelComp}$ within the same topical regime. Unfortunately, crowdsourcing a well-suited anonymity-sensitivity score relying on many ($n=89$) coders [Wang et al., 2014] is not possible in our case. Foreshadowing our categorization on hashtags in (Section B.6), we nonetheless conclude that found prominent interaction situations are in line with the key design feature of being anonymous. Individuals may find safety behind the veil of anonymity, allowing for free speech about personal experiences, wishes, questions, or possibly controversial opinions.

B.5.4.2 Hyperlocality

Our generic intent $\mathcal{I} \times \Theta$ topic schema does not allow for a distinguished evaluation whether a post refers to *anything* local, which turns out to be a challenging question. With qualitative feedback from our annotators, we conclude that e.g., a larger part of *I-Seeking Information & Interaction* actually refers to local matchmaking, events, local services, or educational institutions. By focusing on the platforms content and driving factors first, we leave distinguished analyses of a *well-suited sense of locality* to future work. Though, we discuss this topic in more detail within the next Section B.5.5 comparing intents and topics in Jeddah \times Riyadh, and provide deeper insights via hashtags in Section ??.

| | | | | | | | | | | | | | |
|-------------|-----------------|---------------|---------------|-----------|-------------|---------|---------------|-------------|------------|---------------|-------|------|----------------|
| EntObserv | 4 | 4 | 6 | 10 | 4 | 15 | 6 | 15 | 7 | 12 | 9 | 35 | 57 |
| DistRelComp | 24 | 21 | 56 | 38 | 15 | 72 | 50 | 22 | 22 | 32 | 44 | 263 | 186 |
| GenEntert | 1 | 5 | 4 | 7 | 7 | 13 | 7 | 19 | 394 | 135 | 38 | 10 | 48 |
| Info | 7 | 31 | 14 | 23 | 20 | 56 | 65 | 22 | 25 | 100 | 38 | 341 | 134 |
| SocVentComp | 52 | 23 | 21 | 32 | 27 | 51 | 62 | 101 | 35 | 101 | 48 | 35 | 402 |
| Other | 3 | 8 | 2 | 6 | 12 | 12 | 10 | 75 | 18 | 51 | 477 | 16 | 42 |
| Self | 15 | 44 | 23 | 55 | 46 | 71 | 102 | 74 | 127 | 96 | 83 | 193 | 413 |
| Seek | 49 | 92 | 141 | 187 | 245 | 189 | 204 | 281 | 79 | 369 | 248 | 200 | 679 |
| | IllegalViolence | AnimalsNature | FitnessHealth | FoodDrink | ProdService | EduWork | FashionBeauty | SocialMedia | BeliefsPol | EntertCulture | Other | Self | PeopleRelation |

Figure B.2: Overall annotation counts. Intents $\mathcal{I} \times \Theta$ topics.

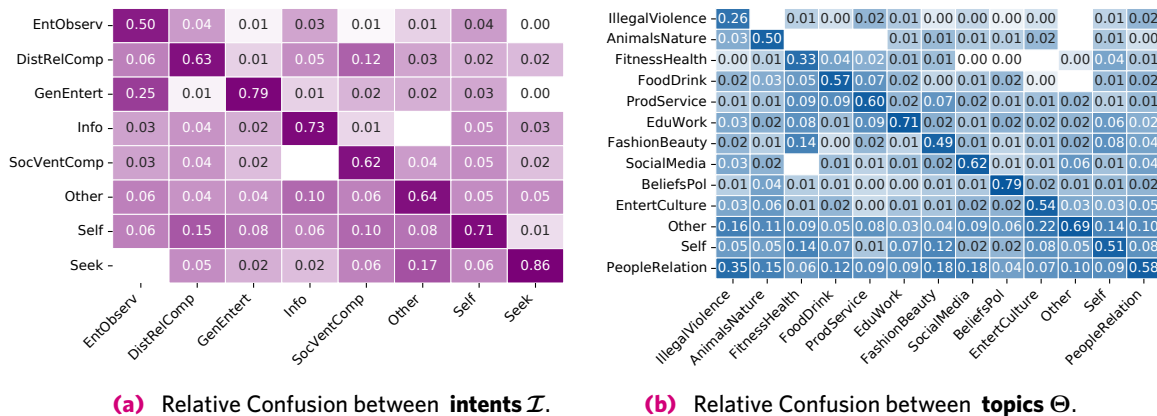


Figure B.3: (Multi-) Labeling Confusion. (Left, right:) Relative Confusion, column wise normalization, sums to 100%. De-Biased multiset join by weighting factor $(n \times m)^{-1}$. Log color scale.

B.5.4.3 Summary

(\mathcal{I}) Seeking Information & Interaction and Self Expression are the predominant drivers for creating content within the Jodel KSA communities. (Θ) Discussions and statements largely revolve around People & Relationships and personal statements (Self). Besides a significant amount of entertaining content, Jodel’s anonymity promotes personal and respectively potentially sensitive content (non-illegal).

B.5.5 City-Level Perspective on Jodel Content

By design, Jodel users can only communicate with other Jodel users in their direct proximity—no country-wide communication is possible. Thus, all posts carry a bias towards their local community. We next study whether and to which extent this bias is measurable across two major cities in SA: Riyadh (the capital) vs. Jeddah about 1000km away and believed to be more liberal. Both largest SA cities account for 38% (28%, 10%) total Jodel platform interactions in our meta data set. As we have sampled our annotation posts accordingly, we can draw a clear picture of driving intents and discussed topics for both communities—and shifts in local topical preferences.

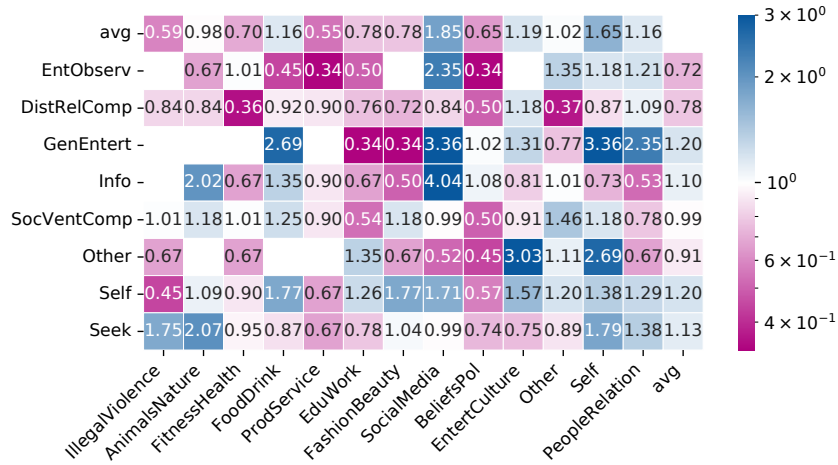


Figure B.4: Community differences, Jeddah × Riyadh PDF relative delta. Relative difference, factor between city PDFs, intents $\mathcal{I} \times \Theta$ topics. Values above 1 (blue) represent more items in Jeddah than Riyadh, likewise below 1 (pink) less.

B.5.5.1 Jeddah × Riyadh

We overall collect $\Sigma_J 1541$ ($\Sigma_R 1100$) annotations for Jeddah (Riyadh) across $\Sigma 1768$ posts with substantial annotator agreement (Jeddah: $\alpha_M^{\mathcal{I}}=0.71$, $\alpha_M^{\Theta}=0.62$, Riyadh: $\alpha_M^{\mathcal{I}}=0.75$, $\alpha_M^{\Theta}=0.63$).

To ensure that we do not only observe random noise, we measure the pairwise R2 scores between per city (Jeddah, Riyadh, random) Intents $\mathcal{I} \times \Theta$ Topic-pdf; resulting scores $R^2=\{\text{Jeddah-Riyadh: } 0.75, \text{Jeddah-random: } 0.44, \text{Riyadh-random: } 0.66\}$ increase confidence in varying biases in discussed contents across the country.

As we are particularly interested in distribution changes, we present the relative changes between both cities as a heatmap across Intents and Topics in Figure B.4. Note the log color scale. Comparing a Jeddah Probability Distribution Functions (PDF) baseline to Riyadh results in figures above one (blue) indicating more annotations for Jeddah, and vice versa.

To draw a better picture, we added column- and row-wise averages across Intents and Topics (avg). We find very similar intent measures in \mathcal{I} -*SocVentComp* (factor 0.99; $\Sigma 15\%$ of total \mathcal{I} -annotations). Non-differing topics may be considered in Θ -*AnimalsNature* (0.98; $\Sigma 4\%$), and *Other* (1.02; $\Sigma 14\%$) in 18% overall Θ -annotations.

Content which is more popular in Jeddah (blue). In Jeddah, we find more \mathcal{I} -*Self* Expression (1.20; $\Sigma 18\%$), *GenEntertainment* (1.20; $\Sigma 10\%$) and individuals *Seeking Information & Interactions* (1.23; $\Sigma 49\%$) in 77% \mathcal{I} -annotations. Increased topic figures are rather personal and casual around Θ -*SocialMedia* (1.85; $\Sigma 9\%$), *Self* (1.65; $\Sigma 13\%$), *EntertCulture* (1.19; $\Sigma 11\%$), and *PeopleRelation* (1.16; $\Sigma 29\%$) within 63% of total Θ -annotations.

Content which is more popular in Riyadh (pink). In Riyadh, we observe an overall shift towards \mathcal{I} -*EntObserv* (0.72; $\Sigma 3\%$) and *DistRelComp* (0.78; $\Sigma 14\%$), both accounting for 17% total \mathcal{I} -annotations. There appears to be a less heavy tailed broader spectrum of content: More popular topics are Θ -*ProdService* (0.55; $\Sigma 6\%$), *IllegalViolence* (0.59; $\Sigma 2\%$), *BeliefsPol* (0.65; $\Sigma 11\%$), *FitnessHealth* (0.70; $\Sigma 4\%$), *EduWork* (0.78; $\Sigma 6\%$), and *FashionBeauty* (0.78; $\Sigma 7\%$). They account for another 36% of overall Θ -annotations.

B.5.5.2 Findings

By comparing the two largest—spatially distinct—communities Jeddah and Riyadh, we show local biases in content occurrences. Riyadh experiences a broader less skewed spectrum of

| Aggregation (Topics) [†] | #A | #B | Σ | #A/ Σ | #Posts |
|---------------------------------------|-------|-------|----------|--------------|--------|
| Jeddah \cup Riyadh \times Country | 2,641 | 3,679 | 6,320 | 0.72 | 4,552 |
| Jeddah \times Riyadh | 1,541 | 1,100 | 2,641 | 0.58 | 1,768 |
| Early \times Established | 809 | 1,832 | 2,641 | 0.31 | 3,158 |

Table B.3: Classifications by comparison aggregate. Annotations partition sizes for relative comparison. Imbalanced subset sizes are due to sampling.

topics, whereas the Jeddah community focuses more on \mathcal{I} -(Self Expression, Entertainment) along Θ -(Social Media, People & Relationships).

B.5.6 Reactions upon Content by Jodel Users

So far, we studied what Jodel users in the KSA are talking about. In the next step, we extend this topic with a perspective on the community reactions upon content. That is, we raise the subsequent question whether certain intents or topics experience the same appreciation in terms of votes and replies. Thus, we gathered the total counts of up- and downvotes, both accumulated called Karma, posts and repliers per annotated thread. From these figures, we derive two scores: *i) Conversationness* [JH5] as an indicator for thread participation homogeneity. It represents the ratio of #repliers to #replies, i.e., higher values equal more people with fewer responses. And *ii) Vote-consensus* as an indicator for community voting confidence. We define it as the ratio of (#upvotes, #downvotes) to the total votes, while mirroring downvote-dominated values, i.e., values approaching (negative) one depict better consistency in upvoting (downvoting) behavior; neutral at zero in between.

Buried deep within per Intent and per topic distributions, we observe similarities across most metrics; yet they differ in cut-offs or variance. To picture an aggregated, but overall representative baseline, we present Cumulative Distribution Functions (CDFs) of Karma, replies and votes counts in Figure B.5a; whereas we show CDFs of both metrics *i) & ii)* in Figure B.5. To be brief, we only highlight and discuss distributional outliers.

B.5.6.1 Replies

The amount of replies is generally heavy-tailed as shown in Figure B.5a, that might be due to app design and feed presentation; in line with other research on the structure of OSN. However, we identify different skews within \mathcal{I} Intent distributions—the least replies can be expected for *GenEntert* (up to $\uparrow 25$ for 97%), whereas others already reach up to $\uparrow 25$ replies for between 80% to 90%. Within Θ -topic distributions, the least replies can be expected for *BeliefsPol* (up to $\uparrow 25$ for already 90%), appearing far less discussed than *IllegalViolence*, *PeopleRelation* & *FashionBeauty* ($\uparrow 25 \approx 70$ -80%).

B.5.6.2 Conversationness

In case of a low conversationness score, only few participants contribute to a long discussion. The distributions of intents and topics are almost linear from the origin individually cutting-off ($\uparrow x=1.0$) as overall shown in Figure B.5b. Most heterogeneous discussions appear for *\mathcal{I} -Seek Info* and *SocVentComp* (cut-offs at $\uparrow 0.75$ to $\uparrow 0.8$), versus more homogeneous conversations in *GenEntert* & *EntertObs* ($\uparrow 0.43$, $\uparrow 0.58$). Θ -*BeliefsPol* ($\uparrow 0.52$) remains most homogeneously discussed at two replies per participant on average.

B.5.6.3 Votes

Similar to replies, vote count distributions for Intents and topics remain heavy-tailed as shown overall series in Figure B.5a.

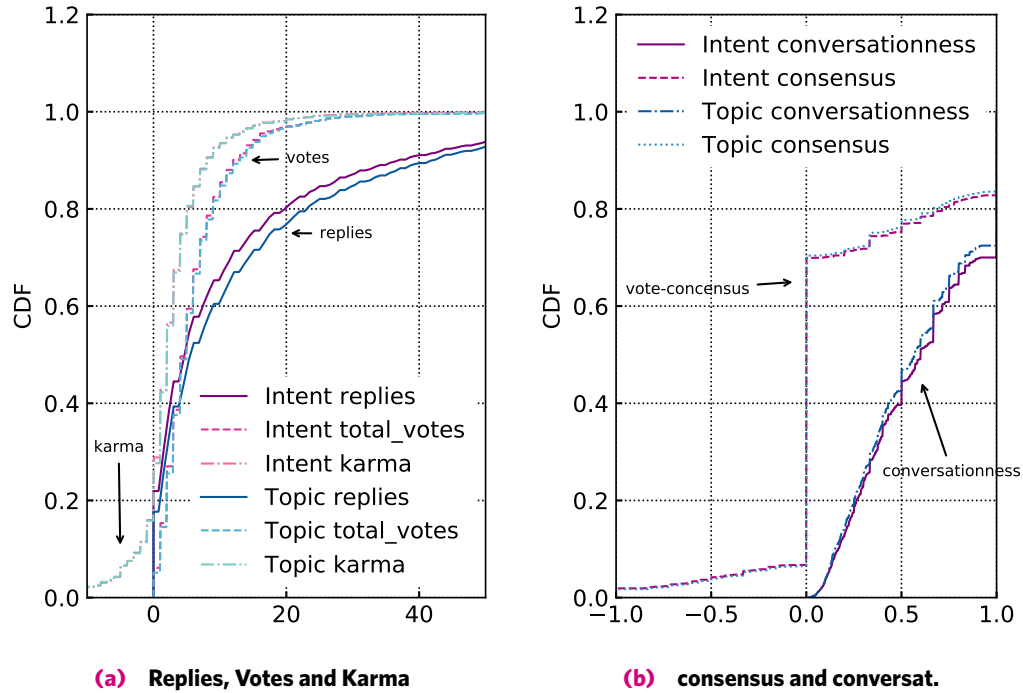


Figure B.5: Received Platform Reactions. Vote-Related: (left) Votes, Karma, (right) Vote-consensus. Reply-Related: (left) Replies, (right) Conversationness.

B.5.6.4 Karma

Karma describes the accumulated vote score between up- and downvotes. As can be seen in Figure B.5a, Jodel Karma is long-tailed to positive votes, whereas disliked posts naturally fall off around the post-remove-threshold [JH5]. We identify higher scores, indicating appreciation, in \mathcal{I} -*GenEntert* ($\uparrow 10$ for only up to 60%) and Θ -*BeliefsPol* ($\uparrow 10$ 78%). This evaluation also reveals that overall 17% posts are disliked, which is slightly deceiving as most Intents and topics are significantly below this 20% threshold.

B.5.6.5 Vote-consensus

The vote consensus is bound to $[-1,1]$, of which extremes indicate a high confidence in the community down- vs. upvotes. As can be seen in Figure B.5b, we observe an S-shape indicating that most posts experience equal up- & downvotes (including mostly none). However, there is an apparent skew towards positive consistency ($\approx 15\%$ of all posts have score 1.0). Whereas for most \mathcal{I} Intents cut off the S-shape between $\downarrow 0.02$ to $\downarrow 0.1$ on the lower end, they reach between $\uparrow 0.61$ to $\uparrow 0.79$ at the upper end; \mathcal{I} -*(Other, Info)* being outliers at (0.82, 0.88). This general observation likewise holds true for Θ topics. Nonetheless, we find most controversially down-voted posts within \mathcal{I} -*Other* ($\downarrow 0.10$, $\uparrow 0.75$), whereas Θ -*BeliefsPol* ($\downarrow 0.03$, $\uparrow 0.61$) are most controversial within upvotes.

B.5.6.6 Findings

From overall distributions across the board, conversationness is rather linear with an upper cut-off, the vote-consensus is s-shaped, whereas others are commonly heavy-tailed. Nonetheless, we find systematic differences in community reactions upon certain $\mathcal{I} \times \Theta$ combinations; confirmed by only few two-sided Kolmogorov-Smirnov tests between CDFs confidently confirming similarity over multiple metrics. We want to highlight various opposing outliers to $\Delta := (\mathcal{I}\text{-GenEntert} \times \Theta\text{-BeliefsPol})$ As for conversationness, we find most

| Type | #Hashtags | #Jodels pdf | DatingFlag |
|------------------------------|-----------|-------------|------------|
| Personal Information Sharing | 501 | 6.36% | 25% |
| Confessions | 76 | 1.17% | - |
| 18+ | 70 | 1.11% | 74% |
| Matchmaking | 43 | 1.02% | 100% |
| Debates & Opinion | 92 | 0.86% | 52% |
| Upvote Campaigns | 14 | 0.55% | - |
| Other | 219 | 2.85% | 3% |

Table B.4: Classification of the top 1015 hashtags.

homogeneous discussions especially within $\Delta \cup \mathcal{I}$ -*EntertObs*. The community often appreciates Δ -content with high confidence. In conclusion, Δ stands out: While being less discussed, discussions are more homogeneous. On the contrary, not spotting consistent outliers across all metrics, Θ -(*IllegalViolence, SocialMedia*) are discussed in longer threads with fewer participants, Θ -(*IllegalViolence, ProdService, FashionBeauty*) can expect most replies, \mathcal{I} -(*Seek, DistRelComp*) \times Θ -(*ProdService, Self, FitnessHealth*) receive the fewest votes. While the latter topics are most controversially dis-/liked, the same holds true for \mathcal{I} -*Info*. Lastly, the communities enjoy replying to \mathcal{I} -(*Self, Other*) \times Θ -(*IllegalViolence, FashionBeauty, FitnessHealth*).

B.5.7 How to Not Scale Out

Given our results, we also attempted to leverage supervised learning via SOTA pre-trained attention-based transformer masked language models, i.e., AraBERT [Antoun et al., 2020], and used data augmentation for increasing our sample size to create automatic classifications at scale. However, simple cross-entropy loss classification fine-tuning using an extensive hyperparameter search resulted in only 42% accuracy for topics on our imbalanced data set. Yet, being well *above chance*, we consider achieved results as insufficient for deeper reliable insights. Most downstream tasks base upon large amounts of data, whereas our expert classifications tend to be on the few-shot learning side. Currently being a hot research topic, it has been shown to generally perform rather poor compared to large-scaled counterparts [Luo et al., 2021, Wang et al., 2021]. Thus, we believe that more data may improve results—as has our dataset room for quality improvements as well. While computer-aided classification at large scale would be a desirable outcome, we argue that random sampling location and time creates a suitable representation of Jodel contents at substantial coder agreement to grasp users’ communication intents and discussed topics.

B.6 Jodel SA Hashtags

After a deep dive into quantitative insights and empirically peeking into community reactions upon content, we identified driving factors intents and discussed topics; yet we realized that it was missing crucial qualitative aspects of our annotators’ experience.

B.6.1 Qualitative Hashtag Classification

That is, we now add another vector of understanding: We provide deeper insights by leveraging that hashtags inherently carry categorical information [Ferragina et al., 2015]. Driven by observations and to better understand the impact of anonymity to the platform, we created a domain-specific annotation schema to capture sensitive contents and qualitatively coded the 1k topmost hashtags accordingly. As discussed earlier, elegantly crowdsourcing a well-suited anonymity-sensitivity score relying on many coders [Wang et al., 2014] is not possible in our case. From a random 1.12M thread subsample, we extracted all hashtags. Selecting the top used 1015 hashtags, we employed a single coder (age 20-25, male, Egypt)

to first qualitatively screen corresponding complete conversation threads to acquire personal impressions of typical associated contents and situations. In a next step, we created a domain-specific hashtag-annotation schema as shown in Table B.4, which we finally used to annotate our selected hashtags. While we show the absolute counts of #hashtags within each class, we also provide corresponding occurrences within the subsample (#Jodels PDF). According to the coder's feedback, he recognized a central recurring motif of vivid match-making, *dating and flirting*; thus, we added a *DatingFlag* correlating with this theme.

Our most prominent two hashtag categories confirm our previous findings that a user's intent often is driven by *Self* or *Seeking Info & Interaction*, which is also in line with the topics *Self* and *People & Relationships*, or combinations of both topics with *Distress Release* and *Social Ventening*. Albeit being very broad, most defined categories may be sensitive to anonymity giving *Debates & Opinions* the 18+ mark; except *Upvote Campaigns*³ and *Other*. We largely observe posts under such hashtags sharing personal experiences or confessions; other *DatingFlagged* topics often discuss love, sex, marriage, playful matchmaking, or games.

B.6.2 Findings

Based on this qualitative insight, we conclude that a considerable amount of 80% topmost hashtags relate to personal information or opinions that might not be posted in a real name environment, which is in line with previously shown main driving intents and topics—e.g., Seeking Information & Interaction, Dating & Flirting, sharing stories and questions about People & Relationships, while also using Jodel as a personal and social vent.

B.6.3 Selected Taboo Picks

Within our qualitative Hashtag study, we came across various topics, that may be considered as taboo. That is, we find evidence for:

- Self-relief & confessions about sexual harassment, encouraging others to share their experience ($p \approx 0.54\%$),
- Questioning forced wear of the Niqab ($p \approx 0.54\%$),
- Sparkling discussions about women driving ($p \approx 0.45\%$),
- Controversial discussions and questions about homosexuality and corresponding what-if scenarios ($p \approx 0.36\%$),
- Words of racism against foreigners ($p \approx 0.22\%$).

We find strong evidence of concrete discussed topics on Jodel within the KSA that probably would not have happened on any non-anonymous platform due to possible neglect, social pressure, and others out of manifold reasons.

B.6.4 The Story of Dating and Happy Marriages.

In light of the main driving intent being *Seeking Information and Interaction*, with Topics w.r.t. *People & Relationships* and *Self Expressions*, we investigate the before mentioned recurring motif: *dating & flirting*. Within our analyzed top 1k hashtags, 275 were annotated with the *DatingFlag* accounting for 45k Jodels ($p \approx 4\%$); digging deeper reveals a complete storyline along getting to know each other via games, dating, questions around kissing, marriage conditions, and intercourse.

³Invitation for gathering *Karma*, a lightweight in-app gamification.

B.6.5 Findings

Our observations are no exception to shown results on Whisper [Wang et al., 2014]. Also Jodel as an anonymous platform promotes sensitive content and provides a sphere where people are free in expression and more likely engage controversial discussions and opinions—one main reason using the application as concluded from interviews [Kang et al., 2016].

Furthermore, qualitative annotator feedback concludes that Jodel also allows for any question, giving advice—or provides ventilation for personal or social distress; Yet being a source of (local) contacts, potential matches, information, good stories, and jokes.

B.7 Conclusions

We created a schema and present our methodology for assessing *why* and *what* humans talk about in the anonymous and hyperlocal Jodel messaging app in the Kingdom of Saudi Arabia.

Unlike common beliefs and in line with research on other anonymous location-based platforms, anonymity does not necessarily lead to toxic content at large (e.g., hate speech). Popular topics in Jodel focus on information seeking, entertainment, people & relationships. Arguably, some mentioned topics can benefit from anonymity in a society that establishes certain taboos, e.g., casual discussions about the other sex or flirting. An anonymous platform can support such topics and enable an atmosphere in which users are free in their expressions as also shown in [Correa et al., 2015]. What they discuss can differ between cities, as shown by comparing Riyadh and Jeddah, with Riyadh having a broader spectrum of topics available. By evaluating votes (content appreciation) and replies (reactions), we show that the communities react differently to different topics; e.g., entertaining posts are much appreciated through votes, receiving the least replies, while beliefs & politics receive similarly few replies but are controversially voted.

Our study shows a lower-bound on the prevalent topics in an anonymous and hyperlocal messaging app. Our classification scheme enables future work to assess topical preferences more broadly.

“As for love, it still might always struggle to come out into the light of day in Saudi Arabia. You can sense that in the sighs of bored men sitting alone in cafés, in the shining eyes of veiled women walking down the streets [...], and in the heartbroken songs and poems, too numerous to count, written by the victims of love unsanctioned by family, by tradition, by the city: Riyadh” [Alsanea, 2007, pp. 313-314].

Acknowledgements

We are deeply thankful to all our students for their exceptional effort enabling this study: Ahmed Soliman, Haitham Almasri, Hana Al Raisi, Dalia & Mahmoud Moussa, Marcel Kröker, and Shams Dulaimi.

C The Role of Emoji

While our investigation so far has uncovered information diffusion properties within Jodel's independent community landscape, and providing insights to message contents and intents, today's social media communication and variation thereof is likewise interesting, such as emoji enlarging the expressiveness of text with a multitude of cues that can be encoded. We set out and present **empirical findings** of emoji usage across Jodel in the German and Saudi communities. Most emoji are used at the end of sentences presumably providing an **emotional connotation** combined with a heavy-tailed selection in emoji usage preferring only few specifically from the Emoji Unicode group of *Smileys & Emotions*. Interestingly, emoji color skin modifiers are used more often in the Saudi communities; in both countries---if used---we find light color to be most popular.

We train **word-emoji embeddings** on large scale messaging data obtained from the Jodel online social network. Our data set contains more than 40 million sentences, of which 11 million sentences are annotated with a subset of the Unicode 13.0 standard Emoji list. We explore **semantic emoji associations** contained in this embedding by analyzing associations between emojis, between emojis and text, and between text and emojis. Our investigations demonstrate anecdotally that word-emoji embeddings trained on large scale messaging data can reflect real-world semantic associations.

Further, we study the extent to which emoji can be used to **add interpretability to embeddings** of text and emoji. To do so, we extend the POLAR-framework that transforms word embeddings to interpretable counterparts and apply it to word-emoji embeddings trained on four years of messaging data from the Jodel social network. We devise a *crowdsourced human judgement* experiment to study six use-cases, evaluating against words only, what role emoji can play in adding interpretability to word embeddings. Leveraging **semantic differentials**, we use a revised POLAR approach interpreting words and emoji with words, emoji or both according to human judgement. We find statistically significant trends demonstrating that emoji can be used to interpret other emoji very well.

Introduction




As of now, we have looked into the concept of information diffusion through the Jodel community landscape, and uncovered about what and why the users in Saudi Arabia talk about. Often being neglected, today's Social Media communication uses more than words: emoji are all over the place enabling subtle cues. This is not a recent phenomenon, but roots within the 90's of using emoticons, where sequences of strings have been used to portray certain emotions [Albert, 2015, Read, 2005, Park et al., 2013], like the winking face smiley ;-). Some of these emerged within the Asian culture, like ^_^ which later translated to the beaming face with smiling eyes emoji. In fact, it has been shown that han-emoji have lots in common with emoji [8. Lee et al., 2019]. Research has developed a good understanding how emoji are used e.g., w.r.t. sentiment [Kimura and Katsurai, 2017], or semantics [Felbo et al., 2017a, Zhao et al., 2018]. With new capabilities and an extended expressiveness of newer Unicode Emoji Versions, e.g., skin tone modifiers were of specific interest. It has been researched how they affect semantics [Barbieri and Camacho-Collados, 2018], self-identifications [Robertson et al., 2018] and reader perception [Robertson et al., 2021b, Berengueres and Castro, 2017]. To mention few others, influences of context and culture [Takahashi et al., 2017], expression and perception [Li et al., 2019, Berengueres and Castro, 2017] and possibly misinterpretations [Miller et al., 2016b] or irony [Gonzalez-Ibanez et al., 2011] have been dissected. While the Unicode Consortium successively extends the Unicode Emoji palette, recent trends show users adopting a richer approach of Graphicons [Zhang et al., 2022]. We complement empirical measures in emoji usage [Lu et al., 2016, Ljubešić and Fišer, 2016] with insights into two spatially distinct countries in

the Jodel platform: Germany and Saudi Arabia in [★ C.1\) Social Media Emoji Usage](#) .

To make emoji tangible in computer processing methods, e.g., predicting emoji [Barbieri et al., 2017], work has focused on enriching emoji with meaning [Wijeratne et al., 2017a]. Enabled about a decade ago by advances within Natural Language Processing (NLP), word embeddings mapping words into a vectorspace [Mikolov et al., 2013a, Mikolov et al., 2013c, Pennington et al., 2014] were the new kids on the block, usually regularized on a word co-occurrence objective, yet still providing solid results in various downstream tasks as of today. Applications may be e.g., semantic similarity or sentiment [Berengueres and Castro, 2017, Kimura and Katsurai, 2017]. Emoji as a regular part of Unicode text, likewise have been projected into e.g., existing vector spaces [Eisner et al., 2016a], or directly incorporated into embedding training [Illendula and Yedulla, 2018, Ai et al., 2017]. They have been used to showcase semantic shifts in emoji semantics over time as well—learning from the model [Robertson et al., 2021a]. With expected rich semantic information encoded in Word-Emoji embeddings, we complement research w.r.t. emoji embeddings on the German Jodel dataset as an example of day-to-day contemporary casual communication. I.e., we extend this perspective with a qualitative and quantitative analysis of semantic emoji and text associations within [★ C.2\) Making Sense out of Emoji](#) .

There exists a dire need for introducing, adding or re-thinking explainable Artificial Intelligence methods. That is, interpretability of used tools can increase trust in models, prevent errors, or allows for better conclusion to be derived from the model itself. In the realm of Natural Language, neural embeddings have become state of the art. We finally show how we can leverage in-embedding semantic differentials to transform a classical word embedding into an interpretable counterpart. As has been shown for sentiment analysis, emoji introduce emotional expressiveness to the written language. Thus, we set out to apply an improved POLAR [Mathew et al., 2020] technique to a Word-Emoji Embedding on the Jodel DE dataset. Via crowdsourced campaigns, we find evidence that our approach is well in line with human judgement in [★ C.3\) Interpreting Emoji](#) .

Structure

-  [C.1: Social Media Emoji Usage](#)
-  [C.2: Making Sense out of Emoji \[JH7\]](#)
-  [C.3: Interpreting Emoji \[JH10\]](#)

| Country | Emoji Group | # [M] | counts | | textpos* | |
|---------|-------------------|--------|--------|--------|----------|------|
| | | | cdf | pdf | avg | std |
| DE | Smileys & Emotion | 61.13 | 1.0000 | 0.7603 | 0.72 | 0.27 |
| | People & Body | 10.68 | 0.2397 | 0.1328 | 0.59 | 0.33 |
| | Symbols | 1.68 | 0.1069 | 0.0209 | 0.28 | 0.26 |
| | Animals & Nature | 1.61 | 0.0860 | 0.0200 | 0.47 | 0.34 |
| | Travel & Places | 1.47 | 0.0660 | 0.0183 | 0.47 | 0.33 |
| | Food & Drink | 1.44 | 0.0477 | 0.0180 | 0.52 | 0.32 |
| | Objects | 1.20 | 0.0297 | 0.0150 | 0.51 | 0.34 |
| | Activities | 0.95 | 0.0148 | 0.0118 | 0.58 | 0.30 |
| | Flags | 0.24 | 0.0030 | 0.0030 | 0.37 | 0.29 |
| SA | Smileys & Emotion | 168.78 | 1.0000 | 0.8396 | 0.62 | 0.16 |
| | People & Body | 18.46 | 0.1604 | 0.0918 | 0.61 | 0.14 |
| | Travel & Places | 7.51 | 0.0686 | 0.0374 | 0.57 | 0.14 |
| | Animals & Nature | 2.02 | 0.0312 | 0.0101 | 0.61 | 0.17 |
| | Activities | 1.77 | 0.0212 | 0.0088 | 0.59 | 0.13 |
| | Objects | 1.07 | 0.0124 | 0.0053 | 0.60 | 0.17 |
| | Food & Drink | 0.80 | 0.0071 | 0.0040 | 0.62 | 0.16 |
| | Symbols | 0.41 | 0.0031 | 0.0020 | 0.66 | 0.18 |
| | Flags | 0.22 | 0.0011 | 0.0011 | 0.73 | 0.17 |

Table C.1: Emoji Usage by Unicode Group - DE & SA. *Normed values for arabic *left-to-right* for better comparison.

C.1 Social Media Emoji Usage

Within nowadays Social Media and personal communication, emoji find widespread use. While emoji are primarily used to connote texts with emotion, emoji can also replace text elements. The Jodel users add emoji to about 11.80% of all posts within Germany, while this value increases for Saudi Arabia to 34.93% total. That is, within this section, we will explore emoji usage on Jodel within Germany and Saudi Arabia from an empirical perspective w.r.t. pure popularity, emoji per post, emoji text position, and emoji skin tone modifiers.

C.1.1 Emoji Usage by Type (Unicode Group)

The Unicode standard for emoji provides a hierarchical class for emoji: the emoji subgroup and group providing a rough categorization into proposed emoji functions/semantics. That is, e.g., the group of *Smileys & Emotion* contains several subgroups for facial emoji, such as *face-smiling* or *face-affection*.

We count all used emoji within Germany and Saudi Arabia according to emoji group and determined the occurrences within the corpora as presented in Table C.1. Further, we deduct frequencies via the Probability Distribution Function (PDF), and the Cumulative Distribution Function (CDF) for each of the countries separately.

For both countries, we observe a heavy bias towards *Smileys & Emotion* being used for 76% (84%) in DE (SA). While *People & Body* is the second favored emoji group at 13% (9%) in DE (SA). This is followed by *Animals & Nature* and *Travel & Places*, while the popularity of other emoji groups differs between both countries, except for flags. However, our findings suggests that emoji usage beyond the top-2 groups remains generally small at cumulative top2 percentages of 89.31% (93.14%) in DE (SA). I.e., in both countries, most users prefer using emoji to express emotion or adding further context, e.g., via hand-signs, or people/professions.

C.1.2 Emoji Popularity

While we have discussed a deeper insight into a group perspective, we are next interested in a more detailed perspective incorporating single emoji. That is, we present the distribution in single emoji usage within Figure C.1a as a Complementary CDF. While the logarithmic

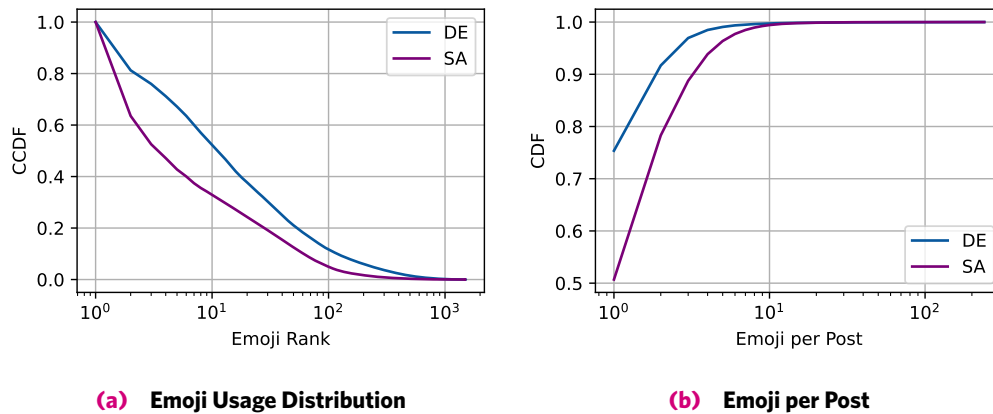


Figure C.1: Emoji Distribution & Emoji per Post - DE & SA. (a) *Left:* Overall emoji usage appears much broader within the DE communities, while most popular emoji within SA are even more heavy tailed. (b) *Right:* On average, the SA users use more emoji per post.

x-axis denotes the emoji rank, the y-axis denotes the proportion along all used emoji. First, we observe heavy-tailed distribution in emoji usage in Germany as well as in Saudi Arabia, while these distributions show a heavier skew in SA; I.e., while the top 10 emoji in DE only account for about 50% of all emoji, they account for about 70% in SA. The vast majority of emoji is only used rarely in comparison.

Further, going into more detail within Table C.2, we showcase the top 10 emoji for DE and SA. Here, we not only show the frequency of each emoji via the PDF, we also include measures for each emoji's occurrences within its Unicode Group (G) and Subgroup (S). To provide an example for the German communities, we observe that 😊 enjoys much popularity at a total of about remarkably 19% of all emoji; while it accounts for about 49% of all within-Group emoji (*Smileys & Emotion*), this also accounts for about 19% within its subgroup (*face-smiling*). Correspondingly, the second most popular emoji 😄 makes up for about 5% in total and within-subgroup, while accounting for 14% within its group.

Overall, the distribution of emoji within their respective Unicode Group and Subgroup also follows a heavy-tailed distribution (not completely shown) and focuses on *Smileys & Emotion*. Related work confirms this finding for other platforms, e.g., Twitter [Ljubešić and Fišer, 2016], and specifically also between western and eastern cultures [Guntuku et al., 2019].

As seen within the discussed CDF, emoji usage is even more skewed towards single items within the Saudi communities compared to the German user base. While the top 10 emoji in DE all belong to the same group of *Smileys & Emotion*, this is similar for the Saudi users. However, the specific ranking of single emoji has similarities, but occurrences of specific items are dissimilar. That is, the Saudi user base uses more emoji having a rather negative sentiment [Novak et al., 2015]: 🤔, 😞, and 😓.

C.1.3 Emoji per Post

So far, we have seen which emoji are used to what extent. However, it still remains unclear how many emoji the users add to posts. As stated earlier, the SA users prefer adding emoji to their contents about three times as much as the DE users. We confirm this trend with Figure C.1b showing how many emoji are being used on a per-post basis as a CDF. While the x-axis denotes the amount of emoji per post, the x-axis denotes the corresponding cumulative amount if posts.

While about 75% of all emoji posts in Germany contain a single emoji, up to 90% of all posts include up to two emoji. In contrast, the Saudi users use emoji a lot more: while only 50% of emoji posts contain a single emoji, about 80% of these posts contain up to two emoji.

| Group | G# [k] | Subgroup | S# [k] | Emoji | # [k] | S pdf | G pdf | pdf |
|------------------------|---------|----------------|--------|--------|--------|--------|--------|--------|
| DE | | | | | | | | |
| Smileys & Emotion | 61,133 | face-smiling | 30,849 | 😊 | 15,076 | 0.1875 | 0.4887 | 0.1875 |
| | | | | | 4,240 | 0.0527 | 0.1374 | 0.0527 |
| | | face-affection | 5,497 | 😘 | 3,841 | 0.0478 | 0.6988 | 0.0478 |
| | | | | | 3,658 | 0.0405 | 0.8911 | 0.0405 |
| | | monkey-face | 3,658 | 🙈 | 3,260 | 0.0405 | 0.8911 | 0.0405 |
| | | | | | 3,658 | 0.0405 | 0.8911 | 0.0405 |
| | | face-smiling | 30,849 | 😊 | 2,883 | 0.0359 | 0.0935 | 0.0359 |
| | | | | | 3,285 | 0.0350 | 0.8558 | 0.0350 |
| | | face-hand | 3,285 | 🙌 | 2,812 | 0.0350 | 0.8558 | 0.0350 |
| | | | | | 3,285 | 0.0350 | 0.8558 | 0.0350 |
| face-smiling | 30,849 | 😊 | 2,458 | 0.0306 | 0.0797 | 0.0306 | | |
| | | | 2,016 | 0.0251 | 0.0653 | 0.0251 | | |
| face-neutral-skeptical | 4,677 | 😐 | 1,834 | 0.0228 | 0.0594 | 0.0228 | | |
| | | | 1,543 | 0.0192 | 0.3300 | 0.0192 | | |
| SA | | | | | | | | |
| Smileys & Emotion | 168,778 | face-smiling | 84,637 | 😊 | 73,311 | 0.3647 | 0.8662 | 0.3647 |
| | | | | | 43,617 | 0.1099 | 0.5067 | 0.1099 |
| | | emotion | 43,617 | ❤️ | 10,837 | 0.0539 | 0.2485 | 0.0539 |
| | | | | | 19,732 | 0.0440 | 0.4481 | 0.0440 |
| | | face-concerned | 19,732 | 😞 | 8,843 | 0.0440 | 0.4481 | 0.0440 |
| 84,637 | 0.0271 | | | | 0.0643 | 0.0271 | | |
| Travel & Places | 7,513 | sky & weather | 7,275 | 🌤️ | 5,316 | 0.0264 | 0.7306 | 0.0264 |
| | | | | | 19,732 | 0.0184 | 0.1873 | 0.0184 |
| Smileys & Emotion | 168,778 | face-concerned | 19,732 | 😞 | 3,696 | 0.0184 | 0.1873 | 0.0184 |
| | | | | | 4,215 | 0.0137 | 0.6530 | 0.0137 |
| | | face-affection | 4,215 | 😘 | 2,752 | 0.0137 | 0.6530 | 0.0137 |
| | | | | | 6,339 | 0.0122 | 0.3881 | 0.0122 |
| face-neutral-skeptical | 6,339 | 😐 | 2,460 | 0.0122 | 0.3881 | 0.0122 | | |
| | | | 43,617 | 0.0118 | 0.0546 | 0.0118 | | |
| emotion | 43,617 | 💜 | 2,382 | 0.0118 | 0.0546 | 0.0118 | | |
| | | | 2,382 | 0.0118 | 0.0546 | 0.0118 | | |

Table C.2: Top 10 Most Popular Emoji - DE & SA. The emoji usage distributions are heavy tailed, but differ in their magnitude. For both countries, facial emotion emoji dominate the top 10. While DE users mostly use positive expressions, the SA platform content appears more balanced w.r.t. positive/negative emotion.

Hence, the Saudi users not only use emoji more often—they also use emoji more extensively within a single posts due to lots of multiple occurrences.

C.1.4 Emoji Text Position

Next, we focus on in-text emoji positions that convey the emoji purpose to a certain extent. We measure the relative text position of emoji according to total posts lengths; note that we adjusted for Arabic text in Saudi Arabia by flipping these values (due to right-to-left text). While we have added aggregated averages and respective standard deviations of the relative in-text positions in Table C.2 per emoji group, we display the detailed distributions in Figure C.2.

Across the board, most emoji are almost exclusively added to the end of a post within both countries. We presume that most of these emoji serve the purpose of adding connotations to the text, which becomes very apparent especially for *Smileys & Emotion* and *People & Body*, while this is not the case for *Symbols* (at least in DE); *Flags* are also better distributed across posts. Specifically within SA, we observe a heavier shift towards the end of posts.

This does not come unexpected: Research over the past decade has shown that emoji carry rich emotional information often being used and improving sentiment analysis [Felbo et al., 2017a, Kimura and Katsurai, 2017].

C.1.5 Emoji Skin Tones

While first Unicode Emoji versions were limited, the consortium has successively added more emoji aiming for better inclusivity. Emoji version 2.0 [Unicode Consortium, 2016] introduced skin tones. Later in 2016, Emoji version 4.0 [Unicode Consortium, 2016] added Zero Width Joiner Sequences (ZWJS). They enable coupling multiple emoji—e.g., defaulting

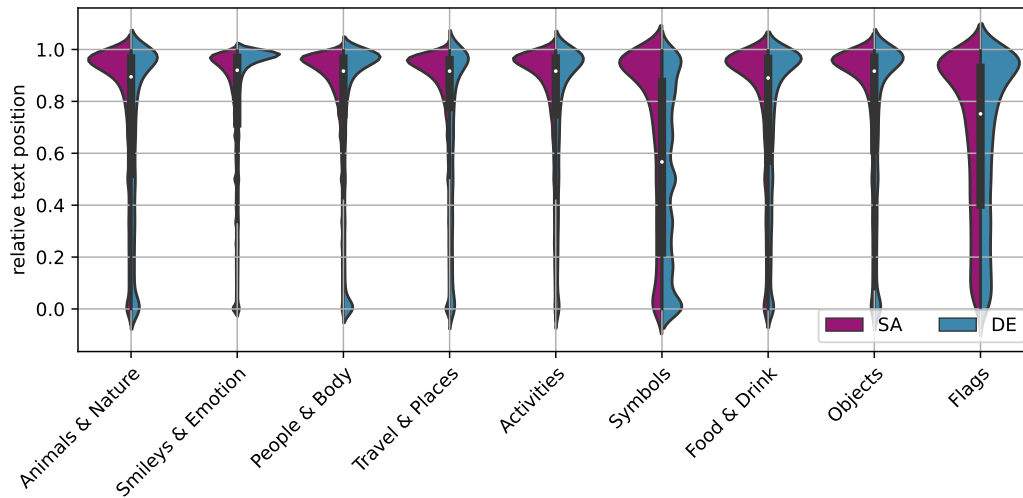


Figure C.2: Emoji relative text position by Unicode Group - DE & SA. 10M Random Sample. For enabling a qualitative comparison, the SA series have been normalized due to Arabic *right-to-left* text. We find most user placing emoji at the end of a post across all groups for both countries.

to neutral, a distinction between men and women, various skin tones, or complex family compositions (possibly two parents and two children times skin tone times gender).

While detailing any of such sequences would be interesting, we take a first look into skin tones in line with other recent research: Specifically skin tone modifiers have been of interest, e.g., how skin tone impacts semantics on Twitter [Barbieri and Camacho-Collados, 2018], or variation in emoji usage by Robertson et al. [Robertson et al., 2020]. Moreover, using a skin tone has influence on both: the sender and the receiver. That is *i*) they also analyzed the extent to which these modifiers function as a tool for self-representation [Robertson et al., 2018], and *ii*) dissected how readers perceive the usage of skin tone modifiers [Robertson et al., 2021b].

We have aggregated all emoji occurrences that allow for a skin tone modifier on Jodel across Germany and Saudi Arabia. In Figure C.3, we present the frequencies (pdf) of the neutral versions, i.e., without modifier, and the distinct other options: *light*, *medium-light*, *medium*, *medium-dark*, and *dark*.

Half of the German user base do not use skin tone modifiers, but prefer the neutral versions. Lighter skin colors are equally popular accounting for 20% each; dark color is rarely used. Interestingly, there is an apparent difference to the SA usage. Here, we find only about 18% of neutral usage, whereas specifically the light color is most popular at about 56%; followed by the medium modifier accounting for another 15%. This is interesting as Robertson et al. have shown that chosen skin tones often correspond to the users actual skin tone as a function of self representation. The difference between DE and SA for (not) using skin tones at all, or specifically if—why—would be very interesting and provides a perfect stage ground for more in-depth research. However, we will leave this topic for future work.

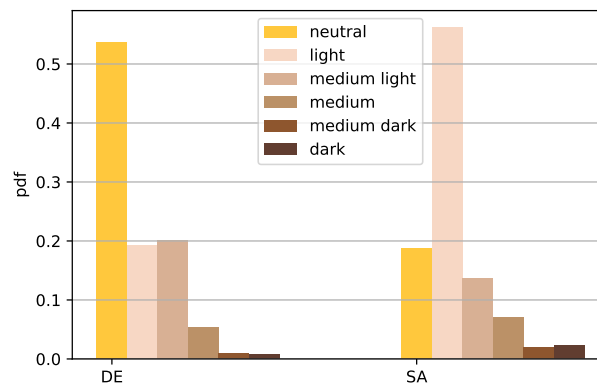


Figure C.3: Emoji Skintone Modifier Usage - DE & SA.

C.2 Making Sense out of Emoji

While we have focused on information diffusion through the Jodel community landscape and uncovered actual message intents and topics, we next are curious about social media language variations, i.e., emoji usage. Having provided evidence of emoji prevalence and given proven strengths in adding value to e.g., sentiment analysis, we make emoji computationally feasible by training word-emoji embeddings on large scale messaging data obtained from the Jodel online social network. Our data set contains more than 40 million sentences, of which 11 million sentences are annotated with a subset of the Unicode 13.0 standard Emoji list. We explore semantic emoji associations contained in this embedding by analyzing associations between emoji, between emoji and text, and between text and emoji. Our investigations demonstrate anecdotally that word-emoji embeddings trained on large scale messaging data can reflect real-world semantic associations. To enable further research we release the Jodel Emoji Embedding Dataset (JEED1488) containing 1488 emoji and their embeddings along 300 dimensions.

C.2.1 Introduction

To make word usage and contents technically tangible, word embeddings [Mikolov et al., 2013a, Mikolov et al., 2013c, Pennington et al., 2014] are a common tool. However, given the prevalence of emoji within today’s social media, we target to add emoji to such word embeddings, usually regularized on a word co-occurrence objective, provide good results for various downstream tasks; e.g., sentiment [Berengueres and Castro, 2017, Kimura and Katsurai, 2017]. Being part of text, emoji have been embedded in various research advances [Eisner et al., 2016a, Illendula and Yedulla, 2018, Ai et al., 2017], followed by downstream tasks, such as sentiment analysis [Hu et al., 2017b, Novak et al., 2015]. Embedding enable identifying semantic shifts in emoji semantics over time [Robertson et al., 2021a]. Yet, context and culture determine emoji semantics [Takahashi et al., 2017] or expression and perception [Li et al., 2019, Berengueres and Castro, 2017].

With expected rich semantic information encoded in word-emoji embeddings, we complement research w.r.t. emoji embeddings on the German Jodel dataset as an example of day-to-day contemporary casual communication.

Structure [JH7]

- ☰ [C.2.2: The Jodel Emoji Embedding Dataset \(JEED1488\) \[SD1\]](#)
- 📁 [C.2.3: Related Work](#)
- 🗨️ [C.2.4: Word-Emoji Embeddings](#)
- 🔗 [C.2.5: Emoji2Emoji Associations](#)
- ☰ [C.2.6: Emoji2Text Associations](#)
- 😊 [C.2.7: Text2Emoji Associations](#)
- 📌 [C.2.9: Conclusions](#)

C.2.1.1 Research Questions

How well do emoji reflect real-world semantics on a Social Media word-emoji embedding? To which extent can emoji add value to embedding spaces, given its prevalence in social media usage?

C.2.1.2 Approach

We use four years of complete data from the online social network provider Jodel to explore semantic associations of emoji, i.e., expressive icons, embedded in textual messaging data. The dataset data contains 48M sentences derived from the users' public posts, of which 11M have at least one emoji. We deploy Word2Vec to generate combined word-emoji embeddings and use the obtained embeddings to gauge the ability of word-emoji embeddings to capture different kinds of semantic emoji associations.

Associations between emoji. What kind of associations between emoji are reflected by word-emoji embeddings of messaging data? We explore t-SNE projections of emoji associations (Figure C.4) and interpret them qualitatively.

Associations between emoji and text. What words are associated with a given emoji in a word-emoji embedding of messaging data? We explore the textual semantics of emoji by deriving top k words that are most similar to a given emoji.

Associations between text and emoji. What emoji are associated with a given word in a word-emoji embedding of messaging data? We train machine learning models to predict an emoji for a given word and evaluate our results employing k-fold cross-validation.

C.2.1.3 Results

Our qualitative results show that emoji to emoji embeddings reveal insightful semantic associations beyond the Unicode standard. Our results highlight that quality emoji to text translations can be obtained from embeddings, e.g., to improve typing prediction on mobile devices or to inform social network users on emoji meanings in their network. Our results show that for text to emoji, machine learning improves accuracy compared to a naive direct embedding approach at the cost of additional training.

These associations reflected by word-emoji embeddings trained on large scale message data open up a range of interesting downstream tasks and prospects, such as text to emoji translations, or emoji recommendation and replacement.

C.2.2 The Jodel Emoji Embedding Dataset (JEED1488) [SD1]

To enable further research, we release a subset of our embeddings to encourage and support further research into real-world semantic emoji associations. This Jodel Emoji Embedding Dataset [SD1] containing 1488 emoji and their embedding along 300 dimensions based on word-emoji co-occurrence in a large messaging corpus.

C.2.3 Related Work

emoji are widely studied, e.g., analyzing their semantics via embeddings [Ai et al., 2017]. To mention few others, empirical measures on emoji usage [Lu et al., 2016, Ljubešić and Fišer, 2016], influences of context and culture [Takahashi et al., 2017], expression and perception [Li et al., 2019, Berengueres and Castro, 2017] and possibly misinterpretations [Miller et al., 2016b] or irony [Gonzalez-Ibanez et al., 2011]. Another topic is sentiment analysis on social networks that often is performed on a word level, but has also attracted incorporating emoji [Hu et al., 2017b, Novak et al., 2015].

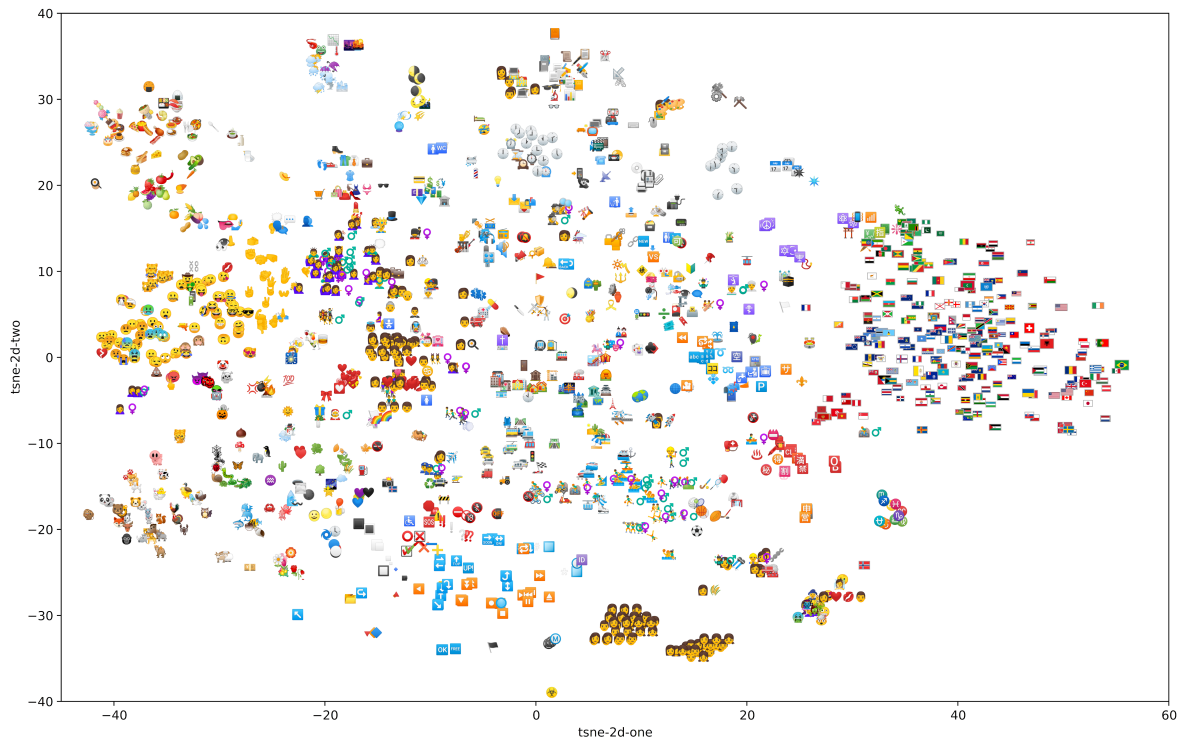


Figure C.4: Emoji2Emoji map. It shows emoji associations as a 2D projection by their corresponding embedding vectors using the t-SNE algorithm. The figure highlights well formed semantic-based clusters of emoji associations: e.g., food (-40, 25) and fruits (-35, 20), animals (-35, -20), negative emotions faces (-40, 0), positive emotion faces (-30, 5), love, marriage, babies and families (-15, 0), country flags (50, 0), weather (-20, 35), or various alike signs (-5, -30). We observe that the embedding can uncover semantic relations between emoji from social network postings.

C.2.3.1 Emoji2Emoji

The Emoji2Emoji task was analyzed on e.g., Twitter data [Illendula and Yedulla, 2018, Barbieri et al., 2016]. Barbieri et al. define two questions: *i*) topical similarity—do two emoji occur at the same time? And *ii*) functional similarity—can two emoji be used interchangeably? Its evaluation leverages human judgment for both questions of 50 selected emoji pairs and achieves an accuracy of 78% for both tasks. A qualitative evaluation is given by a clustered t-SNE evaluation. In [Wijeratne et al., 2017b], Wijeratne et al. extend their prior work on a knowledge database of emoji semantics [Wijeratne et al., 2017a]. They add *sense* information into an emoji embedding, which is validated on 508 emoji pairs in terms of *i*) and *ii*) via crowdsourcing. Their model mostly is strongly correlated to human judgment. As they have published the annotation results, we find our used embedding instance providing similarities only moderately in line with human judgment.

C.2.3.2 Emoji2Text

[Eisner et al., 2016a] use textual descriptions of emoji to map them into the Google News Dataset word embedding. By doing so, they obtained 85% accuracy in predicting keywords for manually labeled emoji. An emoji knowledge base extracted from Google News and Twitter data including sense labels has been presented in [Wijeratne et al., 2017a]. Within a multi-staged crowdsourced human evaluation, they show 83.5% valid sense assignments.

C.2.3.3 Text2Emoji

E.g., [Felbo et al., 2017b] create a word-emoji embedding using LSTM on Twitter data. They showcase a Text2Emoji downstream task predicting one out of 64 emoji by applying deep

learning and achieve a Top 1 (5) accuracy of 17.0% (43.8%) outperforming fasttext-only which yielded accuracy values of Top 1 (5) of 12.8% (36.2%) on a balanced subset. Another work [Zhao et al., 2018] use a multi-modal deep learning model that predicts an emoji and its position within the text also leveraging image and user demographic information. This approach predicts the correct emoji out of 35 most frequent emoji with a Top 1 (5) accuracy of 38% (65%). Further, e.g., [Barbieri et al., 2017] used an LSTM for prediction, whereas [Guibon et al., 2018] use tf-idf and argue that an enrichment with sentiment features improves prediction quality.

We complement related work by performing all 3 tasks to study real-world semantic associations on a large-scale messaging data.

C.2.4 Word-Emoji Embeddings

Word embeddings are semantic vector models that map written language information into an n -dimensional vector space representation. They have become a popular tool for both, industry and academia, e.g., for finding word associations or sentiment analysis.

Beyond classical embedding of words, emoji can likewise be incorporated in this process [Eisner et al., 2016a, Wijeratne et al., 2017a]. Different off the shelf approaches exist to create word embeddings (e.g., GloVe, fasttext, or Word2Vec) of which we use the gensim Word2Vec implementation due to its simplicity and popularity; the algorithm was proposed in [Mikolov et al., 2013b]. Word embedding is thus a promising approach to study emoji associations on Jodel.

Our approach involves two steps: *i*) data preprocessing and cleaning, and *ii*) creating the embedding.

C.2.4.1 Step 1: data preprocessing and cleaning

First, we dropped any skin color modifier to predict the type of emoji only. While preserving all emoji, we next cleaned the data by applying stemming and lemmatization using spaCy (built-in German model) and filtered out any non-alpha characters. We discard words less than 3 characters. Sentences consisting of less than two words were also discarded as they do not convey any information for our word embedding later. This resulted in 42.2M usable sentences.

C.2.4.2 Step 2: embedding creation

We create the word embedding by applying the gensim Word2Vec library to the entire corpus of usable sentences. Beyond default settings, we apply higher sampling to decrease the weight of most frequent words. Then, we trained the embedding model for 30 epochs; note that we discuss the variation of training in the next Section-*Varying the number of epochs*.

Next, we show that the resulting embeddings can be applied to accomplish three associative downstream tasks: *Emoji2Emoji*, *Emoji2Text*, and *Text2Emoji*.

C.2.5 Emoji2Emoji Associations

We begin by analyzing how emoji are semantically associated with each other on Jodel. That is, can we explain an emoji purely using other emoji. These associations can help social-network users to understand the *subjective* meanings of an emoji in their local user community.

C.2.5.1 2D view on emoji associations

We show the Emoji2Emoji associations as a 2D projection of all emoji by their embedding vectors using a t-SNE visualization in Figure C.4 [Maaten and Hinton, 2008]. The t-SNE algorithm iteratively clusters objects within a high dimensional vector space according to their distance and projects them to the selected number of dimensions—here 2D. The shown Figure is a hand-picked version of multiple projections with different random seeds for projection. Note that the rendering of complex emoji may be executed sequentially, e.g., 🍷 is represented as 🧑❤️🍷, or the profession 🧑 is mapped to 🧑👨.

We observe semantic-based clusters of emoji associations as they occur on Jodel. The distance within the projection often *can* be interpreted as semantic proximity, e.g., there exist groups of almost all facial emoji (-35, 5). To mention few other examples, clusters (and their coordinates) involve food (-40, 25) and fruits (-35, 20), animals (-35, -20), negative emotions faces (-40, 0), positive emotion faces (-30, 5), families (-15, 0), country flags (50, 0), weather (-20, 35), or various alike signs (-5, -30).

Next to the cluster of families (👨, 👩) (-15, 0), we find love (❤️) (-20, -5), LGBT (🏳️🌈, 👨, 👩) (-15, -5), marriage (👰, 💍, 🎊) (-10, 5), pregnancy and children (👩, 🧒, 🍼) and even grandparents (👴, 👵) next to it.

For another example, the cluster of sports emoji (e.g., 🏀, 🏊) (10, -15) and especially water related sports (🏊, 🏄, 🏊) show that *holidays* (🏖️) may also closely be associated with flights (✈️, 🛫), sea-born emoji 🚢, the globe (🌍, 🌎), other attractions (🎡, 🎢) and adventures (🏠, 🏠). Further, next to any sort of biking (🚲) as a sport, we observe transportation related emoji (🚗, 🚆) that may also indicate holidays.

This represents a rich set of semantic associations derived from millions of posts reflecting real-world emoji semantics and there are lots of details to be discovered in each tSNE projection instance. However, while these associations are promising, the involved dimensional reduction *may* oversimplify the underlying vector space, consequently, it *may* pretend semantic associations that do not exist. Thus, we next take a deeper look into emoji similarity by analyzing their distances within the embedding space.

C.2.5.2 Emoji similarity

To analyze emoji similarity, we first present a selection of emoji from different emoji groups (according to the Unicode definition) and their top k most similar emoji from the embedding having a document frequency above 100 in Table C.3.

Most semantic associations are matching quite good, e.g., 🍹 to other drinks, good vibes and party; 🏕️ to a tent, traveling and outdoor activities; 😊 to other positive emotions; etc. However, some symbols may have several semantic associations: while 🛠️ is related to other tools, the German word for hammer may also be used in explicit language as a synonym for the alternate meaning of 🍆 as a phallic symbol. Another example can be seen in 🏰, which relates to mountains and attractions, but to presumably Harry Potter 🧙 as well. The 🍏 is mostly matched with other fruits. However, the association to 🍆 relates to the mentioned alternative usage of both as symbols unrelated to the actual fruits, which may be in line with 🍷, 💧 and 💋. In fact, the top 20 set contains more possibly fruit-unrelated emoji such as 👁️, 🛠️ and 😊. Other fruits in this set (🍏, 🍉, 🍎 and 🍊) may thus be reinterpreted in a different context as well; Anyhow, this is a good example where the raw embedding is not well suited to distinguish between multiple semantics for a single item.

C.2.5.3 Aggregation by Unicode groups

To better understand these similarities, we next aggregate emoji into their (sub-)group according to the Unicode definition. Therefore, we show the confusion matrix of the topmost

| Emoji | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-------|---|---|---|---|---|---|---|---|---|----|
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |

Table C.3: Emoji top 10 similarity examples We show a hand-picked selection of emoji and their closest emoji according to their vector cosine distance in our embedding. For most emoji, we observe good semantic relations. However, some emoji may have multiple semantics (castle---lock, home, traveling) or no apparent topic (plus sign).

similar emoji pairs aggregated into their groups in Figure C.5. That is, we calculate the most similar emoji to each emoji contained in the dataset. Then, we count the in-/ correct mappings whether the most similar emoji's (x-axis, most similar) group is equal to the group of the compared emoji (y-axis, target). Due to the imbalance within the number of available emoji in each group, we further normalized the mappings to the total number of results emoji per group.

We observe a strong straight diagonal indicating that most emoji are associated with other emoji in the same Unicode group (average precision of 0.8466 ± 0.0932). Deviations can mostly be explained by associated emoji located in different Unicode (sub-)groups (most notable for the Activities subgroup that has the amongst the lowest similarity scores). Example associations between different groups (noted in parentheses) include 🌲 (Activities) to 🧑 (person), ✨ (Activities) to 🌟 (Animals & Nature), 🎀 (Activities) to 🍷 (Smileys & Emotion), 🧶 (Activities) to 🍷 (Objects), ⚽ (Activities) to 🇩🇪 (Flags). These results show that emoji embeddings on social network posts can reveal semantic associations beyond those captured by Unicode groups.

Repeating this evaluation on a subgroup level results in the same observation (precision of 0.5918 ± 0.3152). Not surprisingly as seen within the t-SNE visualization in Figure C.4, the main driver of confusion are the facial emoji in different subgroups.

C.2.5.4 Comparing embedding similarities with human judgment

As last step, we compare our embedding and obtained emoji similarities to human perception leveraging the EmoSim 508 dataset from [Wijeratne et al., 2017b]. It was obtained by asking 10 subjects to provide a similarity score for 508 well-selected different emoji pairs. To evaluate the suitability of an alike emoji-embedding, they apply the spearman rank correlation test achieving significant correlations for different model instances between 0.46 and 0.76.

Applying the very same test to our embedding, our instance resulted in a moderate correlation of 0.5349 with a very high significance, which is in line with Wijeratne et al. presented results. Anyhow, the asked user-base is not platform-specific to Jodel mismatching the used embeddings, which may make this comparison less representative. Eventually, by

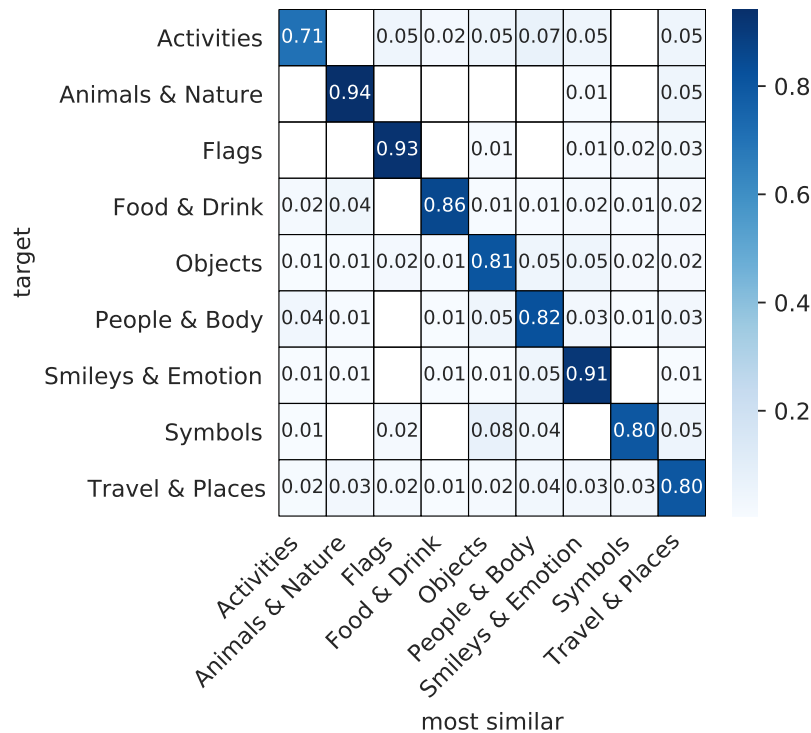


Figure C.5: Emoji2Emoji Unicode group association confusion matrix. This heatmap shows emoji top1 associations from our embedding aggregated to their Unicode group as a confusion matrix. The values are normed row-wise. We observe an overall good similarity for the groups from our embedding depicted by the strong diagonal line.

adjusting model parameters, this value can likely be improved.

C.2.5.5 Varying the number of epochs

While common wisdom may suggest *more is better*, our evaluation shows that the number of trained epochs impacts the results. By comparing the Emoji2Emoji and especially emoji2Text results from embeddings that have been trained different numbers of epochs (5..100), we observed a negative shift in perceived capability to reflect *multiple semantics* / synonyms with more training. I.e., more training seems to focus the embedding to more distinct topics and tends to remove synonyms / emoji with multiple semantics. Thus, depending on the application, both variants may be desired and fine-tuned.

C.2.5.6 Summary

Our results show that Emoji2Emoji embeddings on social media posts reveal insightful associations that go beyond semantic groups in the Unicode standard. We posit that these associations are useful, e.g., to understand the usage of each emoji in a given social media platform (e.g., for its users).

C.2.6 Emoji2Text Associations

Next, we use our embedding to associate Emoji2Text. One use case of this association is to improve keyboard predictions on mobile devices. Another one is to provide social network users a better understanding of the meaning of emoji in the target social media platform, by explaining emoji with words—which can be more descriptive the previously presented Emoji2Emoji associations.












| Class | Emoji | Top matching words (<i>left to right</i>) |
|----------|---|--|
| single |  | bier, prost, nächst, mal, mein, gut, erstmal, letzt, schön, abend <i>beer, cheers, next time, my, well, first, last, fine, evening</i> |
| single |  | camping, campen, zelten, hurricane, fernweh, festivals, reisetipps, verreisen, bereisen, urlaubsziel <i>camping, camping (v), tents, (festival), wanderlust, festivals, travel, travel tips, traveling, visit, travel destination</i> |
| single |  | enttarnen, verdächtigen, entlarven, undercover, auffällig, sherlock, mysteriös, ermitteln, handeln, geheime <i>expose, suspect, unmask, undercover, conspicuous, sherlock, mysterious, investigate, act, secret</i> |
| single |  | all, schön, the, gleichzeitig, sogar, bestimmen, and, bunt, ehe, gerne <i>all, beautiful, the, simultaneous, even, determine, and, colorful, marriage, gladly</i> |
| single* |  | booty, butt, po, hintern, boobs, ass, brüste, tanga, datass, brüste <i>(buttocks coll.), (breasts coll.), thong, (buttocks coll.), breasts</i> |
| multi |  | hammer, hämmern, hörmalwerdahämmert, lörrer, presslufthammer, dlrh, werkzeug, vorschlaghammer, nageln, ding <i>hammer, hammering, (TV show), (penis), pneumatic hammer, (coitus, platform), tool, sledgehammer, nail, thing</i> |
| multi |  | burg, hogwarts, ilvermorny, heidelberg, schloss, prinz, königreich, insel, emoji, beasts <i>castle, hogwarts, ilvermorny, heidelberg, palace, prince, kingdom, island, emoji, beasts</i> |
| omni |  | hab, mal, weiß, einfach, sehen, eigentlich, fragen, echt, denken, grad <i>just, sometimes, white, just, see, actually, ask, really, think, just</i> |
| omni |  | zusammenfügen, vollständig, erhalten, einstellungen, aktuell, auswählen, klickt, siehe, vorhanden, hinzufügen <i>merge, complete, get, settings, current, select, click, see, available, add</i> |
| omni |  | späß, gerne, bitte, falls, suche, ps, suchen, danke, schön, zumindest <i>fun, gladly, please, if, search, ps, search, thanks, beautiful, at least</i> |
| platform |  | manni, busfahrer, racingteam, mannis, linie, kvgracingteam, busse, racing, hvvracingteam, formelaseag <i>(name), bus driver, racing team, (name), bus line, (local transportation), busses, racing, (local transportation)</i> |

Table C.4: Emoji2Text association examples This table shows the top 10 words to a set of emoji filtered by a minimum document frequency of 500 in our dataset. We show the original German words and an English translation in italics below. This hand-picked selection aims to cover a broad range of emoji in four classes of emoji semantics: *i)* single-semantic emoji, *ii)* multi-semantic emoji, *iii)* omni-semantic emoji where the associations do not make an apparent sense, and *iv)* emoji associated with platform-specific idioms. The resulting word lists often provide a good textual emoji representation, whereas multiple semantics may mix.

To give a first insight into the Emoji2Text associations, we study the top matching words for a set of emoji from different groups. Our first observation is that some of the top matching words are very specific to the Jodel platform, e.g., 🍷 → *roterweihnachtsjodel* (red x-mas Jodel), or ⚽ → *grünerfußballjodel* (green soccer Jodel). This reflects specific user habits of posting a word/noun as a Jodel post in relation to an underlying in-app post background color. Evaluated by ourselves, other words can often be semantically linked to the specific emoji, whereas some cannot.

To generalize and improve results, we filtered the resulting words by their document frequency. By adjusting this value to 500, we achieve better matchings according to our interpretation, yet we still find platform-specific and composite words within these sets as they occur quite often. Some emoji are used in a more specific sense than others, which we want to clarify by giving hand-picked examples in Table C.4. This table shows an emoji with its Top 10 words (frequency greater than 500) that are closest within the embedding according to the cosine vector distance. While we provide the actual German words, we also give translations below each of these words.

To perform a first qualitative Emoji2Text analysis, we introduce 4 *exemplary* emoji classes based on their semantic variability in the top 10 words and their Emoji2Emoji association as determined by us: *i)* *single semantic emoji* have only a single association, *ii)* *multi-semantic emoji* have multiple associations (meanings), *iii)* *omni-emoji* having no specific apparent semantic, and *iv)* *platform-specific semantic emoji* have a Jodel-specific meaning that does not exist outside of Jodel. Next, we will discuss an explain the given example and our choice of class.

C.2.6.1 Single semantic emoji

For this class, we found only semantically matching top 10 words within our embedding. That is, 🍷 refers to drinking, alcohol and beer, 🏕️ relates to camping, festivals and traveling. 🕵️ relates to suspicion and detecting and related instances, such as Sherlock Holmes. Interestingly, 🍑 does not relate to its sense as a fruit as given by the Unicode definition

within the top 10 words but is associated with various colloquial terms for buttocks. This phenomenon is not specific to the Jodel, but an established synonymous usage in personal digital communication.

C.2.6.2 Multi semantic emoji

Other emoji have multiple semantics, e.g., 🌈 is associated with gay pride and LGBT, whereas it naturally also simply describes rainbows. For 🏰, we observe castles and kingdoms, the city of Heidelberg, and it also relates to Harry Potter. Looking deeper into this particular example, we also find references to the shire from LOTR. The 🛠️ relates to other tools, but may also have a colloquial different interpretation.

C.2.6.3 Omni semantic emoji

Other emoji do not convey specific semantic associations, such as the symbol ✚. The example of 🤪 is associated with embarrassment, awkwardness and weirdness. That is, it may be used in various (also possibly platform-specific) contexts. Quite generic emotion emoji experience heavy usage resulting in high frequencies within our dataset. Finally, 😊 shows that there is no apparent semantic linked to it except for positivity represented by all face-positive emoji.

C.2.6.4 platform-specific emoji semantics

Some emoji develop a special semantic within a platform, which is reflected in our embedding. A good example is 🚌. The top 10 words refer to busses, the German forename “manni”, and different public transportation providers as a meme referring to their service quality (e.g., “racing team” to reflect slow running busses): here, we find the local service corporations KVG (Cologne), ASEAG (Aachen) and HVV (Hamburg) linked to the mentioned meme. The name “manni” is platform-specifically used as a synonym for bus drivers.

Summary and Limitations. By showcasing emoji from a broad set of different groups, we find strong evidence for good semantic associations between emoji and words within the embedding. Although some emoji may inherently not convey particular semantics, most do, which is reflected within the given examples. We also find multiple semantics for a given emoji due to lacking capabilities for any context in such classical word embeddings. Note that the preliminary distinction between the introduced classes is not always straight forward and limited to our interpretation. While this is a first look into the Emoji2Text associations, a broader evaluation incorporating human judgment would be the next step, which we leave for future work.

C.2.7 Text2Emoji Associations

As the last application, we aim at using embeddings to associate emoji to a given word (Text2Emoji). This association likely can be used in several applications such as giving users predictions of emoji while writing or to translate text to emoji. Yet again, these predictions might also help choosing a suitable emoji, given platform particularities. Further, they may help understanding the perceived meaning a text may convey within a specific community.

We decided to use a quantitative analysis for the Text2Emoji associations to give more variance in presenting possible applications. To evaluate the applicability of leveraging word embeddings for the Text2Emoji association, we first define our target task. For keeping this downstream task simple, we define this task as *predicting the first occurring emoji*

within a given text (disregarding others). Other target functions may be approached likewise.

Our evaluation consists of two different approaches: *i*) Naive direct prediction directly using the word embedding, and *ii*) applying statistical machine learning tools as a smart indirection layer on this problem.

Data preparation and selection. As described in Section *Approach: word-emoji Embeddings*, our data has been cleaned and lemmatized before doing any further processing.

To enable a viable evaluation, we first needed to balance our data since the emoji popularity is heavy-tailed. We selected all emoji whose overall share in usage is higher than 0.1‰. This leaves us with 117 different emoji each having a first occurrence in at least 11.5k of distinct sentences. From our dataset, we then randomly selected this number of sentences per emoji that match our problem definition of the selected emoji being the first one occurring in each sentence. These emoji sub-datasets ($\approx 1.3\text{M}$ sentences) are then split into 5 partitions ($\approx 270\text{k}$ sentences) enabling a 5-fold cross validation in our ML approach.

Test setup. For both evaluation approaches, we create a *base* word embedding from all previously non-selected sentences ($\approx 40.9\text{M}$) masking any emoji that might occur, such that this embedding only contains words. For all embeddings, we used the word2vec implementation with 300 dimensions. Then, for each fold, we individually continue training on the base embedding with 4 out of 5 emoji data subsets resulting in 5 different embeddings—each excluding a single subset that is later on used for validation.

Feature selection. To generate features from an input sentence s , we mask all non-first occurring emoji and then calculate an aggregate over all word vectors $v \in V_{s'}$ from the used embedding as proposed in [De Boom et al., 2016]: $f_{s'} = \text{mean}(v(w), w \in s')$.

For our dataset, the mean performs slightly better than the median for the naive approach, whereas min or max achieve worse results with a higher variance; therefore, we decided to use a mean aggregation.

Learning methods. We apply two methodically different approaches to our prediction task. *i*) We implemented a naive method that calculates the top k most similar emoji directly within the word embedding by the cosine distance. In this case, there is no need to train an additional indirection layer, i.e., for each of the 5 embeddings, we can evaluate the non-matching other 4 subsets. *ii*) We further applied a set of commonly available machine learning techniques (e.g., RandomForest, LogisticRegression) with a limited set of hyperparameters. For all of the latter, we created 5 training and validation sets, while randomly shuffling the training set between different folds. We used the resulting probability matrix for calculating the top k predictions. While development, we noticed that training and validation on only a small subset provide quite similar results. Yet for this section, we present only results on the full training and test set.

C.2.7.1 Results

Baseline. A classical ZeroR baseline would choose the class that is predominant within the set, however, due to our balanced data subset, chances are equal. Thus, the probability in a first try is $1/|\text{classes}|$, which is ≈ 0.0085 for our $n = 117$ classes. In case we are having multiple consecutive guesses, we intuitively compute the probability as $p(n, k) = p(n, k-1) + (1 - p(n, k-1)) \cdot (n-k)^{-1}$, $p(n, 0) = n^{-1}$.

| Method | Top1 | Top2 | Top3 | Top4 | Top5 |
|---------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| ZeroR | 0.0085 | 0.0171 | 0.0256 | 0.0342 | 0.0427 |
| Naive | 0.0847 ± 0.0014 | 0.1340 ± 0.00125 | 0.1679 ± 0.0011 | 0.1964 ± 0.0011 | 0.2208 ± 0.0012 |
| MLP | 0.1292 ± 0.0001 | 0.1932 ± 0.0004 | 0.2335 ± 0.0005 | 0.2660 ± 0.0072 | 0.2935 ± 0.0008 |
| LogRegression | 0.1221 ± 0.0008 | 0.1828 ± 0.0005 | 0.2224 ± 0.0006 | 0.2545 ± 0.0004 | 0.2821 ± 0.0005 |
| RandForest | 0.1140 ± 0.0005 | 0.1645 ± 0.0003 | 0.1972 ± 0.0005 | 0.2241 ± 0.0006 | 0.2483 ± 0.0008 |
| AdaBoost | 0.0926 ± 0.0014 | 0.1437 ± 0.0001 | 0.1792 ± 0.0009 | 0.2079 ± 0.0007 | 0.2323 ± 0.0006 |
| GaussianNB | 0.0797 ± 0.0003 | 0.1207 ± 0.0004 | 0.1490 ± 0.0002 | 0.1731 ± 0.0006 | 0.1943 ± 0.0008 |

Table C.5: Text2Emoji prediction results. This table summarizes the first occurring emoji prediction precision results for the top 1 to 5. The ZeroR baseline depicts random choice, our naive approach directly uses the embedding, whereas others apply another indirection layer of machine learning. While the naive approach outperforms random choice, additional machine learning *can* significantly improve results. The best performing algorithm was the MLP being slightly better than LogisticRegression and RandomForest, whereas others compete with our naive approach. Most results are consistent across all top k predictions.

Naive approach. Our first attempt on directly matching a best suited emoji for sentences directly within the embedding by the cosine distance yields an accuracy of about $8.47 \pm .14\%$ for an exact prediction as shown in Table C.5 (Top 1 column, second row). Comparing this result to the baseline (ZeroR, first row), our naive approach performs an order of magnitude better. By loosening the problem allowing a set of ranked predictions, we also show further Top 2..5 results. Here, the delta to the baseline gets smaller for the top k predictions. Presumably due to our large dataset, the standard deviation across the folding sets is quite small. The accuracy of the top k predictions increases almost linearly, such that we can predict the first emoji within the top 5 set with a precision of $22.08 \pm 0.12\%$.

To get a better insight into where this algorithm fails in particular, we also analyzed the resulting confusion matrix shown in Figure C.6a. This Figure shows the emoji group of the actual-true-first emoji of a sentence (y-axis) in relation to the predicted emoji's group (x-axis). We normed the values, such that each row sums up to 1. A perfect match would result in a straight diagonal, whereas a random prediction would yield an equal distribution across the heatmap.

Although many predictions seem reasonable on an emoji group level, mispredictions predominantly towards *Smileys & Emotion*, and *People & Body*, are apparent. On a deeper level—the confusion between subgroups—we observe the same picture, in particular a shift towards *Face-Concerned*, *Face-Smiling* and *Hand-Signs* (not shown). Although this needs to be analyzed deeper in the future, we believe that this is a result of the applicability and usage of these emoji in many different contexts. Further, the grouping defined by the Unicode standard may not be optimal in a semantic sense as shown by [Barbieri et al., 2016], who propose using data-driven clustering of emoji.

Machine Learning. Secondly, we applied off the shelf machine learning techniques to the problem of predicting the first occurring emoji within a text. Our choice of implementation was the Python sklearn package due to its simplicity and popularity. This allows us to plug in our task into various algorithms easily. Our set of choices consists of GaussianNB, AdaBoost, RandForst, LogRegression, and MLP.

We ran a grid search over several hyperparameter combinations of which we only show the best results for each algorithm in Table C.5. This table shows the accuracy of prediction in two different perspectives: *i*) the top 1 column depicts the resulting prediction of each algorithm, whereas *2*) the top 2..5 columns depict the accuracy of each algorithm according

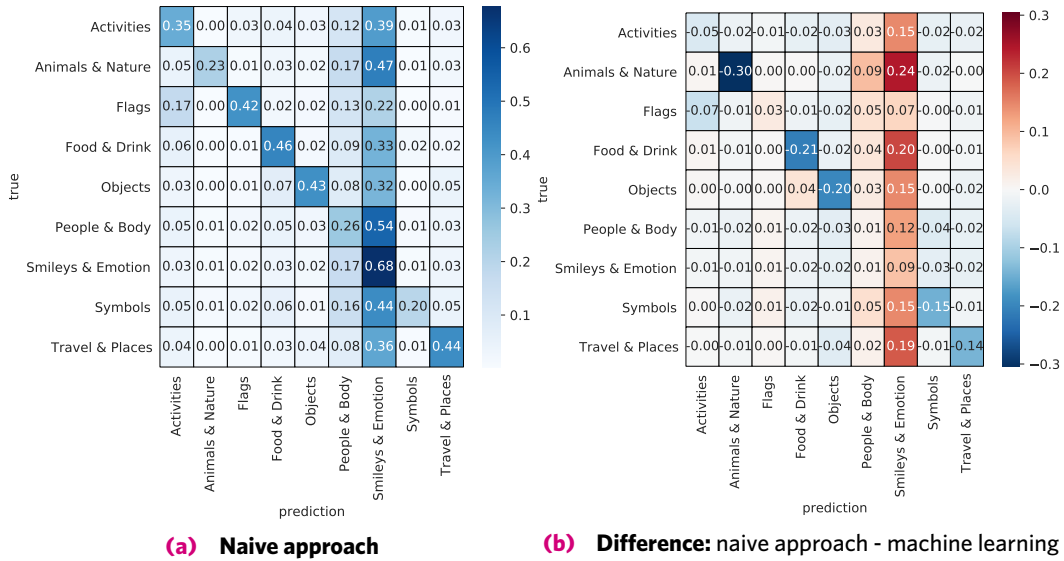


Figure C.6: Text2Emoji top 1 prediction confusion matrices by Unicode groups. The given heatmaps depict the confusion between the Unicode emoji group for the top 1 prediction task. Each actual emoji group (y-axis, true) is mapped to a group of a predicted emoji (x-axis, prediction). Subfigure C.6a shows the prediction confusion for the naive approach (normed row-wise to 1). We observe a diagonal that reveals mostly a good match, however, there is a strong tendency towards predicting emoji from the group *Smileys & Emotion*. Subfigure C.6b depicts the same analogy, but shows the difference from the naive approach to the best performing machine learning variant. Values below 0 imply a higher value in comparison; values greater 0 likewise a lower value. E.g., the predictions towards *Smileys & Emotion* are consistently less frequent. As the ML approach performs better, the difference is almost consistently negative on the diagonal, whereas other values are mostly positive.

to its resulting probability table. That is, whether a correct prediction was amongst the k most probable results. Each value is represented by the mean and standard deviation across the 5 folds.

All classifiers outperformed the ZeroR baseline clearly. The best performing algorithm was Multi-Layer Perceptron classifier with a top 1 (5) accuracy of 12.9% (29.4%), yet the achieved precision is very close to LogisticRegression and RandomForest throughout any top k prediction. Other classifiers performed worse and are comparably good to our naive approach.

As seen in the naive approach, the prediction task is hard for specific groups of emoji (cf. Figure C.6a). We also present these metrics of our best performing machine learning approach in difference to the naive approach in Figure C.6b. E.g., the predictions towards *Smileys & Emotion* are consistently less frequent. In comparison, the better performing machine learning algorithm provides better results for the predominantly mispredicted groups as the difference is almost consistently negative on the diagonal, whereas other values are mostly positive. Still, we draw the same conclusion: there is a shift towards *Smileys & Emotion* and *People & Body* on a group level. In particular, the subgroup confusion matrix reveals that commonly used emotion and hand-sign emoji are the main drivers of mispredictions (not shown).

Summary. Machine learning can be used to predict an emoji for a given word (text to emoji translation) that improves accuracy over the naive approach. While some machine learning techniques perform better than the naive approach, they need an additional considerable amount of computing power for training.

An intense grid search over the hyperparameter space might provide better results. However, since the aim of our study is to demonstrate feasibility, we leave this task open for future work.

C.2.8 Future Work

Besides technical improvements, future work should explore real-world semantic associations in a more principled manner by, e.g., incorporating human subjects as evaluators or annotators. The Jodel app has the unique features of being location-based and anonymous. This enables interesting questions of sorting out semantic differences between locations; the anonymity may also introduce specific semantic aspects different to other Social Media.

We enable such investigations partly by releasing the JEED1488 emoji-subembedding with the publication of this section, and hope to inspire more research into emoji related downstream tasks.

C.2.9 Conclusions

We showed that embeddings are useful to study real-world semantic associations of emoji and words on large-scale messaging data. Our results indicate that word-emoji embeddings reveal insightful Emoji2Emoji and Emoji2Text associations on social media posts going beyond semantic groups defined by the Unicode standard. We show that emoji prediction directly from the embedding may work reasonable well; however, machine learning can improve the results significantly. We posit that such associations are key to understand the usage of each emoji in a given social media platform (e.g., for its users).

While our work demonstrates the potential usefulness of word-emoji embeddings for large scale messaging data, it is exploratory and uses qualitative inspections as a major instrument for our investigations. We used Word2Vec for creating the embedding, other embedding approaches like FastText [Bojanowski et al., 2017b]. Leveraging context, specifically trained sentence-embedding models [Kenter et al., 2016], or, e.g., Bert [Devlin et al., 2019b], may further improve results and possibly catch multiple semantics better. Further, it is still unclear how exactly the amount of training influences semantics that can be extracted from the embedding; we find that more is not always better, depending on the desired application.







C.3 Interpreting Emoji

We study the extent to which emoji can be used to add interpretability to embeddings of text and emoji. To do so, we extend the POLAR-framework that transforms word embeddings to interpretable counterparts and apply it to word-emoji embeddings trained on four years of messaging data from the Jodel social network. We devise crowdsourced human judgement experiment to study six use-cases, evaluating against words only, what role emoji can play in adding interpretability to word embeddings. That is, we use a revised POLAR approach interpreting words and emoji with words, emoji or both according to human judgement. We find statistically significant trends demonstrating that emoji can be used to interpret other emoji very well.

C.3.1 Introduction

Word embeddings create a vector-space representation in which words with a similar meaning are in close proximity. Existing approaches to make embeddings interpretable, e.g., via contextual [Subramanian et al., 2018b] sparse embeddings [Panigrahi et al., 2019], or learned [Senel et al., 2018] transformations [Mathew et al., 2020], focus on text only. Yet, emoji are widely used in casual communication, e.g., Online Social Networks (OSN), and are known to extend textual expressiveness, demonstrated to benefit e.g., sentiment analysis [Novak et al., 2015, Hu et al., 2017b].

Structure [JH10]

-  [C.3.2: Related Work](#)
-  [C.3.3: Creating Interpretable Embeddings](#)
-  [C.3.4: Embedding and Polarization](#)
-  [C.3.5: Human Evaluation](#)
-  [C.3.5.2: Results](#)
-  [C.3.6: Conclusions](#)

C.3.1.1 Research Questions

We study what role emoji's expressiveness beyond words can play in adding interpretability to word-emoji embeddings. I.e., can we adopt word embedding interpretability via leveraging semantic polar opposites (e.g., cold / hot) to emoji (e.g., ❄️ / ☀️, or 😡 / 😊) for interpreting words or emoji w.r.t. human judgement.

C.3.1.2 Approach

Motivated and based upon POLAR [Mathew et al., 2020], we deploy a revised variant POLAR^ρ that transforms arbitrary word embeddings into interpretable counterparts. The key idea is to leverage semantic differentials as a psychometric tool to align embedded *terms* on a scale between two polar opposites. Employing a projection-based transformation in POLAR^ρ, we provide embedding dimensions with semantic information. I.e., the resulting interpretable embedding space values directly estimate a *term*'s position on a-priori provided polar opposite scales, while approximately preserving in-embedding structures (Section C.3.3).

For studying the role of emoji in interpretability, we create a word-emoji input embedding from on a large social media corpus. The dataset comprises four years of complete data

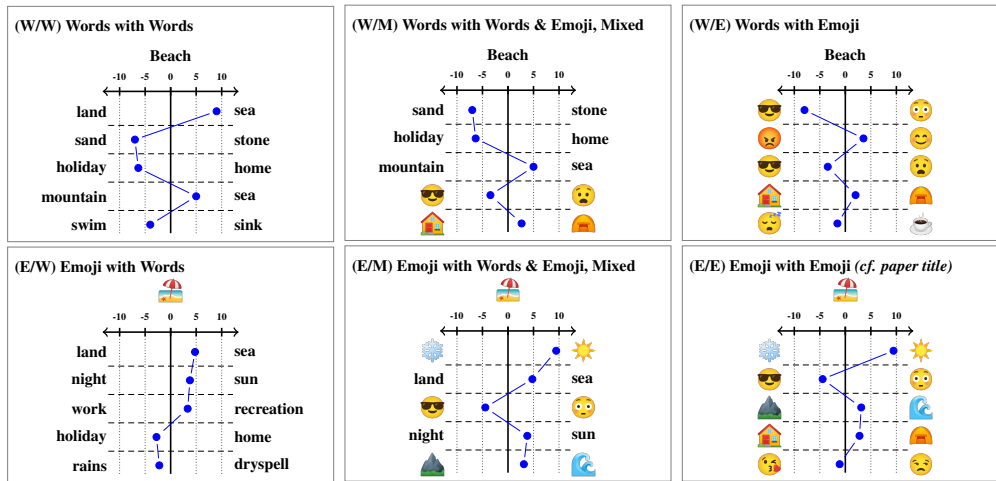


Figure C.7: The POLAR-framework [Mathew et al., 2020] makes word embeddings interpretable leveraging polar opposites. It provides a new interpretable embedding subspace with systematic polar opposite scales: Along six use-cases, we evaluate which role emoji expressiveness plays in adding interpretability to word embeddings. I.e., how well can our adopted POLAR^ρ interpret (W/*) words or (E/*) emoji with words, emoji or both (*M), *Mixed*. **We test POLAR^ρ alignment with human judgement as represented in shown semantic profiles above.**

in a single country from the online social network provider Jodel (48M posts of which 11M contain emoji). For subsequent main evaluation, we make this embedding interpretable with word and emoji opposites by deploying our adopted tool POLAR^ρ (Section C.3.4).

Given different expressiveness of emoji, we ask **RQ1)** How does adding emoji to POLAR^ρ affect interpretability w.r.t. to human judgement? I.e., do humans agree on best interpretable dimensions for describing words or emoji with word or emoji opposites? And **RQ2)** How well do POLAR^ρ-semantic dimensions reflect a *term*'s position on a scale between word or emoji polar opposites?

We design a crowdsourced human judgement experiment (Section C.3.5) to study if adding emoji to word embeddings and POLAR^ρ in particular increases the interpretability—while also answering how to describe emoji best. Our human judgement experiment involves six campaigns explaining *Words* (W/*) or *Emoji* (E/*) with *Words*, *Emoji*, or both *Mixed*. We evaluate two test conditions to answer both research questions: (*RQ1*) a *selection* test studies if human subjects agree to the POLAR^ρ identified differentials (e.g., how do emoji affect POLAR^ρ interpretability?), and (*RQ2*) a *preference* test that studies if the direction on a given differential scale is in line with human judgement (e.g., how well does POLAR^ρ interpret scales).

① The projection approach POLAR^ρ is the contribution of one of my co-authors Sandipan Sikdar.

C.3.1.3 Results

POLAR^ρ identifies the best interpretable opposites for describing emoji with emoji, yet generally aligning well with human judgement. Except interpreting words with emoji only probably due to lack of emoji expressiveness indicated by coder agreement. Further, POLAR^ρ estimates an embedded *terms*' position on a scale between opposites successfully, especially for interpreting emoji.

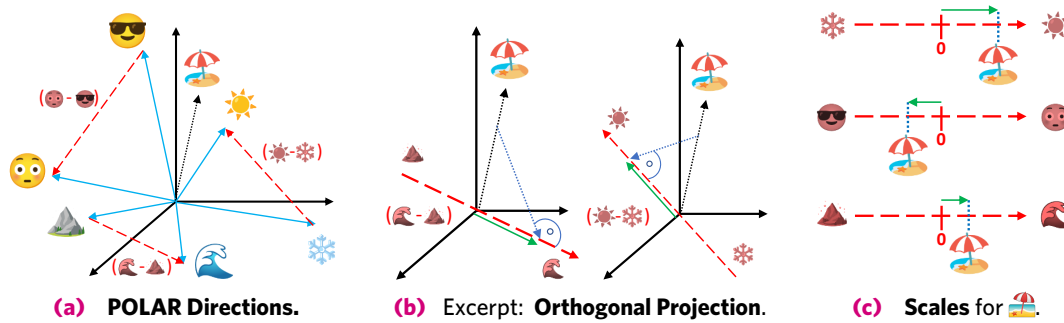


Figure C.8: POLAR [Mathew et al., 2020] with Projection in a nutshell. We showcase POLAR^p interpreting emoji with emoji (E/E). (a) We leverage polar opposites (here: ☀️/❄️, 😊/😞, 🏔️/🌊) to provide embedding dimensions with semantic information. By using opposite differential directions (red dashed vectors), we create a new interpretable subspace. (b) Orthogonal projection (blue dotted vectors) of an embedded term (here: 🌂) onto this subspace (e.g., left: 🏔️/🌊, right: ☀️/❄️) yields a direct scale measure between both opposites in the adjacent leg (green vectors, directed alike the differential). (c) The resulting interpretable embedding now contains a tangible position estimation along employed polar dimensions for each embedded term (here: 🌂).

C.3.2 Related Work

No universal meaning of emoji Prior work showed that the interpretation of emoji varies [Miller et al., 2016a, Kimura-Thollander and Kumar, 2019], also between cultures [Guntuku et al., 2019, Gupta et al., 2021]. Even within the same culture, ambiguity and double meanings of emoji exist [JH7] and differences exist on the basis of an individual usage [Wiseman and Gould, 2018]. These observations motivate the need to better understand the meaning of emoji. Currently, no data-driven approach exists to make emoji interpretable—a gap that we aim to close.

Interpretable word embeddings Word embeddings are a common approach to capture meaning; they are a learned vector space representation of text that carries semantic relationships as distances between the embedded words. A rich body of work aims at making word embeddings interpretable, e.g., via contextual [Subramanian et al., 2018b], sparse embeddings [Panigrahi et al., 2019], or learned [Senel et al., 2018] transformations [Mathew et al., 2020]—all focus on text only. Recently, [Mathew et al., 2020] proposed the POLAR that takes a word embedding as input and creates a new interpretable embedding on a polar subspace. The POLAR approach is similar to SEMCAT [Senel et al., 2018], but is based on the concept of semantic differentials [Osgood et al., 1957] for creating a polar subspace. It measures the meaning of abstract concepts by relying on opposing dimensions associated (good vs. bad, hot vs. cold, conservative vs. liberal). In this work, we extend and use POLAR. Note that the framework has recently been extended to contextual embeddings [Engler et al., 2023].

Emoji embeddings Few works focused on using word embeddings for creating emoji representations, e.g., [Eisner et al., 2016b] or [JH7]. [Barbieri et al., 2016] used a vector space skip-gram model to infer the meaning of emoji in Twitter data [Barbieri et al., 2016]. Yet, the general question if the interpretability of word embeddings can be improved by adding emoji and if different meaning of emoji can be captured remains still open. In this work, we adapt the POLAR interpretability approach to emoji and study in a human subject experiment if word embeddings can be made interpretable by adding emoji and how emoji can be interpreted by emoji.

C.3.3 Creating Interpretable Embeddings

We explain next our deployed tool for creating interpretable word-emoji embeddings: POLAR [Mathew et al., 2020]; and provide detail on a revised POLAR extension via projection.

C.3.3.1 POLAR Approach

Semantic Differentials. Based upon the idea of semantic differentials, denoted as z , as a psychometric tool to align a word on a scale between two polar opposites (Fig. C.7), POLAR [Mathew et al., 2020] takes a word embedding as input and creates a new interpretable embedding on a polar subspace. This subspace, i.e., the opposites ($z, -z$) used for the interpretable embedding are defined by an external source.

That is, starting with a corpus and its vocabulary \mathcal{V} , a word embedding created by an algorithm a (e.g., Word2Vec or GloVe (cf. C.3)) assigns vectors $\overrightarrow{\mathbb{W}}_v^a \in \mathbb{R}^d$ on d dimensions to all words $v \in \mathcal{V}$ according to an optimization function (usually word co-occurrence). This pretraining results in an embedding:

$$\mathbb{D} = [\overrightarrow{\mathbb{W}}_v^a, v \in \mathcal{V}] \in \mathbb{R}^{|\mathcal{V}| \times d}$$

Instances from this embedding space carry a semantic structure between embedded words according to applied pretraining. We can leverage the semantic structure between words to transform the embedding space to carrying over meaning into the dimensions: POLAR uses N semantic differentials/opposites that are itself items within the embedding, i.e., $\mathbb{P} = \{(p_z^i, p_{-z}^i), i \in [1..N], (p_z^i, p_{-z}^i) \subseteq \mathcal{V}^2\}$.

As shown in Fig. C.8a, given two anchor points for each polar opposite, a line between them represents a differential—which we name POLAR direction (red dashed vectors):

$$\overrightarrow{dir}_i = \overrightarrow{\mathbb{W}}_{p_z^i}^a - \overrightarrow{\mathbb{W}}_{p_{-z}^i}^a \in \mathbb{R}^d$$

Original POLAR. Naturally, we can use these differentials to create a new basis for the interpretable embedding \mathbb{E} , where each dimension represents a scale measure of the semantic z .

Summarizing [Mathew et al., 2020], they first gather all directions in a matrix $dir \in \mathbb{R}^{N \times d}$, where \overrightarrow{dir}_i denotes the coordinates of each differential, obtaining for all embedded terms $v \in \mathcal{V}$: $dir \cdot \overrightarrow{\mathbb{E}}_v = \overrightarrow{\mathbb{W}}_v^a$. In most practical cases, this set of equations is over- or under-determined, which is why they use the Moore-Penrose pseudoinverse to calculate a least squares solution for the interpretable embedding vectors as follows: $\overrightarrow{\mathbb{E}}_v = (dir^T)^+ \overrightarrow{\mathbb{W}}_v^a$, yielding an interpretable subspace along the differentials dir that carries over specific geometric semantics from the input embedding. I.e., for each word $v \in \mathcal{V}$ within the resulting interpretable embedding \mathbb{E} , its embedding vector $\overrightarrow{\mathbb{E}}_v$ now carries a heuristic measure w.r.t. each polar dimension's semantics.

Limitations. Polar opposites being very close in the original embedding space might tear apart. From a technical perspective, due to the used pseudo inverse⁴ suitable numerical stability remains to be shown and hence there remain concerns, especially if $d \approx N$ [Mathew et al., 2020].

C.3.3.2 POLAR^ρ Extension: Projection

While the base change approach seems natural, its given limitations lead us to propose a variant that comes with several benefits. Instead of creating a new interpretable vector space, we take measurements on the differentials dir defined as before (Fig. C.8a, red

⁴we used a python numpy SVD-based implementation

dashed vectors). However, we now project each embedding vector $\overrightarrow{\mathbb{W}}_v$ for v orthogonally onto the differentials as shown in Fig. C.8b (blue dotted vectors). This leads to a smallest distance between both lines w.r.t. the differential, yet simultaneously allows for a direct scale measure on the differential vector as shown in Fig. C.8b and Fig. C.8c (green vectors). Thereby, we also *decouple* the transformation matrix, which eases later add-ins to the interpretable embedding.

Orthogonal projection (blue dotted vectors) of each input embedding vector $\overrightarrow{\mathbb{W}}_v^a$ onto a differential i provides us the adjacent leg vector as follows:

$$\text{oproj}_{dir_i}(\overrightarrow{\mathbb{W}}_v^a) = \underbrace{\frac{\overrightarrow{\mathbb{W}}_v^a \cdot \overrightarrow{dir}_i}{|\overrightarrow{dir}_i|}}_{\text{scalar}} \cdot \underbrace{\frac{\overrightarrow{dir}_i}{|\overrightarrow{dir}_i|}}_{\text{direction}}$$

As this adjacent leg (green vectors)'s direction naturally equals the differential, we focus only on the scalar part representing a direct scale measure. By normalizing the differential vector lengths $\hat{dir} = \left\{ \overrightarrow{\hat{dir}}_i = \overrightarrow{dir}_i \cdot |\overrightarrow{dir}_i|^{-1}, i \in [1..N] \right\} \in \mathbb{R}^{N \times d}$, the projected scale value conveniently results in the following transformation per differential and embedded term: $|\text{oproj}_{dir_i}(\overrightarrow{\mathbb{W}}_v^a)| = \overrightarrow{\hat{dir}}_i \cdot \overrightarrow{\mathbb{W}}_v^a$.

This transformation allows to create a new interpretable embedding for each term's embedding vector $\overrightarrow{\mathbb{W}}_v^a$ (exemplified in Fig. C.7) by generalizing the projection into a single matrix operation as follows:

$$\overrightarrow{\mathbb{E}}_v = |\text{oproj}_{dir} \cdot (\overrightarrow{\mathbb{W}}_v^a)| = \hat{dir} \cdot \overrightarrow{\mathbb{W}}_v^a \in \mathbb{R}^N$$

Computationally, this requires an initial matrix multiplication for each embedded *term*; Dimension increments require a dot product on each *term*.

Downstream Tasks. Other experiments indicate POLAR ^{ρ} downstream task performance being on par with the input embedding, and an edge over base change POLAR in numerical stability.

C.3.3.3 Measuring Dimension Importance

There can be many possible POLAR dimensions, which requires to select the most suitable ones. That is, we want to define a limited set of opposites that best describes words or emoji w.r.t. interpretability across the whole embedding.

Extremal Word Score (EWSO). We propose a new metric to measure the quality of polar dimensions complementing heuristics from [Mathew et al., 2020]. It measures the embedding confidence and consistency along available differentials. The idea of POLAR ^{ρ} is that directions represent semantics within the input embedding. We determine embedded *terms*' shortest distance to these axes via orthogonal projection; we use resulting intersections as the position w.r.t. the directions.

That is, as a new heuristic, for each of our differentials dir_i , we look out for $k = 10$ embedded words at the extremes (having the highest scores in each direction) and take their average cosine distance within the original embedding \mathbb{D} to the differential as a measure. This results in the average similarity of existing *extremal* words on our scale—a heuristic that represents the skew-whiffiness within the extremes on a differential scale.

C.3.4 Embedding and Polarization

We next propose an approach to improve the interpretability of word embeddings by adding emoji. It uses our extended version POLAR^ρ and adds emoji to the POLAR space by creating word embeddings that include emoji.

| Description | # |
|-----------------------------|------------|
| Sentences | 48,804,375 |
| ~ after cleaning | 42,245,200 |
| Total Emoji | 19,911,632 |
| Sentences w/ Emoji | 11,283,180 |
| Emoji per Sentence | 1.76 |
| Unique Emoji w/ modifiers | 2,620 |
| Unique Emoji w/o skin color | 1,488 |

Table C.6: Corpus Statistics. The data set contains all posts in Germany since late 2014 until August 2017.

Corpus. The network operators provided us with data (cf. Tab. C.6). of content created in Germany from 2014 to 2017. It contains 48M sentences, of which 11M contain emoji (1.76 emoji per sentence on average).

Ethics. The dataset contains no personal information and cannot be used to personally identify users except for data that they willingly have posted on the platform. We synchronize with the Jodel operator on analyses we perform on their data.

C.3.4.1 Semantic Differential Sources

POLAR^ρ can create interpretable embeddings w.r.t. a-priori provided opposites. We next describe how we select these opposites to make POLAR^ρ applicable to our data. Most importantly, the approach requires being part of or locating desired opposites within the original embedding space.

Words. As we extend the word embedding space with emoji, we still want to use words. We find common sources of polar opposites in antonym wordlists [Shwartz et al., 2017] as used in the original POLAR work. To fit our German dataset, we translated and manually checked all pairs keeping 1275 items. From GermaNet [Hamp and Feldweg, 1997], we extracted 1732 word pairs via antonym relations leading to $|\mathbb{P}_{\text{words}}| = 1832$ word pairs.

Emoji. Being not ideal, but due to lack of better alternatives, we ended up heuristically creating semantic opposites from emoji through qualitative surveys across friends and colleagues resulting in $|\mathbb{P}_{\text{emoji}}| = 44$ emoji pairs, cf. Tab. C.10. While we could use far more opposites especially of facial emoji, due to emoji clustering in the input embedding, spanned expressive space would arguably become redundant at similar EWSO scores for many directions. Effectively it may bias interpretability over proportionally towards facial emoji.

C.3.4.2 Polarization

Preprocessing. We tokenize sentences with spaCy and remove stop words. To increase amounts of available data, we remove all emoji modifiers (skin tone and gender): $\{\text{👤}, \text{👤}, \text{👤}\} \rightarrow \text{👤}$. Due to German language, we keep capitalization.

| interpret | with | | |
|-----------|-------|-------|-------|
| | W | Mixed | Emoji |
| Words | (W/W) | (W/M) | (W/E) |
| Emoji | (E/W) | (E/M) | (E/E) |

Figure C.9: Campaigns Overview. We interpret Words and Emoji with likewise Words, Emoji, and Mixed (both).

Please choose 5 Pairs that characterize 🌈 best!

black - white

female - male

slow - fast

fork - spoon

⋮

Which term describes 🌈 better?

← = →

black ○○○○○○○○○ white

female ○○○○○○○○○ male

slow ○○○○○○○○○ fast

fork ○○○○○○○○○ spoon

⋮

(a) Selection Task for Emoji/Mixed (E/M).

(b) Preference Task for Emoji/Mixed (E/M).

Figure C.10: Crowdsourced Questions. (a) We conduct six campaigns measuring human interpretability for including emoji to the POLAR^p embedding space. Exemplified with the Emoji Mixed campaign (E/M): interpreting emoji with emoji and words. (b) In the Selection test, coders choose suitable differentials for describing a given term. (c) In the Preference test, coders provide their interpretation of a given term to a differential scale.

Original Embedding. We use `gensim` implementation of Word2Vec (W2V). A qualitative investigation suggests that skip-gram works better than CBOW (better word analogy). We kept training parameters largely at defaults including negative sampling, opting for $d=300$ dimensions.

Interpretable Embedding. We create the matrix of differentials dir , the POLAR subspace, according to our antonym-set $\mathbb{P}_{\text{words}} \cup \mathbb{P}_{\text{emoji}}$ (Section C.3.4.1). After normalizing the subspace vectors, we create all embedding vectors via projection $\vec{\mathbb{E}}_v = \hat{dir}^T \vec{\mathbb{W}}_v, \forall v \in \mathcal{V}$. Though normalization requires careful later additions to the POLAR space, we opted for standard normalization, $\mathbb{E}_{\text{stdnorm}} = [\mathbb{E} - \text{mean}(\mathbb{E})] \cdot \text{std}(\mathbb{E})^{-1}$, to ensure that the whole embedding space aligns properly around the center of gravity on each differential scale. We select the best suited opposites for a given embedding space by using the Extremal Word Score (Section C.3.3.3) for $d=500+44$ dimensions (words + emoji).

C.3.5 Human Evaluation

While we have now created a supposedly interpretable embedding, it remains to be seen how well it is *perceived* by humans. That is, we next evaluate our two key RQs, discuss significance, and provide further details: *RQ1*) How well does POLAR^p with EWSO perform in selecting most interpretable dimensions at varying expressiveness of words and emoji? *RQ2*) How well do POLAR^p scalar values reflect directions on the differential scales? *i*) Do humans prefer emoji to words? *ii*) How well do human raters align w.r.t. interpretability? *iii*) What impact do demographic factors play in interpretability with or without emoji?

C.3.5.1 Evaluation design

To gather human judgement, we employ crowdsourcing on the Microworkers platform.

Questions and Evaluation Metrics Our evaluation of the POLAR^p approach including emoji to the differentials bases on two main questions next to demographics.

| Task | Random | POLAR | (W/W) | (W/M) | (W/E) | (E/W) | (E/M) | (E/E) | |
|------------|--------|-------|--------------|-------------|-------------|-------|--------------|-------|-------------|
| Selection | Top 1 | 0.500 | 0.876 | 0.79 | 0.83 | 0.60 | 0.81 | 0.79 | 0.88 |
| | Top 2 | 0.222 | 0.667 | 0.62 | 0.61 | 0.35 | 0.67 | 0.68 | 0.77 |
| | Top 3 | 0.083 | 0.420 | 0.45 | 0.42 | 0.15 | 0.54 | 0.57 | 0.67 |
| | Top 4 | 0.024 | 0.222 | 0.30 | 0.18 | 0.07 | 0.37 | 0.37 | 0.59 |
| | Top 5 | 0.004 | 0.086 | 0.14 | 0.08 | 0.01 | 0.22 | 0.19 | 0.38 |
| Preference | 0.500 | - | 0.740 | 0.672 | 0.576 | 0.800 | 0.848 | 0.832 | |

Table C.7: Crowdsourcing Campaign results. Random and original POLAR baseline. Selection and Preference results across campaigns. Words are better described by word dimensions, and emoji are better described by emoji dimensions.

SELECTION TEST. Analogous to the original work, we want to find out whether humans agree on best interpretability of POLAR^ρ selected differentials with a word intrusion task. The question asks our coders to select five out of ten differentials that describe a given word best as shown in Fig. C.10a. We select half of these dimensions according to the highest absolute projection scale values (most extreme). The other half consists of a random selection from the bottom half of available differentials. I.e., if the projection approach determines interpretable dimensions well, humans would choose all five out of five POLAR^ρ chosen differentials.

As any user might choose differently, we count how often coders choose certain differentials. The resulting frequencies immediately translate in a ranking that we leverage for calculating the fraction of Top 1..5 being POLAR^ρ chosen differentials.

PREFERENCE TEST. Additionally, we introduce the preference test evaluating whether the direction on a given differential scale is in line with human judgement. That is, for the same words from the selection test, we display the same ten dimensions (5 top-POLAR^ρ, 5 random bottom) where coders select their interpretation of the given word on scales as shown in Fig. C.10b. Typical for semantic differential scales [Tullis and Albert, 2008, Osgood et al., 1957], we deliberately use a seven point scale representing -3 to 3, allowing more freedom than 3 or 5 points [Simms et al., 2019]. Further, we specifically allow a center point—being equal—as it might indicate both being *equally well* or *not good at all*.

Due to scale usage heterogeneity [Rossi et al., 2001], we normalize coder chosen directions (shift+scale according to mean) prohibiting disproportional influence of single coders. We evaluate the coder agreement by counting direction (sign) non-/alignment with the POLAR^ρ projection scale.

DEMOGRAPHICS. There is a multitude of other external factors that might have impact on coders’ choices. To better understand participant background, we ask for their education, emoji usage (familiarity), smartphone platform (different emoji pictograms), and if they had used Jodel before.

Evaluation Setup

CROWDWORKER CAMPAIGNS We run a campaign for each of the cross product between words only, emoji only, and mixed Tab. C.9 and Fig. C.8. (W/W) word/word sets a baseline comparison to results from the original POLAR work, albeit now using the projection approach. (W/M): word/mixed uses not only words, but includes emoji to the POLAR subspace. (W/E): word/emoji uses only emoji to describe words. (E/W): emoji/word provides another baseline as to how well emoji may be interpreted with words only. (E/M): emoji/mixed uses both, emoji and words to interpret emoji. (E/E): emoji/emoji may be the most interesting as we only use the expressiveness of emoji to describe emoji.

For mixed cases (emoji and words within the POLAR subspace), we create rankings from absolute scale values on both types (words/emoji) separately and then select them equally often to achieve similar amounts of word and emoji differentials.

USED WORDS AND EMOJI. We selected 50 words and emoji to be described in each campaign. To ensure that *i*) we only use common words that are very likely known to our coders, and *ii*) these words are captured well within the underlying embedding, we pick them out of the upper 25% quantile by occurrences in the corpus ($n \geq 1.6k$). I.e., we chose emoji and words that appear frequently and should therefore be well-known. For words, we ensured that they are part of the German dictionary *Duden*.

TASKS SETUP. Within our six campaigns, we now have each 50 emoji or 50 words to be interpreted. We bundled this into 5 tasks each consisting of 10 emoji/words—resulting in 30 different tasks. Each of these tasks contains the Selection test, Preference test, and demographics.

SUBJECTS. Human judgement and crowdsourced evaluations are noisy by nature. While it is usually sufficient to employ few trusted expert coders, it is suggested to use more in the non-expert case [Snow et al., 2008]. Thus, we assign 5 different annotators to each of the 30 tasks. At estimated 10-15min duration, we provide 3\$ compensation for answering a single task, above minimum wage in our country.

QUALITY ASSURANCE. Any crowdsourcing task offers an incentive to rush tasks for the money, which requires us to employ means of quality assurance (QA). As we have an uncontrolled environment and thus untrusted coders, we handcraft test questions for the selection and preference test. This task is non-trivial as we require unambiguity in *correct* answers (we ensured this with multiple qualitative tests among friends and colleagues), while simultaneously not being too obvious. We place one test question for selection and one for preference randomly into each task (ending up in 11 words or emoji per task). This also means that each coder can only participate in up to 5 different tasks within a single campaign before re-seeing a test question.

We define acceptance thresholds of four out of five correct answers for both, the selection test and the correct direction for the preference test.

C.3.5.2 Results

Within the crowdsourcing process, we rejected about 10% of all tasks according to our QA measures, which then had to be re-taken. We ended up with 6 campaigns each having 50 words/emoji answered by 5 coders; summing up to completed 150 tasks. In total, 16 different coders accomplished this series of which 4 completed $\Sigma \geq 100$ tasks.

Interpreting Emoji First we focus on the describing emoji campaigns (E/*). We present our main evaluation results in Tab. C.7. Within columns, we show results for random, original POLAR, and our six campaigns. We split the rows into results from the selection test across Top1..5 entries and the preference test.

SELECTION TEST. We find very good results along all emoji campaigns (E/*) being consistently better than any campaign describing words (W/*). The best performance was achieved for explaining emoji with emoji (E/E); others are on par.

We want to note however, that the small size of used emoji-differential set may ease selection. E.g., facial expression emoji regularly achieve higher embedding scores than

| Task | | (W/M) | (W/E) | (E/W) | (E/M) | (E/E) |
|------------|-------|-------|--------------|-------|--------------|--------------|
| Selection | Top 1 | 0.605 | 0.037 | 0.812 | 0.941 | 0.229 |
| | Top 2 | 0.945 | 0.008 | 0.593 | 0.527 | 0.094 |
| | Top 3 | 0.800 | 0.001 | 0.345 | 0.216 | 0.025 |
| | Top 4 | 0.157 | 0.004 | 0.464 | 0.452 | 0.003 |
| | Top 5 | 0.391 | 0.023 | 0.291 | 0.485 | 0.005 |
| Preference | | 0.095 | 0.000 | 0.111 | 0.003 | 0.012 |

Table C.8: Chi-square χ^2 statistics. $p = 0.05$ (bold) across campaigns compared to words-words (W/W) as a baseline.

others, which thus may bias the bottom control half (Section C.3.5.1). However, interpreting emoji or words with words only, (E/W) and (W/W), achieve comparable performance.

PREFERENCE TEST. Here, we make the same observation; The projected scales on the differentials are mostly well in line with human judgement.

Interpreting Words Again, we refer to Tab. C.7, but now change our focus to describing words, campaigns (W/*).

SELECTION TEST. Albeit not being directly comparable, using POLAR^{*p*} in compaigns: describing words with words (W/W), or describing words with words and emoji (W/M) achieved performance well on par with POLAR. Noteworthy, describing words with emoji (W/E) yielded the worst results. The projection scale values for the emoji dimensions were mostly lower compared to words. I.e., according to POLAR^{*p*}, for words only few emoji differentials would be among the top 5 opposites.

PREFERENCE TEST. As for the preference test, describing words yield the best results using word opposites only (W/W). Explaining words with emoji (W/E) performs particularly worse.

Result Significance To test for differences within the coder alignment with POLAR^{*p*}, we model both, the selection and preference test. With our primary goal to understand the impact of including emoji to a POLAR^{*p*} interpretable word embedding, we anchor to the (W/W) campaign as a baseline.

For the selection test, we count if coders aligned with POLAR^{*p*} or chose any of the random alternatives across the Top 1..5 selection. For the preference test, we count whether coders aligned with POLAR^{*p*}'s scale direction. We apply double-sided chi-square tests χ^2 with $p < 0.05$ between the interpreting words with words (W/W) baseline and the remaining five campaigns as shown in Table C.8.

We identify significant differences in coder-POLAR^{*p*} alignment to the (W/W) baseline when describing words with emoji (W/E) over Top1..5 selection and preference. Counts from explaining emoji with emoji (E/E) signal significance for preference and selection Top3..5. Coder-POLAR^{*p*} alignment in preferences is also significant for describing emoji with emoji and words (E/M).

Observations

EMOJI. As a byproduct, we also show if emoji opposites are preferred over words. That is, we focus on the mixed campaigns describing words and emoji with words and emoji (*M).

| α | (W/W) | (W/M) | (W/E) | (E/W) | (E/M) | (E/E) |
|--|-------|-------|-------|-------------|-------|-------------|
| Selection | 0.44 | 0.35 | 0.24 | 0.46 | 0.39 | 0.55 |
| Preference | 0.57 | 0.41 | 0.34 | 0.61 | 0.54 | 0.60 |
| Preference <i>POLAR^p only</i> | 0.65 | 0.52 | 0.40 | 0.70 | 0.64 | 0.68 |
| Preference <i>random only</i> | 0.31 | 0.17 | 0.25 | 0.31 | 0.22 | 0.22 |

Table C.9: Inter-rater agreement. Krippendorff’s α across campaigns. Coders achieve the best agreement in selection test of emoji-based campaigns (E/*) and generally within the preference test measuring differential scales.

We establish a baseline by filtering the counts for all non-POLAR^p randomly chosen dimensions being word or emoji representing a Bernoulli experiment. I.e., along the random dimensions, our coders chose 228 vs. 221 and 167 vs. 187 words over emoji. Applying chi-squared statistics indicates, that both types (words and emoji) are chosen equally often at least cannot be rejected.

We next analyze the POLAR^p chosen dimensions in the mixed campaigns. Here, coders chose words over emoji as follows: 465 vs. 336 in the (W/M), and 414 vs. 482 in the (E/M) campaign. We find statistically significant favors for words to interpret words and emoji to describe emoji.

SCALE USAGE. We find no evidence for any directional biases within our preference test (cf. C.10b).

CODER AGREEMENT. While the aggregate results are compelling, we use the Krippendorff-alpha metric to measure coder agreement along all six campaigns as shown Tab. C.9; higher scores depict better agreement. We split the overall results by test first (Selection and Preference), but also show additional agreement results for preferences along POLAR^p chosen dimensions and their random counterpart.

Most agreement is within the *moderate* regime. This observation does not come unexpected from our five non-expert classifiers per task. Overall, we find that coders agree better for well-performing campaigns. We identify the best agreement scores for interpreting emoji with emoji (E/E); coders agree least in the worst performing explaining words with emoji campaign (W/E).

For the preference test, we subdivide our results into POLAR^p chosen differentials and compare them to the randomly chosen ones. While the agreement on the random opposites is only *fair*, the agreement on POLAR^p chosen opposites is consistently better: Estimating differential scale directions via POLAR^p for words yields *moderate* agreement, whereas coders consistently align *substantially* in interpreting emoji. We presume emoji may convey limited ideas, but are easier to grasp, have better readability; the campaigns interpreting emoji (E/*) were generally accomplished faster.

Demographics Though we are confident in applied QA measures, none of the demographics can be confirmed. The annotator sample-size is small and thus most likely not representative. Further, we find most workers providing contrasting answers across multiple tasks they participated in, rendering collected demographic information unusable.

C.3.6 Conclusions

We raise the question whether we can leverage the expressiveness of emoji to make word embeddings interpretable. Thus, we use the POLAR framework [Mathew et al., 2020] that creates interpretable word embeddings through semantic differentials, polar opposites. We

employ a revised POLAR^ρ method that transforms arbitrary word embeddings to interpretable counterparts to which we added emoji. We base our evaluation on an off the shelf word-emoji embedding from a large social media corpus, resulting in an interpretable embedding based on semantic differentials, i.e., antonym lists and polar emoji opposites.

Via crowdsourced campaigns, we investigate the interpretable word-emoji embedding quality along six use-cases (cf. Fig. C.7): Using word- and emoji-polar opposites (or both *Mixed*), to interpret words (W/W, W/E, W/M) and emoji (E/W, E/E, E/M), w.r.t. human interpretability. Overall, we find POLAR^ρ's interpretations w/wo emoji being well in line with human judgement. We show that explaining emoji with emoji (E/E) works statistically significantly best, whereas describing words with emoji (W/E) systematically yields the worst performance. We also find good alignment to human judgement estimating a *term*'s position on differential scales, using the POLAR^ρ-projection.

That is, emoji can improve POLAR^ρ's capability in identifying most interpretable semantic differentials. We have demonstrated how emoji can be used to interpret other emoji using POLAR^ρ.

Acknowledgements

We thank Felix Dommes, who was instrumental for this work by helping develop and implement the POLAR^ρ projection approach and the Extremal Word Score in his Master Thesis.

| p_{-z} | p_z | p_{-z} | p_z | p_{-z} | p_z | p_{-z} | p_z |
|----------|-------|----------|-------|----------|-------|----------|-------|
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |

Table C.10: Used heuristically identified polar emoji opposites, $(p_{-z}, p_z) \in \mathbb{P}_{\text{emoji}}$. We opted for a diverse set of opposites selecting only few facial emoji differentials.

Chapter Summary

With our established understanding of adoption processes and insights to user interactions within DE & SA, we have detailed various aspects of abstractions of ideas, to specific contents, and diverse methods of communication. That is, we explored ★ **A) Information Spreading** represented by hashtags across the Jodel platform in Germany. For easing comparison, we employ metrics as used on other platforms [Kamath et al., 2013, Brodersen et al., 2012]. Next, we take a deep dive into actual ★ **B) Message Contents** via an in-depth human crowdsourced content evaluation. Inspired by [Kang et al., 2016] and [Paul et al., 2011, Correa et al., 2015], we developed a rich classification scheme that distinguishes between message *intents* (why) and *topics* (what). Finally, we highlight ★ **C) The Role of Emoji** in today's social media. We provide empirical community insights to ★ **C.1) Social Media Emoji Usage**. By leveraging neural word embeddings, we are ★ **C.2) Making Sense out of Emoji**. With shown success of our emoji embeddings in representing semantic associations, we then apply and measure the POLAR [Mathew et al., 2020] framework that adds ★ **C.3) Interpreting Emoji** to an input embedding space. We provide evidence that the algorithms decisions are well in line with human judgement w.r.t. interpretability.



USER MANAGEMENT

In a nutshell. We have learned countless details of platform adoption, user behavior and communication contents. However, every platform provides and implements only certain functionalities and frameworks steering its users and platforms. Within this chapter, we take a deep dive into the very heat of user moderation on the Jodel platform through **distributed moderation** not only in *voting*, but also the additional *moderation* safeguard layer. We next look into distribution shifts within user bases empirically characterizing **user churn**. By modeling churn features from metadata, we enable **Customer Lifetime Value** and imminent **churn prediction** at high performance leveraging Random Forest Ensembles. Lastly, we provide a methodology outlook as to how to measure **long-term** platform dynamics and **Quality of User Experience** with *data-driven* tools. Identified signals and hypotheses then allow of traditional in-depth controlled lab studies being infeasible for prevalent timeframes.

CONTENT

| | | |
|----------|---|------------|
| A | Distributed Content Moderation | 183 |
| A.1 | Introduction | 183 |
| A.1.1 | Research Question | 184 |
| A.1.2 | Approach | 184 |
| A.1.3 | Results | 185 |
| A.2 | Related Work | 185 |
| A.3 | Empirical Insights | 187 |
| A.3.1 | Blocked Content Prevalance | 188 |
| A.3.2 | Blocked Content Type | 189 |
| A.3.3 | Karma Scores for Post and Replies, Happyratio | 190 |
| A.3.4 | Votes per Posts. Blocked | 190 |
| A.3.5 | Flags | 191 |
| A.3.6 | Estimating Moderation Speed | 192 |
| A.4 | Modeling Disliked and Abusive Content | 193 |
| A.4.1 | Initial Datasets | 193 |

| | | |
|----------|---|------------|
| A.4.2 | Baseline: Naive Bayes | 193 |
| A.4.3 | Attention-Based Masked Language Model | 194 |
| A.4.4 | Modeling Threads only | 197 |
| A.5 | Leveraging the Model as a Baseline | 197 |
| A.6 | Future Work | 197 |
| A.7 | Conclusions | 197 |
| B | User Lifetime Insights and Modeling | 199 |
| B.1 | Introduction | 199 |
| B.1.1 | Research Questions | 200 |
| B.1.2 | Approach | 200 |
| B.1.3 | Results | 200 |
| B.2 | Related Work | 200 |
| B.3 | User Lifetime and Churn | 201 |
| B.3.1 | User Retention and Churn | 202 |
| B.4 | Modeling User Lifetime | 204 |
| B.4.1 | Goal: User Lifetime Prediction | 204 |
| B.5 | Features | 205 |
| B.5.1 | Users home community | 205 |
| B.5.2 | Capturing time | 205 |
| B.6 | Random Forest Implementation | 206 |
| B.6.1 | ML Algorithm Selection | 207 |
| B.6.2 | Independent Communities | 207 |
| B.6.3 | Predictor Sweet Spot | 208 |
| B.6.4 | Feature Subset Analysis | 209 |
| B.6.5 | Generalization | 210 |
| B.6.6 | Country Model in Detail | 211 |
| B.6.7 | Feature Relevance per Community | 212 |
| B.7 | Future Work: Empirical Lifetime Study | 212 |
| B.8 | Binary Lifetime Prediction | 213 |
| B.9 | Conclusions | 214 |
| C | Excursus: Data-Driven Long Term Gaming and QoE | 216 |
| C.1 | Introduction | 216 |
| C.1.1 | Research Questions | 217 |
| C.1.2 | Approach | 217 |
| C.1.3 | Results | 217 |
| C.2 | Related Work | 217 |
| C.3 | Long-Term Gaming and Data Collection | 218 |
| C.3.1 | Dataset and Data Collection | 218 |
| C.4 | Long-Term Game Dynamics Analysis | 218 |
| C.4.1 | Player Lifetime | 219 |
| C.4.2 | Inevitable Tension in Expansion | 219 |
| C.4.3 | Player Experience Indicators | 222 |
| C.5 | Conclusions and Future Directions | 223 |

Introduction

We have showcased how online platforms emerge and develop, how users interact, differences between countries, and uncovered properties, intents and topics of platform discourse. This chapter changes focus on **◆ User Management**. That is, we take a deep dive into Jodel’s distributed moderation system including modeling of abusive content, analyze and predict customer lifetime, and finish with an excursus on how to derive valuable user insights applying data-driven methods in long-term gaming and Quality of Experience.

Apparent platform threats: Distributed Content Moderation and Modeling. Especially anonymous platforms, such as Jodel, tend to become toxic [Papasavva et al., 2020, Zelenkauskaitė et al., 2021], promote bullying [Whittaker and Kowalski, 2015], or presents spaces for extremes [Baele et al., 2021]. Operators have learned that some regularization is necessary to keep user contents with a given frame: **★ A) Content Moderation**. While general moderation can be implemented in various fashions, a widespread approach resembles distributed moderation by selected moderators, or whole user bases. Research has provided valuable insights about distributed moderation setups: empirically [Trujillo and Cresci, 2022, Lampe et al., 2014], modeling [Stoddard, 2015], or governance [Fiesler et al., 2018]; while also expressing concerns [Lampe and Resnick, 2004, Zhu et al., 2021, Gilbert, 2013]. We discuss further related work and other influencing factors of user participation. Complementing work on existing platforms, we empirically investigate Jodel’s distributed moderation architecture across the German and Saudi communities. Given the higher blocking prevalence and user voting participating in Germany, we take a deep dive into further metrics w.r.t. voting, blocking, and flagging. By modeling blocked content with a state-of-the-art BERT-alike Masked Language Model, we show that general popularity prediction w.r.t. blocking remains a hard task. Resulting in huge amounts of false positives renders further estimates on the imbalanced actual data unusable; Yet, we describe how to leverage a suitable model as an (disliked/abusive) content-wise baseline, allowing for identifying relatively tolerant, or possibly toxic communities.

Long-term platform threats: Customer Lifetime Value. As all social networking platform depends on an active user-base, they also are threatened by *user churn*. Besides acquiring new users, trying to retain existing users renders a core marketing strategy [Kotler, 2016]. The measurement of **★ B) User Lifetime**, or customer lifetime value (CLV), denotes expected revenue over time in marketing, which introduces a possible value-weighting to user timespans. Knowing, or predicting soon-to-churn customers may allow for preventing losses. The user churn and prediction is a well studied data mining task within various regimes [Óskarsdóttir et al., 2020, Chen et al., 2015, Runge et al., 2014, Chamberlain et al., 2017], usually focusing single communities. Their generalization beyond a single user-base as found on Jodel remains unknown. Thus, we first characterize user lifetime and churn within Jodel empirically. Then we model user lifetime as a regression and classification problem. Though statistical modelling and distribution fitting has shown significant success [Fader et al., 2005b, Dror et al., 2012, Chen et al., 2015], we showcase Random Forests to provide very strong results alike shown on other platforms as well [Dror et al., 2012, Danescu-Niculescu-Mizil et al., 2013, Pudipeddi et al., 2014]. We provide an ablation study w.r.t. time-dependent feature subsets, and present a practical binary decision model (lifetime longer than timespan x). A single countrywide model generalizes well. We showcase and discuss how the model can be used to infer similarities between communities, and how feature importance measure might indicate important churn factors.

Enabling long-term user experience analysis leveraging data. While churn, i.e., a user's decision to leave a platform for good, can be interpreted as a rather extreme manifestation of user experience, the operator usually has interest in uncovering reasons and possibly prevent such happenings. However, as research suggests, also Jodel appears to be quite ephemeral, which is why our empirical investigation of User Lifetime RF model importances have been rather incomprehensive as of now. Thus, we believe that extending the methodology as applied for the churn analysis to a framed timespan would result in less noise. Thus, we discuss a new data-driven approach to **★ C) Long-Term Gaming Quality of Experience**. Some massively multiplayer online games (MMOG) are designed for long-term game rounds that last for months or even years. One example is TribalWars¹. Finishing a round inherently requires a very long-term commitment over years on a regular basis at cumulative costs to quit. Classical interactive tests in the order of few minutes suffice to assess the player experience for *most* games; however, widened timeframes are infeasible and questionnaire-based surveys do not scale to large populations of unknown users. We posit that this well-defined round structure and the surprising commitment of many players render games such as TribalWars prime candidates to study long-term gaming dynamics and user experience. Thus, we characterize the entire longitudinal development of a Tribal Wars game round consisting of thousands of villages and 16k players over 1.5 years. We provide insights to behavioral patterns of all active players. By identifying features that capture the in-game success and relate to the player experience, we show that only successful players keep up playing. We conclude that hypotheses mined and validated through data analysis provide a reasonable starting point for classical studies for Quality of Experience, making the study of long experience timeframes tangible.

¹<https://www.tribalwars.com>

A Distributed Content Moderation

Within this section, we explore the very heart of many online platforms responsible for regularizing whole communities by themselves: distributed moderation. After providing an intensive introduction to related search and other influencing factors of user participation, we empirically investigate the two stages of Jodel's distributed moderation architecture across the German and Saudi communities. While providing insights to the prevalence of blocking w.r.t. various metrics, we also provide insights to user flagged content supposedly violating terms of use, i.e., abusive content. Next, we model abusive content with BERT-like state of the art masked language models in comparison to a Naive Bayes baseline. While content specifically blocked also being flagged allows for better classification, overall model performance remains questionable due to real-world imbalanced data resulting in huge amount of false positives. Yet, we showcase how such a model may be used to establish a textual baseline across various communities that enable relative comparisons w.r.t. the classification target: *here* blocked content.

A.1 Introduction

We showcase how communities adopt platforms and the users interact within a given frame, also providing in-depth content analyzes. However, user and content moderation resembles the very heart of such platforms, regularizing content through distributed moderation and voting mechanisms is a popular, sustainable, and widespread approach.

Platforms may implement feedback signals differently, such as only linking, up- and down-voting, or flagging harmful contents are examples pursuing a similar task: Allow the large user base to provide feedback on specific content aiming to promote mainstream content, or fighting abusive or harmful contributions.

Such systems have been researched empirically [Trujillo and Cresci, 2022, Lampe et al., 2014], modeling popularity dynamics and content quality [Stoddard, 2015], or the governance behind the platform [Fiesler et al., 2018]. Besides dangers in biases [Lampe and Resnick, 2004, Zhu et al., 2021], widespread underprovision, i.e., only few votes, may become problematic [Gilbert, 2013]. Further, not only the voting mechanism itself influences platform contents, but also many other factors, such as external community pressure establishing certain (learned [Lampe and Johnston, 2005]) social norms and self-identifies [De Candia et al., 2022, Heuman, 2020b, Gaudette et al., 2021]. Arguable, gamification and social credit [Bosu et al., 2013, Movshovitz-Attias et al., 2013, Kusmierczyk and Gomez-Rodriguez, 2018] play another important role in user content and behavior [Cavusoglu et al., 2015].

While post quality and popularity has been subject to measurements and predictions [Zhang et al., 2018, Chang et al., 2015, Carmel et al., 2012], many approaches have been showcased on various platforms [Keneshloo et al., 2016, Huang et al., 2018, Kang et al., 2019, Chen et al., 2019, Ding et al., 2019, Mazloom et al., 2018], or incorporated temporal information [Abbar et al., 2018]. A wider focus on text quality is found in Q&A platforms [Arora et al., 2015, Anderson et al., 2012, Baltadzhieva and Chrupala, 2015, Correa and Sureka, 2014]. Other researchers applied neural embeddings [Tóth et al., 2020] or analyze specific architectural neural network depths [Ghosh and Ghosh, 2019].






Within this section, we complement existing works empirically investigating the prevalence of blocked content and its particularities, while later modeling abuse contents with state of the art Masked Language Models.

We identify about 5% blocked posts within the German user bases, whereas there is less blocked content within Saudi Arabia. A major reason might be the comparably few votes being cast in SA—a vote-active user base increases the likelihood of blocking. Nonetheless, post amounts blocked by flagging and subsequent moderation remain equal between both countries. Flagged Posts are more likely to be blocked in SA than in DE, which we estimate

to happening quite fast.

Given the blocking prevalence in DE, we fine-tune language models predicting blocking content. While modeling e.g., blocked content through moderation performs best, other blocked content cannot easily be classified. While an accurate model would allow for a text-quality baseline across the country to enable relative comparisons of community tolerance and *abusiveness*, achieved performance is not sufficient for reliably applying the LM and using its inferences due to a heavily imbalanced real world data w.r.t. blocking.

Structure

-  [A.2: Related Work](#)
-  [A.3: Empirical Insights](#)
-  [A.4: Modeling Disliked and Abusive Content](#)
-  [A.5: Leveraging the Model as a Baseline](#)
-  [A.7: Conclusions](#)

Acknowledgements I thank Leon Wolter who was instrumental for this topic by helping implement and execute MLM training, and providing the Naive Bayes baseline in his Master Thesis.

A.1.1 Research Question

Within this section, we want to shed light on the heart of the community internals: distributed content voting and flagging with subsequent moderation. We are interested in the prevalence of blocked contents by types, across time and community size and other potential influence factors.

Further, we ask whether we could create a suitable state of the art language model (LM) to predict blocked contents. Such a model allows for establishing a textual quality baseline w.r.t. the optimization target. High amplitude discrepancies in predicted and actual blocked contents may indicate wonderful or bad communities being very strict, or more tolerant in accepted contents.

A.1.2 Approach

We set out to provide a unique empirical view into our ground truth information on Jodel w.r.t. platform blocking and detail the community landscape by interaction volume. Further, we develop a state of the art Masked Language Model (MLM), i.e., we fine tune various pre-trained BERT [Devlin et al., 2019a]-compatible models of which GottBERT [Scheible et al., 2020] performed best—and better than a Naive Bayes baseline. While detecting blocked content remains a hard problem, the deep learning approach achieves considerably better performance than a Naive Bayes baseline. Though the model performs reaches around 70-75% accuracy, we encounter strong limitations in actual deployments due to the heavily imbalanced real-world dataset w.r.t. non-/blocked content; the model (naturally) identifies large amounts of false positives. Nonetheless, under the given model's uncertainty, we outline how a language model on a country-wide scope may allow for a textual baseline tool: Are there communities that are considerably worse in text quality, but simultaneously deviate from this in terms of community feedback?

A.1.3 Results

We characterize the prevalence of blocked posts w.r.t. various metrics within Jodel Germany (DE) and Saudi Arabia (SA). Correlating to the user base enjoying voting more than the Saudi counterparts, we find that increased amounts in votes lead to increased figures in blocked content. Blocked content via moderation however is equal in both countries; the moderation system acts quite fast.

We fine-tune a Masked Language Model (GottBERT) to identify blocked contents. Achieved performance remains below expectations at about 73% accuracy. This becomes a problem in subsequent application on real world data due to its imbalanced nature (only 5% of DE posts are blocked)—resulting in huge amounts in false positives. Lastly, we showcase how to establish a qualitative textual baseline from such a model.

A.2 Related Work

Moderating user contents seems to be a necessary step towards functioning communities and have found widespread use. Many platforms use expert moderators, either paid, or on a voluntary basis. Others employ specific companies outsourcing e.g., moderating user provided pictures raising controversy [Newton, 2019].

Distributed Moderation. However, distributed regularization and moderation approaches have shown to be successful and sustainable. This comes at the cost of potential of misuse by platform others or the platform operator [Crawford and Gillespie, 2016]. While *moderation signals* may be implemented differently on various platforms, e.g., up- and downvoting, thanks, or flagging may be implemented platform, they pursue a similar task: providing feedback to the system that leads to some kind of moderation decision usually being potentially removed and the potential of user-based consequences in severe cases. Content moderation can be attributed to various stages and factors of moderation. That is, the community regulates itself *i*) through social credit and reward positively-reinforcing, *ii*) widely distributed not or explicitly disliking unwanted content, user flagging of critical content, and *iii*) smaller groups of community moderators. Such systems in the wild have attracted lots of research in this area. We find empirical studies, e.g., on Reddit investigated moderation effects [Trujillo and Cresci, 2022], popularity dynamics and content quality [Stoddard, 2015], or the governance behind the platform [Fiesler et al., 2018]. Others have analyzed slashdot [Lampe et al., 2014], drawing an overall positive conclusion in effectiveness due to high consensus [Lampe and Resnick, 2004]; however, at times, the moderation processes seems to be too slow, or incorrect decisions are not reversed. They suggest that positional bias highly influences user behavior [Lampe and Resnick, 2004]; or as shown in [Zhu et al., 2021]. Notwithstanding, [Gilbert, 2013] provides evidence of moderation being non-functional due to widespread underprovision of the moderation system receives too little feedback, i.e., too few users actively participate.

Community Pressure. The environment users engage teaches them social norms, identities, liked and (non-) accepted contents by observation and active feedback. Further, social platforms implementing user profiles allow for social credit; or the platform and design may nudge users to behave *well*. We next discuss these factors in more detail.

Leaving moderation up to platforms users also results in (self-inflicted) specific norms [De Candia et al., 2022], also shown for Jodel [Heuman, 2020b], and reflect self-identities [Gaudette et al., 2021], that may be framed by the operator. Platforms users learn such norms by observation and community feedback (distributed moderation) [Lampe and Johnston, 2005]. Besides such passive factors, active interaction on social platforms allow for

gathering social credit and reputation in terms of likes and followers, e.g., on StackOverflow [Bosu et al., 2013, Movshovitz-Attias et al., 2013]. Further, application and UX design may also nudge users towards well behavior, honoring positive interactions, while possibly punishing bad behavior—like having posted content that is being removed afterwards due to a distributed moderation decision. Gathering some kind of currency belongs to the regime of gamification as analyzed with badges in [Kusmierczyk and Gomez-Rodriguez, 2018]. Others take a more general perspective investigating whether gamification may lead to higher user responses [Cavusoglu et al., 2015]. Platforms should further consider being transparent to their users w.r.t. content removal as shown in [Jhaver et al., 2019].

Quality and popularity. Efforts have been made to analyze factors of post popularity and quality across various platform and hence, varying contexts.

While some research provides data-driven approaches determining factors for social media popularity, e.g., [Zhang et al., 2018], amongst surveying aligning processes along Facebook fan pages [Chang et al., 2015], [Carmel et al., 2012] analyze post novelty factors. The landscape of applied approaches is ever-increasing. Classical approaches like various regression techniques on news articles [Keneshloo et al., 2016], using Random Forests [Huang et al., 2018], or boosting algorithms [Kang et al., 2019, Chen et al., 2019] still provide good performance. With adding temporal information, e.g., [Abbar et al., 2018] propose using Latent Dirichlet allocation (LDA) within time series forecasting. Many rely on crafted features, however, recent advances in Natural Language Processing (NLP) lead to using deep neural architectures [Ding et al., 2019] focusing on text embeddings. While the mode text has attracted lots of research, [Mazloom et al., 2018] combine text with images. As for pictures, [Zohourian et al., 2018] have analyzed popularity on Iranian Instagram, whereas multimodal approaches can provide more valuable input signals [Mazloom et al., 2017].

Estimating the popularity of certain posts is a question particularly interesting within Online Social platforms. However, we find the application of distributed moderation within content focused platforms like Stack Overflow [Anderson et al., 2012, Baltadzhieva and Chrupala, 2015] as well, possibly with a focus on deleted questions [Correa and Sureka, 2014] identifying most deletions being manual labor by the moderators. Scoring questions on these networks has likewise been of interest in [Arora et al., 2015]. Another line of research tries to employ quality estimations for automatic moderation or prefiltering [Lampe et al., 2007]. Approaches leveraging pure text using Naive Bayes [Delort et al., 2011], or deep learning methods (e.g., Recurring Neural Networks, RNNs) [Tóth et al., 2020] have been applied to this problem. A more general approach has been analyzed by [Ghosh and Ghosh, 2019] elaborating on neural network depths w.r.t. model prediction performance for predicting deletion on Q&A platforms.

The Moderators. While distributed moderation on a larger scale is performed by all or many users, certain platform still rely on voluntary select community moderators as a second escalation barrier. That is, [Yang et al., 2014] provide insights to behavior of senior users on Stackoverflow, while [Li et al., 2022] highlight the effort that is undertaken by such moderators.

Hate and Fake News. Not only anonymous platform, such as 4chan tend to create toxic environments [Papasavva et al., 2020], but regular public platforms likewise struggle with general hate, filter bubbles, and fake news/misinformation, and propaganda as surveyed in e.g., [Zhou and Zafarani, 2020, Oshikawa et al., 2018].

Fake news detection within media platforms is vital task against such campaigns. That is, research proposed applying supervised learning methods [Reis et al., 2019], applying NB [Granik and Mesyura, 2017], multimodel approaches [Singhal et al., 2019], or multi-

| Country | #* | #out* | #mod* | p(out) | p(mod) | p(blocked) |
|---------|-----|-------|-------|--------|--------|------------|
| DE | 280 | 12.34 | 3.41 | 0.044 | 0.012 | 0.056 |
| SA | 437 | 3.10 | 4.04 | 0.007 | 0.009 | 0.016 |

*Million

Table A.1: Overall Blocked Content - DE & SA. While blocking frequency due to moderation efforts is within the same regime, the distributed voting mechanism leads to less outvoted contents within SA. We overall observe positivity reflected in only about 5.6% (1.6%) total blocked contents in DE (SA)

class classification [Karimi et al., 2018], amongst many others [Jain and Kasbe, 2018, Pérez-Rosas et al., 2018]. While not only prediction is of importance, other work focuses on understanding user structures behind fake news [Shu et al., 2018]; others have applied geometric deep learning [Monti et al., 2019], or tried to make the modeling explainable [Shu et al., 2019]. Above the text level, [Karimi and Tang, 2019] investigated discourse within hierarchies. Also peer to peer approaches have found their place for e.g., fact checking news [Jiang et al., 2014].

Of specific interest e.g., also also been fake news in the realm of the COVID-19 pandemic [Wani et al., 2021], or our analysis on censorship (see 4).

Besides fake news, hate is another serious problem on any internet platform. The authors of [Benítez-Andrades et al., 2022] apply various deep learning architectures (CNN, LSTM, Transformer-BERT) to identify racist and xenophobic contents within Twitter messages. Other work focusing on training state of the art attention-based transformer models are e.g., HateBERT [Caselli et al., 2020] (cf. C.2.2). It has shown good performance on various evaluation datasets [Zampieri et al., 2019, ?, Basile et al., 2019].

(Controversial) Active Measurements. The presented work—as the vast majority—of empirical studies base on observation only, marking it hard to create causality. Albeit being seriously delicate w.r.t. ethics, [Weninger et al., 2015] analyze a butterfly effect: By actively voting Reddit content up- or down within the early lifetime, they provide biases and show that the community reinforces these trends in comparison to a control group.

A.3 Empirical Insights

As discussed in related work, there is a multitude of factors influencing user and community behavior. Thus, we first set out defining blocking and discuss which signals may be obtained from our dataset. We then characterize the German and Saudi Jodel community landscape empirically by size and time.

Defining Blocking On Jodel, posts may be simply disliked by the distributed voting mechanism. As soon as they accumulate *vote* scores of a negative threshold (usually -5), they are not displayed anymore, i.e., they become blocked. However, posts may also violate terms of use and might therefore be flagged for subsequent moderation. That is, posts above this threshold must be blocked due to *moderation*—this is an underestimation, as many posts that would have been blocked are often outvoted by the community voting anyway.

Dataset Limitations Unfortunately, our dataset does not include timestamps for when vote interactions as there were mapped to post creation. Thus, we cannot leverage any voting sequences not qualitatively, nor quantitatively w.r.t. time. Further, as described, the dataset does not contain additional information about the (possible) moderation process.

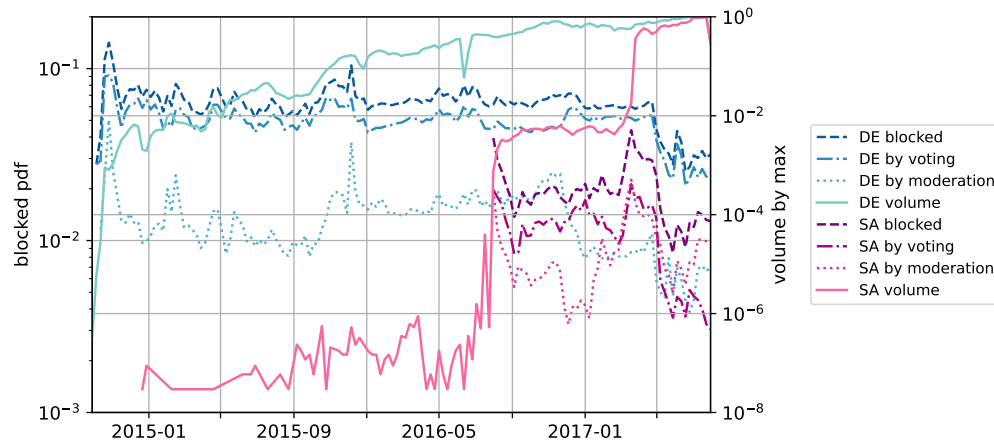


Figure A.1: Blocked Content Ratio by Type over Time/Volume - DE & SA. Weekly. The amount of blocked posts and the fraction of specifically moderated posts remains stable within the German communities. Turbulences within the user base of Saudi communities as describe earlier are reflected by blocked content figures.

A.3.1 Blocked Content Prevalance

A.3.1.1 Overall

First, we want to provide a picture about how many posts at all are blocking the platform as shown in Table A.1. While this amounts to about 5.6% for the German communities, only 1.6% of the Saudi contents end up being blocked.

Dissecting the amount of blocked content by blocking type as define previously, we find that the content frequency of blocked content that has been removed due to moderation is within the same regime at 1.2% in Germany and 0.9% in Saudi Arabia. However, content competing harder for votes, even the downvotes suffer, hence we observe less blocked content determined by the distributed voting mechanism (see A for more detail).

A.3.1.2 Over Time

After drawing the big picture, we are next interested in the development of blocking figures over time. As shown earlier, the communities evolved and naturally encounter lots of churn (see B). Thus, we take a deeper look into the frequencies of content and the blocked posts in Figure A.1 for both countries, Germany (DE, blue) and Saudi Arabia (SA, pink). The x-axis denotes time in weekly aggregates, whereas the logyrrithmic left y-axis the Probability Distribution Function (PDF) denotes. The right y-axis is assigned to the volume normed each country series maximum occurrence.

For Germany (DE, blue), the total amount of posts is ever-increasing on the logarithmic scale as observed earlier (see C). As such, the total amount of blocked posts remains widely stable at 5% to 8%, however we observe a significant change in June 2017, where the figures suddenly drop to a level between 2% and 4%. The amounts of content blocked through moderation likewise remains mostly stable with increased figures between beginning of 2016 to end of 2016; the sudden drop also occocurs for moderated content at high variance.

For the KSA communities, we can spot various stages in platform evolution. While there is almost no interaction on the platform at all until late 2016, first users enjoyed occasional communication at only few hundreds of posts across the country;. This stage gets then replaced with the described heavy influx of new users and new heights in platform interactions (as described in C). At first, amounts in moderation content increases specifically w.r.t. moderated content, however, the same drop that has been observed within the DE

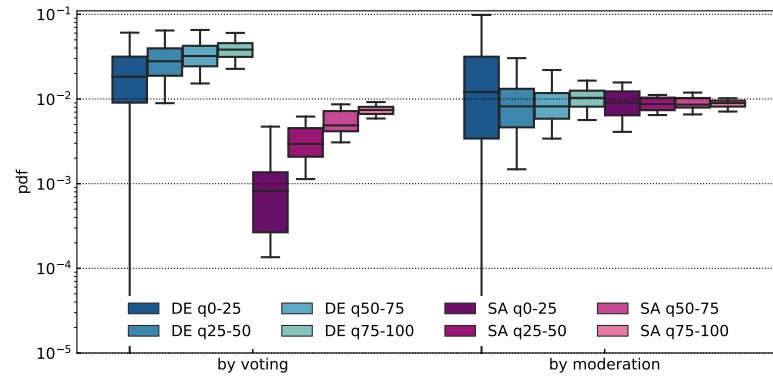


Figure A.2: Blocked Content Ratio by Type per Community - DE & SA. The German users deem voting more popular than the Saudi user base; thus, outvoted content appears to be more prevalent in DE compared to SA. For both countries, we identify a clear upwards trend in blocking figures with increasing community interaction volumes.

data happens here as fell: blocked content volumes drop from 2% down to 1%.

We attribute this distributional shift to a platform-wide change of the out-voting threshold from -5 to -10 at the beginning of May 2017. We keep influences and reasons for this change observed in DE for future work.

A.3.1.3 By Community Size

Next, we were interested whether and to which extent the distributions of blocked contents vary across community size. This, we present a boxplot of blocking PDFs across community quantiles by size for Germany (DE, blue) and Saudi Arabia (SA, pink).

Focusing on blocking by moderation first, the median fraction of blocked posts remains equal throughout all counties and community sizes. However, smaller communities tend to increase variation. The blocked content through distributed voting by them communities themselves is different between DE and SA. The popularity of vote interactions in DE communities leads to greater extents in outvoted content compared to SA. Within both countries, we observe a clear trends in higher outvoting figures in larger communities.

A.3.2 Blocked Content Type

We have now seen how blocked posts are distributed across community by size and how shifts happened at turbulent times. As of now, we have not distinguished between the available content types: Posts, Replies and Images, which we discuss next in terms of frequencies and the blocking method with Table A.2. The upper two rows relate to replies and posts, whereas the lower two rows represent text and images for both countries, DE & SA.

As for replies and posts, we observe a massive overhead in replies, of which however, fewer far get blocked at about 2.9%, whereas the posts dominate with 18.6%. That is almost one out of six posts becomes ultimately blocked on the platform. The most common reason for blocking is through moderation in DE (SA): The replies are moderated at 0.5% (0.8%) totals, whereas specifically moderated posts account for 4.5% (7.5%). As such, we observe an increase in moderating figures in SA compared to DE.

As for images versus posts, general blocking of images occurs to a higher extent with 12.2% (9.7%) than texts at 5.4% (1.4%) for DE (SA). Again, the moderation figures for SA posts are increased compared to DE; nonetheless, images are not as aggressively outvoted as other posts in DE—quite similarly in SA.

| Country | type | #* | #out* | #mod* | p(out) | p(mod) | p(blocked) |
|---------|-------|--------|--------|-------|--------|--------|------------|
| DE | reply | 231.48 | 5.522 | 1.224 | 0.024 | 0.005 | 0.029 |
| | post | 48.26 | 6.822 | 2.191 | 0.141 | 0.045 | 0.186 |
| SA | reply | 382.59 | 0.199 | 2.830 | 0.001 | 0.007 | 0.008 |
| | post | 54.59 | 2.903 | 1.209 | 0.053 | 0.022 | 0.075 |
| DE | text | 270.86 | 11.679 | 2.999 | 0.043 | 0.011 | 0.054 |
| | img | 8.88 | 0.665 | 0.416 | 0.075 | 0.047 | 0.122 |
| SA | text | 421.64 | 1.988 | 3.646 | 0.005 | 0.009 | 0.014 |
| | img | 15.54 | 1.114 | 0.393 | 0.072 | 0.025 | 0.097 |

*Million

Table A.2: Overall Blocked Content By Type - Images, Posts and Replies, DE & SA. While mostly posts dominate blocking figures in a relative scale, images also get more attention on average compared to text.

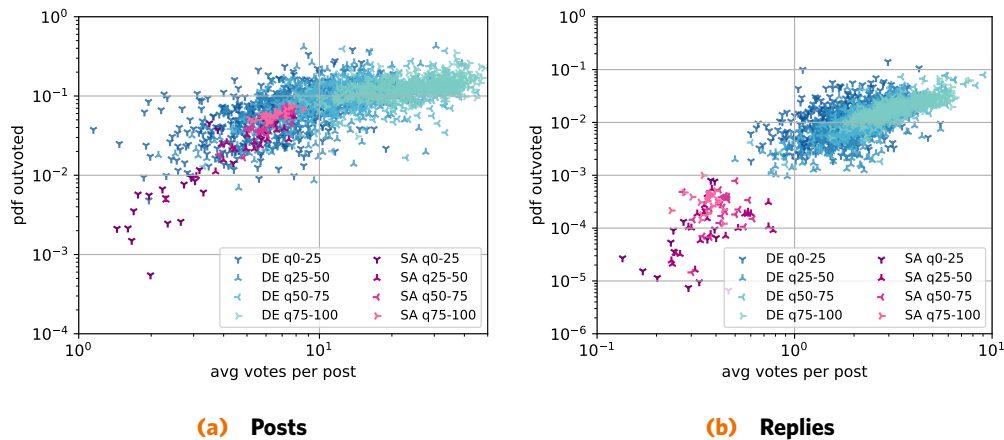


Figure A.3: Outvoted Content Ratio to AVG Votes per Post, DE & SA. DE sampled at $n=500$. With increasing amounts of votes, which happen in the larger communities, also more content gets outvoting by the distributed voting mechanism. However, the overall amount of outvoted and therefore disliked content remains rather small; as shown earlier, specifically within the SA communities.

A.3.3 Karma Scores for Post and Replies, Happyratio

We would like to refer the reader to our earlier discussion about usage differences between the German and Saudi communities in B; Futher, we provide deeper insights to the happyratio on Jodel within SA in C. The average Saudi user votes as much as the German counterpart, however, the volume in posts within Saudi Arabia appears heavily shifted towards content creation. Nonetheless, overall voting behavior remains positive across any communities size and country.

A.3.4 Votes per Post s. Blocked

Next, we incorporate another dimension into our analysis. While distinguishing between both countries and communities by size w.r.t. volume, we provide a scatter plot of the community relation of average votes per post to blocking frequencies in Figure A.3. Note that we compare to outvoted content only, focusing on the distributed moderation via voting only. We subdivide our investigation further into posts in Figure A.3a and replies in Figure A.3b, of which the German communities are sampled to $N=500$, which we deem sufficient to convey qualitative insights.

Within the sampled set of German communities, we observe more votes per post within larger communities (as discussed in B). However, there is a clear trend: With more votes per posts, the amount of outvoted content increases. The same observation holds true for SA posts and to a lesser extent for replies.

These observation may indicate an ongoing under-provision in votes as discussed in [Gilbert,

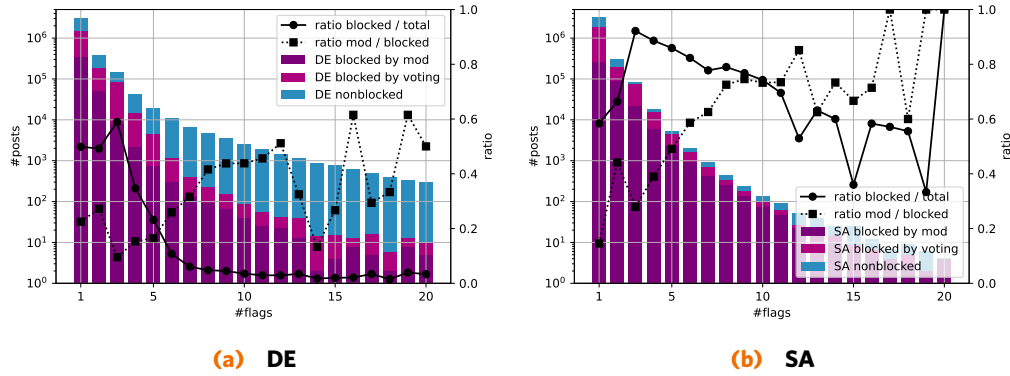


Figure A.4: Flagging vs. Blocking - Totals and Ratios, DE & SA. *left: (a)* Within the German communities, non-moderated posts remain in the system for quite some time gathering even more flags. *Right: (b)* In the very vivid Saudi environment, posts are unlikely to gather comparably much more flags. The overall influence of flagging is larger in SA.

2013], yet the amounts of blocked content through moderation appear equal between the DE and SA communities. We would like to refer to future work for deeper insights into this topic.

A.3.5 Flags

Besides discussed blocking and voting, we next discuss flags in particular as a necessary requirement for a post to be moderated.

A.3.5.1 Blocked Posts vs. Flagging

We start with an insight into overall casted flags and the correlation to actual blocking results in Figure A.4. We distinguish between the German user base in Figure A.4a, and the Saudi communities in Figure A.4b. The x-axis denotes the amount of received flags of a post, whereas the left logarithmic y-axis describe the amount if posts within that category. The right y-axis denotes relative ratios of overall blocked posts and the amount of moderated posts w.r.t. all blocked posts. Note that we have added three series of flagged posts: *i*) posts blocked by moderation, *ii*) posts blocked by distributed community voting, and *iii*) posts flagged, but not blocked as a result. *Note:* The logarithmic y-scale may appear deceiving; The ratios provide a good baseline for comparison.

While for the DE communities, the amount of blocked contents from all flagged is at about 50% for few flags covering the cast majority of cases. However, with increasing amounts in gathered flags, the relative amounts of non-blocked content increases dramatically. Simultaneously, the amount of moderated content increases up to 50%. We generally assume the Jodel moderation being quite fast within its vivid communities. Therefore, most posts get either voted out by the community when violating terms, or might end up in moderation after only a single flag. Posts that might be controversial (possibly w.r.t. topic without necessarily violating terms), but may already have passed moderation, likely may encounter even more flags. Unfortunately due to our dataset, we cannot infer any insights on the timeline of flagging interactions.

The user base in SA tells a different story. Overall figures in blocking resulting from flagging are comparably high and only increase in the amount of received flags. We attribute the low amount of non-blocked flagged posts to the high frequency that is apparent within the large SA communities.

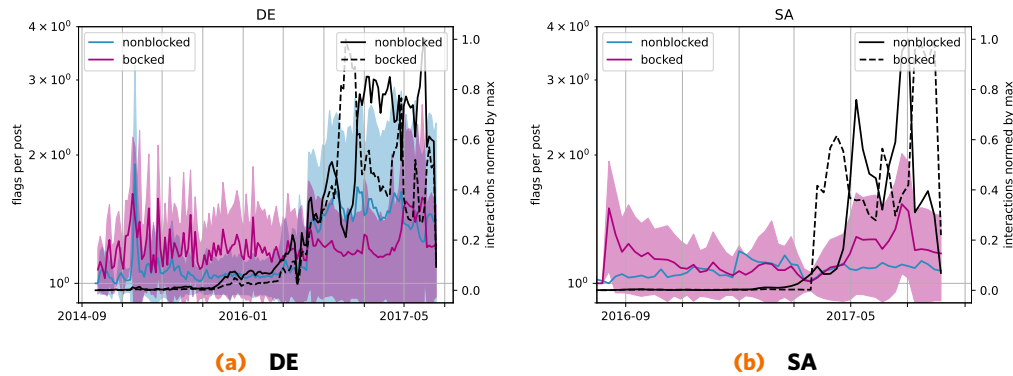


Figure A.5: Flags per Post, DE & SA. While ultimately blocked content consistently receives 2 to 5 flags, non-blocked content gather even more flags within the German communities. Changes in post volume do not have an apparent effect on flagging ratios; albeit we observe a pattern change due to particularities in SA (see 4).

A.3.5.2 Flags per Post over Time

To provide a broader understanding on flags happening within the platform, we next analyzed the average amount of flags per post over time as shown in Figure A.5. We distinguish between DE Figure A.5a, and SA Figure A.5b. While the x-axis denotes time, the left logarithmic y-axis denotes the flag frequency per post. The colored series represent averages with additional standard deviations (area) across blocked (pink) and non-blocked (blue) content. The right y-axis describes the amount of platform interactions normalized by maximum, likewise for non-blocked (black, solid) and blocked (black, dashed) content.

For (a) Germany (DE, left), we observe a long timeframe until end of 2016 where the flagging of ultimately blocked content happens slightly more often than those of non-blocked contents. However, this rigorously shifts significant into the other direction: Whereas non-blocked content receives about 2 to 3 flags each, these figures are much lower for blocked content. The latter remains mostly stable. Again, we presume that due to longer visibility in German communities, critical comments may gather more flags, while not violating the platform terms as decided by moderation. Fluctuations in volume have no impact on observed trends.

Within the (b) Saudi Communities (SA, right), we observe similar behavior alike seen in Germany in the early stages. However, after the huge influx of new users, flags on ultimately blocked content increases significantly. Fluctuations in volume have no apparent impact on the observed trends in SA as well.

A.3.6 Estimating Moderation Speed

Given our observation that most posts that are ultimately blocked only receive few flags until being moderated or disliked by the distributed community voting anyway. Lastly, we are now interested in the moderation speed and agility.

Due to lack of specific voting or flagging timestamps, we need to estimate the post lifetime through thread replies. As long as a post has not been blocked, people are able to reply. Hence, we filter for all threads having at least a single reply. From the set of replies, we calculate a lower-bound thread lifetime as the maximum difference between post and reply. Note however that threads become outdated and receive less attention over time.

This still allows for a comparison of average thread lifetimes between blocked and non-blocked threads. We distinguish further into community size by interaction volume in Figure A.6. We provide details for DE in Figure A.6a, and alike insights to the SA user base in Figure A.6b. While the x-axis denotes the categorial community quantile w.r.t. interaction volume, the y-axis denotes a logarithmic time axis for the threads. We add the median

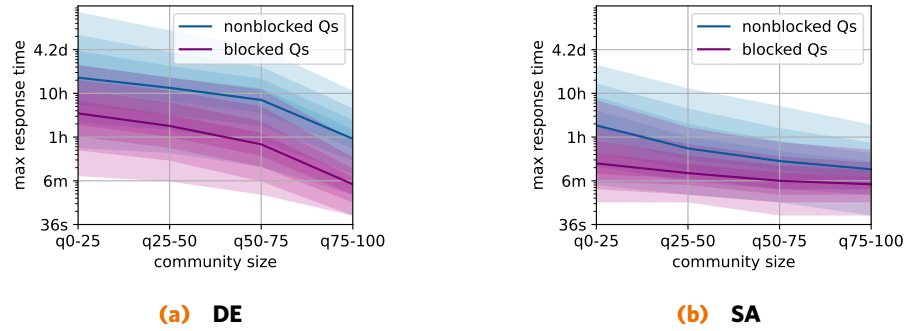


Figure A.6: Thread Max Response Times - Non-/Blocked Content, DE & SA. We observe a significant difference in thread lifetime between blocked and non-blocked content. Content that ultimately has been blocked remains on the platform widely only a couple of hours (minutes) even within the smaller communities in DE (SA).

lifetime series for blocked and non-blocked threads in comparison. Additionally, the areas denote the sibling-quantile (q40-60, q30-70, ...).

First, the thread lifetime within Saudi Arabia are significantly lower compared to the German counterparts. Both countries level within the larger communities. Further, the thread lifetime of ultimately blocked threads is also significantly lower compared to others, non-blocked. The platform reaction speed to block content is larger for the smallest communities, but remains in the realm of hours (minutes) in DE (sA).

A.4 Modeling Disliked and Abusive Content

Our next step is creating a model for content that is disliked or moderated. This process will feature two basic outcomes: *i)* We gather information how well we can model a classifier on a country wide scale, e.g., whether it generalizes. And *ii)* The obtained language model can then be leveraged as a text-quality baseline throughout the country enabling a relative comparison for community-*roughness*, i.e., whether it allows content that would not be accepted by the model, hence other communities.

A.4.1 Initial Datasets

For further model creation and evaluation, we first define used datasets for the upcoming baseline and model fine-tuning. We generally use 20% of the data for evaluation and test; blocked and non-blocked content account for equal shares.

Varying Dataset Size. We use various sizes in datasets for further analysis. While default to the 8M set, we provide insights into apparent scaling laws w.r.t. used dataset size $\mathcal{D}_{size}^{rand}$, $size \in \{2.5M, 5M, 8M, 10M, 20M\}$.

Varying Blocking Reason Further, in anticipation of different *difficulties* in moderation decisions, we employ several datasets representing a random sample across all posts, those blocked by moderation, by voting, or blocked with high and low voting consensus: \mathcal{D}_{4M}^{type} , $type \in \{all, vite, mod, high\ consensus, low\ consensus\}$; for a qualitative outline, we believe that a rather small dataset suffices.

A.4.2 Baseline: Naive Bayes

While we will later on user state of the art Masked Language Models internally relying on attention-based transformers, a Naive Bayes [Manning et al., 2010] approach provides a

strong baseline and very competitive low computational costs as demonstrated in, e.g., [Granik and Mesyura, 2017, Delort et al., 2011]. An NB classifier does not rely on specific text embeddings, but handles words as distinct input tokens. These tokens are taken into consideration in a bag of words approach of which then conditional probabilities are then calculated via leveraging Bayes rule. This is refined until fitting a complete corpus of text: for each class C , word frequencies per class and overall are gathered. Due to belonging to the group of supervised representation learning methods, labelled text data is necessary for training—alike in our later MLM approach. This allows for a computational maximization of achievable probabilities matching a given class.

Though the NB classifier performs well at low computational costs, by working on a word-token level, it inherently lacks the possibility to leverage, e.g., semantics that are salient within given input text via, e.g., grammar.

We will provide NB baseline figures where suitable.

A.4.3 Attention-Based Masked Language Model

Natural Language Processing (NLP) has leapfrogged in recent years with the simple idea of attention and the subsequent introduction of the transformer architecture [Vaswani et al., 2017] and subsequent Language Model designs, such as BERT [Devlin et al., 2019b] or the GPT family [Radford et al., 2019].

A.4.3.1 Choosing Pre-Trained Model

As for simplicity due to availability and proven expressiveness, we decided for a BERT-based model. Such models are pre-trained on large corpora of data gathering major context-dependent statistical corpus knowledge. End-to-end training for each application becomes cumbersome; thus, these models allow for fine-tuning after general training. That is, weights may be adapted to learn a specific downstream task usually within only a few epochs of training.

Within this section, we discuss hyperparameters tested on various models to determine the base model with which we proceed.

Hyperparameters We conduct a hyperparameter search with parameters provided in Table A.3. We test a bunch of various common hyperparameters, of which we detail the most important ones later. Further, we adopted two other techniques that might increase model performance. Proposed in [Zhang et al., 2020] being similar to [Howard and Ruder, 2018], *LLRD* aims at focusing the upper layers within the multi-headed attention hierarchy as they are believed to be responsible for more problem-specific aspects and high-level contexts. Their method implements decaying, but higher learning rates for the upper layers. Similar to dropout, and likewise focusing upper layers, [Zhang et al., 2020] improve fine-tuning by re-initializing whole layers at random.

| Hyperparameter | Search Space |
|--------------------------------|----------------------|
| Batch Size | 8, 16, 32 |
| Weight Decay | 0.01 to 0.3 |
| Initial Learning Rate | 0.0000005 to 0.00006 |
| Head Learning Rate Increase | 1 to 25 |
| Decay Rate | 0.65 to 0.85 |
| Number of Layers Reinitialized | 0 to 8 |
| Epochs | 6 |

Table A.3: Hyperparameter Search Space.

Lightweight Finetuning Results We present our model fine-tuning results in Table A.4. Each row corresponds with a pre-trained model, whereas the columns denote accuracy and loss metrics. The overlined values denote averages, whereas the title denotes the median. We find that the GottBERT [Scheible et al., 2020] model performed best on our dataset with consistent comparably high accuracies up to 73.3%.

| Model | \overline{Acc} | $Acc \text{ max}$ | \overline{Loss} | $Loss \text{ min}$ | #Runs |
|------------------------------------|------------------|-------------------|-------------------|--------------------|-------|
| GottBERT [Scheible et al., 2020] | 0.7305 | 0.7335 | 0.54592 | 0.538 | 10 |
| SentimentBERT [Guhr et al., 2020] | 0.6930 | 0.7154 | 0.5804 | 0.563 | 13 |
| German BERT [deepset AI, 2019] | 0.7179 | 0.7236 | 0.5618 | 0.5488 | 16 |
| MLM-RoBERTa [Conneau et al., 2020] | 0.6349 | 0.7211 | 0.6126 | 0.5533 | 11 |

Table A.4: Various Pre-Trained Models - Preselection results. From tested models with limited fine-tuning efforts, we identify GottBERT performing best for our purpose.

A.4.3.2 Finetuning

Having fixed the use pre-trained model, we next engaged more detailed fine-tuning. That is, besides Adam, we applied regular SGD performing worse at identical runtimes. Increasing the batch size from 64 to 128 did not result in a model improvement.

Dataset Size We next wanted to do an ablation study w.r.t. used dataset sizes. Usually, performance increases only logarithmically with dataset scaling. In Table A.5, we show performance results using various dataset size as defined earlier. As expected, larger dataset sizes tend to generalize better, however performance benefits come at massively increased computational costs. Further, the amount of available data is inherently limited.

| Dataset | Acc | Loss |
|----------------------|---------------|---------------|
| $\mathcal{D}_{2,5M}$ | 0.7231 | 0.5514 |
| \mathcal{D}_{5M} | 0.7282 | 0.5464 |
| \mathcal{D}_{10M} | 0.7340 | 0.5363 |
| \mathcal{D}_{20M} | 0.7398 | 0.5252 |

Table A.5: Dataset Size. Large training datasets yield better performance as expected.

Blocking Type, Difficulty and NB Baseline As described earlier, we curated specific datasets representing various types of blocking. That is, we next train models on datasets containing blocked contents with additional filters: *i)* \mathcal{D}_{all} as a baseline random sample, *ii)* \mathcal{D}_{vote} only containing blocked examples that have been voted out, *iii)* \mathcal{D}_{mod} only containing blocked examples that have been moderated, *iv)* \mathcal{D}_{hc} which only selects posts with high consensus in the voting decision, and *v)* \mathcal{D}_{lc} which only selects posts with low consensus in the voting decision.

The Naive Bayes (NB) classifier achieves 65%-66% accuracy. As expected, the more powerful transformer based neural architecture achieves better results of up to 8%. By comparing model performance on the different datasets, the contents blocked by moderation are apparently easier to grasp for such models as accuracy rose to 77% in this instance. The low consensus dataset was the hardest to model. Albeit his model being of smaller size than previous tests, the performance on the *all* dataset turned out superb.

Our findings confirm the hypothesis that certain contents are easier to model. This does not come unexpected: If the community is undecided (low consensus), how should the model know better? Further, blocking through the moderation process likely violate certain fixed platform rules that follow more homogenous patterns across the Jodel communities. Simple mainstream dis- and liking on the other hand may be highly individual from

community to community, establishing local identifies and enforcing social and behavioral norms.

| Dataset | GottBERT Acc | Naive Bayes Acc |
|---------------------------|----------------|-----------------|
| \mathcal{D}_{all}^{AM} | 0.73890 | 0.6613 |
| \mathcal{D}_{vote}^{AM} | 0.73033 | 0.6546 |
| \mathcal{D}_{mod}^{AM} | 0.77143 | 0.6879 |
| \mathcal{D}_{hc}^{AM} | 0.72393 | 0.6513 |
| \mathcal{D}_{lc}^{AM} | 0.72176 | 0.6501 |

Table A.6: Dataset Blocking Type / Model Difficulty, Naive Bayes Baseline. We observe apparent better performance for the moderated dataset; we presume that moderated content follows a clearer pattern w.r.t. platform terms of use, whereas general voting is based on dis-/linking content. The neural embedding solution works better than the NB counterpart.

A.4.3.3 Complementary Triplet Loss

While our current approach uses a CrossEntropy Loss maximizing the distinction between both classes, further opted for creating a triplet embedding as a pre-step in the hope that this embedding eases decision making. However, with mild training, results were not satisfactory. We conclude that possibly (a lot) more training may help, which we leave to future work.

A.4.3.4 Discussion

As we have trained a model to our restrained capabilities w.r.t. time, performance has merely reaches up to 73% accuracy. To showcase the implications, we discuss confusion matrices of the 50:50 training evaluation set, and the overall dataset.

50:50 training Given an achieved performance of about 70% accuracy, one would expect false classifications. To gain better insights, we plot the confusion matrix of the evaluation set in Table A.7. False Positives and False Negatives are roughly at level across both classes.

| | | predicted | |
|---------|-------|-----------|--------|
| | | False | True |
| actuals | False | 38.51% | 11.49% |
| | True | 14.61% | 35.39% |

Table A.7: Evaluation Set Confusion Matrix. In line with overall performance within 70% accuracy, one would expect deviations; False Positives and False Negatives are roughly at level across both classes.

Error Propagation on Entire Dataset As we are further interested in using the Language Model as a text quality baseline across communities, while being trained on a German random sample, we next analyze the obtained inference from the model across all available data—except training data.

As we have seen earlier, the amounts of content being ultimately blocked is generally quite low at about 5%. Thus, our platform dataset is highly biased. Though our model might perform quite ok-ish, this skew is responsible for huge error counts propagated into the dataset. We provide detailed figure in the confusion matrix in Table A.8.

The observed imbalance in the dataset leads to very huge figures in False Positives, that is, many posts are flagged as block-worthy by our language model, that in fact have not been blocked. At such expected high error rates, this model might be used to suggest posts for moderation, but by no means should be used for any automatic action.

| | | predicted | |
|---------|-------|-----------|--------|
| | | False | True |
| actuals | False | 72.94% | 21.69% |
| | True | 1.52% | 3.85% |

Table A.8: Inference Confusion Matrix. Due to the highly imbalanced dataset in the wild, the model error propagates into totals figures, leading to very large amounts in False Positives.

Effectiveness It is noteworthy that our approach using $\mathcal{O}(n^2)$ attention in a complex neural network is quite costly w.r.t. computational resources, hence energy and CO-2 footprint. The Naive Bayes baseline performed about 10% worse, but represents a very efficient algorithm. This creates interesting tensions within the trade-off between necessary accuracy and cost.

A.4.4 Modeling Threads only

Further, tests in modeling threads only resulted in comparable performance, though the real data set appears less imbalanced in this case.

A.5 Leveraging the Model as a Baseline

Lastly, having a Language Model that is well capable of classifying all platform text being trained across all communities, it may serve as a textual baseline w.r.t. actual contents and the optimization target (*here*: blocked). This enables analyzes for deeper insights that cannot be captured by any empirical approach on meta data.

However, as outlined within our discussion, our achieved model quality only allows for quite uncertain measures. Yet, assuming the model being equally mistaken across the whole population, results may still allow for conclusion. To provide a first peek into the possibilities if our model was performing better, we calculated blocking ratios of German communities according to real data and the model prediction in Table A.9.

Our first observation is that the order for both measures differs, however, we find evidence of correlation. I.e., the model assumes differences in blocking behavior w.r.t. the German average post. We leave further evaluation for future work, such as digging into high amplitude differences between the actual observed figures w.r.t. blocking versus model predictions; this investigation is of interest to either side: communities being stricter, or more tolerant.

A.6 Future Work

Future work should focus on improving the model by e.g., incorporating information about false negatives into the model decision, i.e., punishing false positives within the loss function. There are other approaches to this problem; possibly recent bigger LMs might also perform better than BERT-alikes under investigation.

A.7 Conclusions

We set out to characterize the role of blocked content in DE & SA. We observe that posts are blocked more often than posts; pictures also tend to be blocked more often, while blocked content is more prevalent in the German communities in general. Further, We show a correlation between cast votes per available post to the frequency of blocked contents. I.e., increased voting activities leads to more blocked contents. However, the amounts of moderated content remain equal between DE & SA irrespective of community size. Those posts

| Community | Predicted | Community | Actual |
|-----------------|-----------|----------------|--------|
| Potsdam | 0.2041 | Potsdam | 0.0324 |
| Hildesheim | 0.2125 | Vechta | 0.0359 |
| Koblenz | 0.2132 | Koblenz | 0.0365 |
| Vechta | 0.2143 | Lübeck | 0.0370 |
| Hochfeld | 0.2144 | Harburg | 0.0377 |
| Lübeck | 0.2145 | Krefeld | 0.0382 |
| Mönchengladbach | 0.2191 | Ratingen | 0.0382 |
| Castrop-Rauxel | 0.2201 | Gladbeck | 0.0384 |
| Oldenburg | 0.2207 | Hochfeld | 0.0388 |
| Chemnitz | 0.2219 | Wuppertal | 0.0407 |
| ... | ... | ... | ... |
| Hansaviertel | 0.2780 | Flensburg | 0.0663 |
| Kaiserslautern | 0.2789 | Passau | 0.0687 |
| Oberschleißheim | 0.2832 | Würselen | 0.0691 |
| Mitte | 0.2835 | Darmstadt | 0.0694 |
| Göttingen | 0.2864 | Niederrad | 0.0707 |
| Berlin | 0.2867 | Kaiserslautern | 0.0727 |
| Bayreuth | 0.2872 | Göttingen | 0.0735 |
| Gerlingen | 0.2937 | Ulm | 0.0745 |
| Niederrad | 0.2967 | Offenbach | 0.0801 |
| Offenbach | 0.3184 | Bayreuth | 0.0907 |
| All data | 0.2554 | All data | 0.0537 |

Table A.9: Model Inference blocked Ratios vs Actuals. Due to the imbalanced nature of our dataset, i.e., only few blocked posts compared to benign content, inference on actual data provides huge amounts in false classifications. Assuming that the model misclassifies consistently, inference results may still provide *limited* insights. Overall, we observe correlating amounts in blocked content across the various communities, indicating that the community-internal content-wise baselines w.r.t. bocking appear similar to some extent.

being flagged, get blocked with roughly 50% probability in DE, whereas the SA communities block 60% to 90%. Further, leveraging responses, we estimate that blocked posts remain for significantly shorter timeframes on the platform.

We next train a Language Model (LM) to identify blocked content. After selecting a suitable pre-trained model, GottBERT (a German only RoBERTa alike MLM), we continue fine-tuning this model to our downstream task varying optimizer, and batch size, amongst hyperparameters. We showcase the influence of dataset size and also observe that blocked content via moderation is an easier task for the Lm achieving better performance. We tested an additional triplet loss embedding approach without success. While presented model performance may appear being applicable, it cannot be used in practice due to a heavily imbalanced real world data. The amounts in false positives according to the model are too large. Nonetheless, we showcase how such a model could be used to gain further insights into independent communities below the meta level: establishing a qualitative textual baseline.

B User Lifetime Insights and Modeling

In this work, we predict the user lifetime within the anonymous and location-based social network Jodel in the Kingdom of Saudi Arabia. Jodel's location-based nature yields to the establishment of disjoint communities country-wide and enables for the first time the study of user lifetime in the case of a large set of disjoint communities. A user's lifetime is an important measurement for evaluating and steering customer bases as it can be leveraged to predict churn and possibly apply suitable methods to circumvent potential user losses. We train and test off the shelf machine learning techniques with 5-fold crossvalidation to predict user lifetime as a regression and classification problem; identifying the Random Forest to provide very strong results. Discussing model complexity and quality trade-offs, we also dive deep into a time-dependent feature subset analysis, which does not work very well; Easing up the classification problem into a binary decision (lifetime longer than timespan x) enables a practical lifetime predictor with very good performance. We identify implicit similarities across community models according to strong correlations in feature importance. A single countrywide model generalizes the problem and works equally well for any tested community; the overall model internally works similar to others also indicated by its feature importances.

B.1 Introduction

Every social networking platform depends on an active user-base. This user-base is threatened by *user churn*, which represents users leaving the platform. Retaining existing users is a core marketing strategy [Kotler, 2016] to mitigate potential losses focusing on positive user relationships via data and behavioral analysis. Loyal users (possibly inadvertently) advertise a product freely. More importantly, they tend being more profitable to a company. Beyond our field, customer lifetime value (CLV) denotes expected revenue over time in marketing and may be used to identify high-value and users at risk.

The actual churn prediction's goal is not only limited to predicting a churn event, but also the likelihood or time until a user might churn. Such individual churn probabilities allow for direct timed steering of single users (help, notifications, email) to improve individual retention. In the broader picture, this also allows for steering communities or customer-bases with the envisioned optimal features in mind: Maintaining a healthy user-base, which always is hoped to grow and converge into a well-mixed population. The prediction of user churn is a well studied data mining task. However, these works predominantly focus on predicting churn in a single community typically represented by one platform. The degree to which they generalize beyond a single user-base is thus an open question. New types of location-based networks enable the establishment of many disjoint communities within the same platform. This location-based property forming multiple independent user bases within the same platform constrains thus enables the study of this currently open question on how churn models generalize beyond a single community.

Structure [JH2]

-  [B.2: Related Work](#)
-  [B.3: User Lifetime and Churn \[JH4\]](#)
-  [B.4: Modeling User Lifetime](#)
-  [B.5: Features](#)
-  [B.6: Random Forest Implementation](#)

✂ B.8: Binary Lifetime Prediction

📄 B.9: Conclusions

B.1.1 Research Questions

We generally ask how long users are part of Jodel, i.e., we are interested in churn vs. retention. Furthermore, we ask whether we can model user churn events that would allow for an early classification, and hence possible counter measures.

B.1.2 Approach

We set out and characterize user lifetime empirically for the Jodel users in Saudi Arabia. By leveraging domain specific feature engineering, we train and test various of-the-shelf machine learning algorithms. We further provide ablation studies for time-dependent feature subsets and Random Forest hyperparameters. While results are promising, we finish our modeling efforts by providing an extensive evaluation of a practical binary non-/churn predictor. We lastly compare model internals proxied by feature importances.

B.1.3 Results

We study user lifetime in a location-based, anonymous social messaging application. Our goal hereby is creating prediction models for the lifetime of a user within a specified observation period. We leverage resulting models to implicitly show in-/equalities of these communities w.r.t. churn.

Among tested off the shelf machine learning algorithms, Random Forest provides the best results for predicting a users' expected lifetime, both in the case of a regression problem *and* a classification problem. Our models use two types of features: user and community. We observe the models to perform well for all communities.

Creating a single countrywide model generalizes the problem and works equally well for any tested community; this overall model internally works similar to others as indicated by its feature importances. We argue that model feature importances can provide feedback for empirical patterns pictured by the envisioned ideal community and may help to better understand reasons for users to stay or leave a platform.

At last, we use Random Forest to answer a supposedly simpler—and easier to answer—binary classification problem of practical relevance to network operators: Given an observation time period, will the users' lifetime be longer? This approach achieves even better prediction quality than any other presented classifier.

B.2 Related Work

User churn prediction has been a research topic for decades, yet with new emerging use-cases and technical advancement in data mining, machine learning and explainable AI, research on this topic has not halted by any means. We have seen various settings and applications within, e.g., telecommunication [Óskarsdóttir et al., 2020], social networks [Chen et al., 2015], online/video gaming [Runge et al., 2014, Chen et al., 2018], or online marketing [Chamberlain et al., 2017]. User lifetime has also been modelled as a survival analysis [África Periañez et al., 2016].

While user churn prediction often describes a binary classification, users' retention time might also be of interest. From a marketing perspective, user churn measures are typically weighted into an optimization target of a Customer Lifetime Value (CLV) according to, e.g., profit, yet relying on the same building block [Fader et al., 2005a, Chamberlain et al., 2017].

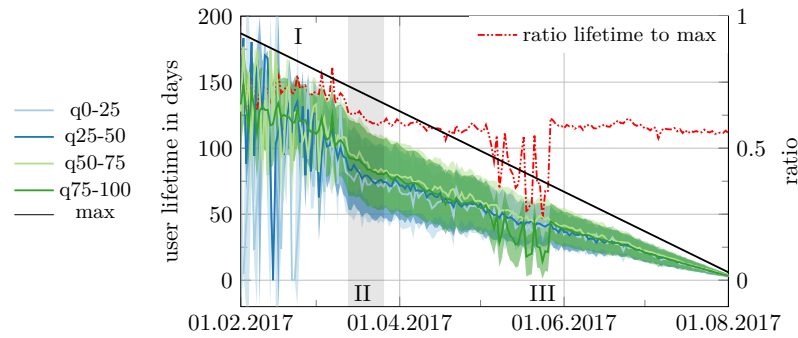


Figure B.1: Users' lifetime. This figure shows the average user lifetime and stddev by their registration date. While being noisy in Phase I, this metric stabilizes equally for all city subsets later, depicted by the lifetime ratio to the upper bound (end of observation) at a constant level above 55%.

Early work focussed less on today's off-the-shelf ML techniques. That is, statistical modelling and distribution fitting has shown significant success [Fader et al., 2005b, Dror et al., 2012, Chen et al., 2015]. Besides applying Markov models [Runge et al., 2014], others have evaluated, e.g., evolutionary [Wai-Ho Au et al., 2003] or relational [Fader et al., 2005b, Óskarsdóttir et al., 2020] minging techniques. Nonetheless, various classical ML approaches have shown promising to very strong results with, e.g., boosting [Lemmens and Croux, 2006], DTs and tree ensembles [Dror et al., 2012, Danescu-Niculescu-Mizil et al., 2013, Pudipeddi et al., 2014]. Neural networks have also been applied to the problem in various architectures: E.g., deep [Yang et al., 2018, Chamberlain et al., 2017] or convolutional [Chen et al., 2018] NNs. Explicit feature engineering for the data-driven methods requires an individual process to the very field of application. However, research suggests that social ties and graphs are an important information carrier [Dror et al., 2012, Yang et al., 2018]. Some research adds specific building blocks into their ML pipeline, such as user embeddings from browsing sessions [Chamberlain et al., 2017]. Although, e.g., RF or DT importances are often given, there is lack of its discussion, i.e., backfed empirical implications are seldomly drawn. Explainable AI is currently more or less tackled by, e.g., applying a user pre-clustering [Yang et al., 2018], however.

B.3 User Lifetime and Churn

The previous analysis showed that larger communities attract more users to participate reflected in the daily active users. Yet, we do not know how this translates into the time users stick to the system, which is why we now evaluate the time for how long users keep using the app. There is a variety of possibilities in the extremes of *a*) a cyclic renewal of the complete user base happens over and over again, or *z*) users are very committed to their community and participate over longer time periods.

To answer this question, we show the user lifetime within the system in Figure B.1. This figure shows the average lifetime and standard deviation of users w.r.t their registration date. While the x-axis denotes time, the y-axis marks the number of days a user is active (registration to last system interaction). Due to the end of our observation period, the active days are bounded (max). The ratio lifetime to max resembles the fraction of overall average user lifetime to this bound.

We make three observations: *i*) We observe that the user lifetime is quite high, but noisy in Phase I, while decreasing within the inception Phase II. Then, it stabilizes within Phase III indicated by the linear trend of the user lifetime. *ii*) The ratio of the overall user lifetime to the given observation bound indicates that on average more than 60% of the users keep using the app until the end of observation. *iii*) There is no qualitative difference between community sizes as although, there are huge differences in absolute numbers, the measured

user lifetime is rather identical—a similarity.

Takeaway *i)* With increasing community activity, the amount of daily active users also increase following a power-law. The amount of active users is in a steady state for Phase III as the communities do not differ qualitatively. *ii)* There is high user retention indicated by $> 60\%$ of the users keeping using the app until the end of our observation. Though user lifetime fluctuates, on average, it remains qualitatively similar throughout the city quantiles. We conclude that many users stick to the system, while there also happens a cyclic renewal of the user base for the remaining 40% users.

B.3.1 User Retention and Churn

Next, we study what makes users to continue using Jodel (retention) or to leave the app (churn). That is, we present a similarity across all communities to study if can we characterize long-time users from their first two days in the platform? Further, we take a closer look into differences of churned users between community sizes from a network perspective.

While the population of user interactions that have dropped out in various time frames between communities has no statistical significant difference, we find that 60% of the users keep using the app after registration, However, about 15% of dropped users do not interact with the platform at all. From our data, we cannot assess whether these users simply do not use the app, or whether they are lurkers who entirely consume content entirely *passive*.

B.3.1.1 User Retention

To study what differentiates users that kept using the app (retention) from users that dropped out (churn), we focus on their interactions with the app. We take users registered after April 1 (Phase III, nationwide establishment) and group them into three groups: *i)* users that were active only for two days, *ii)* only for a week, and *iii)* users that kept using Jodel for more than 30 days. For each group, we extracted the amount of interactions of each user on the Jodel platform within the first 24 hours after registration and determined the user's community set.

First, we compare the total number of system interactions (i.e., creating content and voting) between the different groups and communities. We observe that the user populations do *not* significantly differ (tested with a t-test).

Second, we analyze differences w.r.t. different interaction types (i.e., posting, replying, up- and downvoting, and flagging) in isolation, we arrive at the very same result (again with a t-test): there is no significant difference between our defined groups—a similarity across all community sets.

B.3.1.2 User Churn

There is no obvious difference in the populations of user interactions w.r.t. retention. We next flip the question and study why and how users churn (leaving the app). We begin by studying users that dropped out of the system, e.g., by losing interest. To shed light on this group of users, we analyze behavioral metrics of users who did not interact with the system at all or their lifetime was limited to only at most 24 hours.

First, in Figure B.2, we provide a high-level view on the dropped user base via a Sankey diagram (describing qualitative flows). Most dropped users (about 51.6%, 174k) have no interaction with the system at all. These users installed the application and opened it at least once to trigger a system registration, but did not actively interact by posting, replying, or voting. From our data, we cannot tell whether these users did not use the application at all or used it only by means of passively consuming content (i.e., browsing over and reading

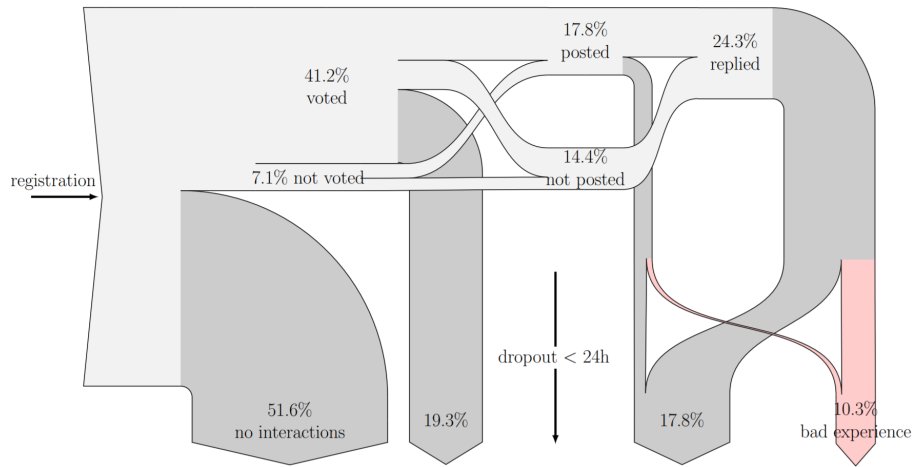


Figure B.2: Qualitative view of overall lost users within 24h. This Sankey diagram shows actions of all users before dropping out of the system. Most users do not interact at all (no interaction), while the others at least vote once (voted) before dropping out. A *bad experience* means getting downvotes on posts.

| community quantile | no interactions | voted | posted | replied |
|--------------------|-----------------|-----------|-----------|-----------|
| q0-25 | 50.1±5.6% | 38.5±7.0% | 18.7±3.6% | 22.5±3.7% |
| q25-50 | 43.7±4.7% | 46.2±5.4% | 19.8±3.3% | 27.2±4.1% |
| q50-75 | 42.1±5.0% | 48.9±5.3% | 20.3±2.7% | 28.1±4.3% |
| q75-100 | 41.8±11.0% | 50.1±9.6% | 21.2±4.1% | 29.5±5.8% |

Table B.1: Qualitative view of users within 24h by community sizes. This table shows the average amount (and stddev) of performed interactions of users that dropped out within 24h. There is a clear tendency of users participating more in larger communities.

posts). Most other dropped users at least voted once, while replying is more popular than posting among them. Still, about 19.3% of the dropped users only voted, while the others created content. Out of these content creators, we counted the users having a *bad experience*, that is getting downvotes to one of their posts or even getting blocked by moderation—accounting for 10.3% of all dropped users.

Second, we split the users into community subsets and analyze the non-/presence of possible interaction types. We provide the results in Table B.1. It provides information of the average amount (and standard deviation) of how many users of each community subset have either not interacted with the platform at all, voted, posted or replied. For all community subsets, we observe a difference in these figures. That is, in smaller communities, more people have not interacted at all and other interaction types are less common than in larger communities. On the contrary, larger communities (in terms of interaction volume) trigger more users to interact. As we observe positive trends on average across community sizes. However, given overlapping standard deviations across averages arguably represents rather similar behavior.

Takeaway *i)* All communities show similar behavior w.r.t. user retention, an *invariant*. That is, all communities behave similar by their interaction volume and interaction types subject to users lifetime. *ii)* 27.6% of all registered users drop out within 24 hours. Although about 50% of the users interacted with their community by voting or posting, half of them created content at least once (25%) of which only about 10% actually make a bad experience from an empirical point of view. *Invariant* to community size, churned users behave similar before leaving the platform.

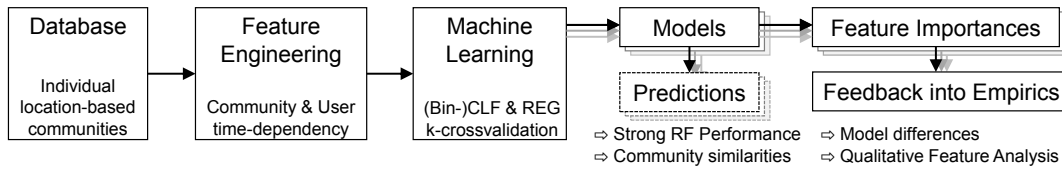


Figure B.3: Machine Learning and Evaluation Pipeline. We use data from the anonymous location-based messaging app Jodel in KSA. We compute various types of features to be fed into various ML tasks leveraging crossvalidation. Our resulting RF models provide string prediction performance for Regression and Classification. Through model similarities, we identify community similarities; Further, we analyze resulting models w.r.t. their feature importances to be fed back into empirical analysis.

B.4 Modeling User Lifetime

To study and model user lifetime, we first describe our dataset and then define how we assess a user’s lifetime. The goal of our work is to automatically detect and classify the user lifetime in the anonymous and local communities within Jodel.

We employ a generic ML pipeline as shown in Figure B.3. Using individual community data, we do domain specific feature engineering, which we use to derive machine learning models applying crossvalidation. Note that we apply data normalization after the test/train split to prevent information leakage.

Dataset limitations. Our dataset only includes the users’ *active* interactions with the system, i.e., registering, creating posts, replying, or voting. Thus, we cannot infer when or how much a user only *passively* participates—lurkers—who only consume content. Further, the vote interactions are always mapped to the date and geoposition of the respective content creation. This prevents us from making detailed analyses depending on the voting time or place. However, due to the vivid usage of the application within larger communities (multiple posts/replies per minute), we generally consider votes to be executed on the same day as their respective content. Especially since posts are only accessible via the three different feeds, where they will only stay for a very limited time, casting votes to content long after creation is usually not possible.

B.4.1 Goal: User Lifetime Prediction

Definition. We define the *lifetime* of a user as the time between the first (automatic account creation) and last system interaction (i.e., posting or voting) in minutes. Note that we can only define the lifetime by system interactions of a user since our dataset does not include passive activities (i.e., only reading). In prior work, churn, i.e., users leaving the system, is often defined as the end of a user’s lifetime. The lifetime enables us to partition users into timespan-dependent activity classes that we later predict, e.g., users that only used Jodel for a short amount of time, or longer.

| class | lifetime | #users | fraction |
|----------|-------------------|--------|----------|
| 1 | 0 . . . 1 days | 135k | 13.3% |
| 2 | 1 . . . 7 days | 123k | 12.1% |
| 3 | 7 . . . 14 days | 75k | 7.4% |
| 4 | 0.5 . . . 1 month | 124k | 12.3% |
| 5 | 1 . . . 3 months | 268k | 26.4% |
| 6 | > 3 months | 288k | 28.4% |
| Σ | | 1 012k | 100.0% |

Table B.2: Defined six churn classes. We subdivide the user population by their active time.

User lifetime distribution. We compute the lifetime for every user in our dataset and group users into six classes shown in Table B.2. There is a broad range of different user lifetimes, ranging from users that only use Jodel for less than a single day and users that stick with the platform for more than 3 months. A user's lifetime can end in two cases: First, the end of the observation period. Second, users that stop using the application (i.e., churn). As in a practical setting the observation period is always finite, prior work *approximates* the churn potential by using a threshold (e.g., no activity within the last n days, where n often is derived empirically); the finiteness naturally introduces a skew towards shorter time periods. If we apply such a threshold of one week $n = 7$ (i.e., users are regarded as churners if there is no activity within the one week threshold margin towards the end of observation), 61% of the users will be defined as churners. We remark that churn prediction is an inherently hard problem since users could become active again after the threshold. Instead of churn, we predict user lifetime, i.e., the chance of a user to use the app for at least k days or a discretized timespan.

Goal. The goal of this section is to predict the lifetime of a user within a specified observation period. Social network operators can use the resulting models as an online algorithm to predict the likelihood of a user to stick with the platform; furthermore, they allow for studying user behavior.

B.5 Features

To predict the lifetime of a user by using a data-driven ML model, we derive features from *i*) the user itself and *ii*) her community.

Engineering Subsets. We introduce two different feature classes to represent the user and her environment: *i*) **User related features:** e.g., user registration event information, down-/upvoting and post/reply behavior of a specific user. *ii*) **Community related features:** e.g., posts/replies, up-/downvotes, average post response time of the users home community.

B.5.1 Users home community

Since Jodel establishes communities relative to the users' locations, users can participate within different communities when sufficiently changing their geographic position. For a stable model, we derive the community features from the users' home community, which we define as the city location with a user's most interactions. For 87% of the users, this home community represents the city in which they initially registered. We use this attribute to determine a user's city throughout this work.

B.5.2 Capturing time

While our features up to now do not catch any time-dependent information reflecting a user's lifecycle, we add timing insights by duplicating the features with time-period bounds (1 day, 2 days, 3 days, 1 week, 2 weeks, 1 month, 3 months, >3 months). We end up with 66 (29) time-(in)dependent features. Noteworthy, none of typical scaling (e.g., std, min-max) or balancing (e.g., SMOTE, random over/undersampling) techniques improved any of our results significantly. Further, we used mean imputing as it provided best overall results.

Imputing. While the feature sets may be incomplete (e.g., including NULL values due to a user not having any data beyond one month), we need to define how to handle them.

| subset | time-independent # | time-dependent # |
|-----------|--------------------|------------------|
| Generic | 5 | - |
| User | 24 | 49 |
| Community | - | 17 |

Table B.3: Feature Subgroup Counts.

The easiest solution would be dropping this user from our ML approach, being rather undesirable. Imputing of data to such missing fields allows us to use all available information. To not introduce a bias into the used data, typical imputing strategies are either mean or median.

Scaling. Engineered features usually are simple numeric values. Some ML methods may not work well with arbitrary ranges, which is solved by applying scaling. With scaling, each feature series is converted into a range of values to ease using them. Typical scaling approaches resemble MinMax-Scaling (relative value adjustment between the series' min and max values onto the interval 0..1), or Standard-Scaling (remove mean from series and norm its values by the standard deviation).

Balancing. Within any dataset, it may be that the population within the optimization target is not equally distributed. E.g., a binary classification problem may have largely different amounts of representatives for a certain class. In some instances, artificially balancing such imbalanced data may lead to better ML performance. There are lots of methods to tackle such problems; However, random over- and under-sampling are the most generic concepts, which try to create and drop data according to specific rules to create a balanced dataset.

Summary. We have gathered wide subsets of features from our dataset covering both, the community and the user, while also maintaining a view on temporal changes within behavior at varying time-periods. Further, we introduced common best-practice techniques for data preparation for any upcoming ML processing.

B.6 Random Forest Implementation

Within this section, we discuss our machine learning approach for a user lifetime predictor. By applying grid searches, after showing the Random Forest providing best results in comparison to other off the shelf ML methods, we dive deeper into the results of individual community models and their best parameters obtained. Then, we show how the prediction quality depends on model complexity and how different time-dependent feature subsets determine performance. Moreover, we look into model generalization and the impact of the amount of input data. Eventually, we derive implicit model (and as such community) similarities.

For any model, we apply grid-searches leveraging a random 5-fold crossvalidation approach for both, formulating a problem for *a*) regression of lifetime in minutes (REG), and *b*) classification into the previously defined classes (CLF). Our main evaluation metrics are: *REG*: The R^2 score measures how equal real and predicted populations are, and *CLF*: the F1 score describes the harmonic mean of precision and recall; both providing an overall picture.

Since the property of establishing communities relative to a user's position, our data set contains a large set of city-level communities throughout the country. To focus our discussion, in comparison to an all data *Country* model, we chose a subset of five cities varying in their size by the amount of users to create distinct prediction models: Riyadh,

Jeddah, Mecca, Al Bahah, and Al Jafr (large to small). Our implementation uses Python scikit-learn off the shelf functionality.

B.6.1 ML Algorithm Selection

First, we ran a grid search for all data over a set of hyperparameters for standard ML methods to obtain ballpark numbers. These grid searches use a *mean* imputing strategy while not incorporating any scaling or balancing. Our used algorithms are Random Forest (RF), Decision Tree (DT), Multi Perceptron (MLP), AdaBoost (AdaB), K-Nearest Neighbors (KNN), and Stochastic Gradient Descent (SGD).

Table B.4 provides an overview of a baseline in comparison to both problem formulations: regression and classification. The baseline is obtained by using the scikit-learn dummy regressor predicting the mean user lifetime in minutes, and the most frequent class for classification. While it yields unusable results for the regression task, an imbalanced dataset naturally provides better figures for classification. Due to multiple crossvalidation runs, we also provide standard deviation figures.

| Algorithm | REG: $R^2 \pm \text{stddev}$ | CLF: F1 $\pm \text{stddev}$ |
|-----------|------------------------------|-----------------------------|
| baseline | -0.0000 ± 0.0000 | 0.1672 ± 0.0007 |
| RF | 0.9822 ± 0.0004 | 0.9668 ± 0.0005 |
| DT | 0.9580 ± 0.0009 | 0.8049 ± 0.0095 |
| MLP | 0.9668 ± 0.0036 | 0.6768 ± 0.0304 |
| AdaB | 0.7654 ± 0.0012 | 0.6720 ± 0.0093 |
| KNN | 0.6764 ± 0.0013 | 0.5077 ± 0.0006 |
| SGD | 0.3422 ± 0.0551 | 0.1686 ± 0.0191 |

Table B.4: Off the shelf ML algorithm results using all data applying mean imputing; no scaling, no balancing. While the RF performs best, DT and MLP achieve similar regression performance falling short in classification.

We observe that both, the regression and classification baseline are easily outperformed by any algorithm except for CLF with SGD and KNN. The best performing algorithm always is a rather complex Random Forest. However, for regression, DT and MLP also perform quite good. The results almost do not fluctuate across multiple crossvalidation instances at all.

Findings The best performing ML algorithm is the Random Forest with very strong regression $R^2 \approx 0.97$ and classification F1 ≈ 0.99 scores. Thus, we will from now on focus on the RF algorithm. Nonetheless, most others also outperform the baseline significantly.

B.6.2 Independent Communities

As we have now determined the best-working algorithm for all data to be Random Forest and its parameters for our regression and classification predictor, we take a closer look into performance of specific independent community models. That is, does the prediction performance differ by community? This evaluation is enabled by the location-based nature of Jodel which allows us to compare independent user bases subject to the same platform constraints.

In Table B.5, we show the best results of each Random Forest grid search instance for our selected cities and an all data country model. The communities are sorted by the amount of users within their community in descending order. We selected these particular examples due to their different amounts of users to cover a wide range from large to small. The R^2 score describes the crossvalidation result for the regression problem, whereas F1 describes the classification results; additionally, we provide the standard deviation across folds.

| Community | REG: $R^2 \pm \text{stddev}$ | CLF: F1 $\pm \text{stddev}$ | #users |
|-----------|------------------------------|-----------------------------|--------|
| Country | 0.9822 ± 0.0004 | 0.9668 ± 0.0005 | 1 012k |
| Riyadh | 0.9728 ± 0.0013 | 0.9531 ± 0.0006 | 284k |
| Jeddah | 0.9667 ± 0.0016 | 0.9372 ± 0.0008 | 101k |
| Mecca | 0.9551 ± 0.0035 | 0.9185 ± 0.0039 | 45k |
| Al Bahah | 0.9457 ± 0.0032 | 0.8752 ± 0.0093 | 11k |
| Al Jafr | 0.8219 ± 0.1115 | 0.5807 ± 0.0594 | 174 |

Table B.5: RF classification and regression results for the Country model and selected individual communities. Results are consistent and stronger for larger communities, except for Al Jafr due to small amount of data.

Generally, all predictions perform well with R^2 scores above 0.94 for all cities except Al Jafr; The F1 score within the classification instances is above 0.87 for most cases. There are negligible fluctuations across folds as seen by the low standard deviation for all cities again except Al Jafr due to the very few data points (only 174 total users); yet, the regression still works surprisingly well, whereas the classification falls short in achieved quality. We observe that all predictors work better on larger communities, hence more data.

Further, besides our best performing Random Forests mostly use larger amounts of estimators, we notice that they are complex being rather deep and leveraging all features (n). We will discuss tree complexity versus performance in Subsection B.6.3.

Findings The overall predictor performance for the regression and classification task is very good for all analyzed independent communities except Al Jafr. Resulting models are quite complex in terms of tree depths, used features and estimators. Best-performing classifications tend to require less complex model instances than the regression.

B.6.3 Predictor Sweet Spot

As we have presented best performing results from the grid search for the Random Forest across our selected communities, we now want to shed light on the relationship between model complexity and prediction quality for two reasons: 1) An overly complex model might tend to overfit our data, 2) Less complex models are usually preferred due to less computation times for both, fitting and application.

We therefore investigated the relationship between used features, estimators, tree depth and resulting model quality for both, CLF and REG. For both problem formulations, model complexities are qualitatively very similar to their achieved performance—expectedly, more complex models perform better.

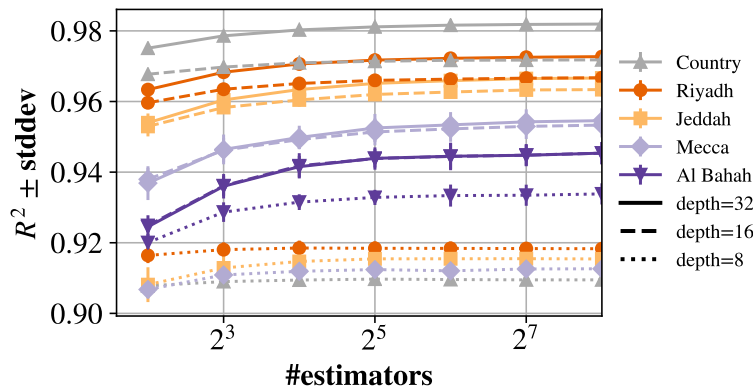


Figure B.4: RF regression performance vs. model complexity. More complex models provide stronger performance with diminishing returns in depth and especially the amount of used estimators.

Exemplary, we show the relation between used estimators, tree depth and quality for

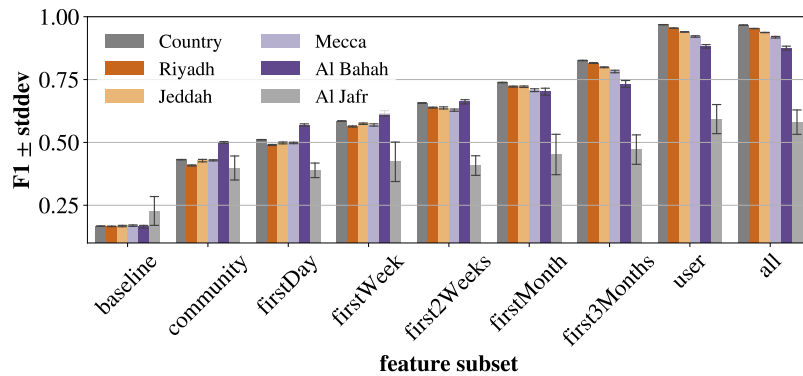


Figure B.5: RF Classification feature subset performance results. With increasing observation time per user, the results improve significantly. Overall, the country model works best, while most others achieve similar performance.

the regression model using all features in Figure B.4, which ultimately allows us to define a sweet spot. The logarithmic x-axis denotes the amount of used estimators (tree instances of the ensemble), whereas the y-axis shows the resulting R^2 score and standard deviation as error bars. There are three series for each city for a tree depth of 8, 16, and 32, respectively; we removed Al Jafr to increase readability.

We observe that increasing the tree depth substantially increases model quality. However, the improvement from a depth of 8 in comparison to 16 is by far larger than the change from 16 to 32. For the smallest community shown, Al Bahah, increasing the tree depth above 16 does not improve performance. While 32 estimators already yield very good results, the quality increase is asymptotically bounded, i.e., there are diminishing returns.

Findings. Although our grid search shows best results with rather complex Random Forest models (cf. Section B.6.2), our in-depth analysis of hyperparameters vs. quality reveals that only few estimators with mediocre tree depths already yield very good results with diminishing returns of increasing model complexity.

B.6.4 Feature Subset Analysis

While our prediction for regression and classification works well, we next want to determine the impact of different feature subsets. That is, which feature subset provides best results, or even better results than using all features? We conduct grid searches for the Random Forest across community and user features for all data (Country) and each of our selected independent communities.

Furthermore, from a practical implementation standpoint, one might think about leveraging a sliding window approach over a single day, a week or longer as model input. Such an approach makes a model more time-invariant. However, limited knowledge may seriously impact the model performance—as it is expected to degrade.

Thus, we show the impact of our feature subsets within Figure B.5 on the classification example. Note that the subset impact is quite similar for the regression problem (not shown). The x-axis describes used feature subsets: community and time-dependent features being cumulative: E.g., the *firstWeek* subset also includes features for the shorter time periods of *firstDay* and *first3Days*. The y-axis denotes the model quality via the F1 metric and the standard deviation across folds.

First of all, we observe that the community features alone provide worst results, but are quite similar to only looking into user data of her individual first days at scores ranging from $\approx 0.4 \dots 0.6$. By increasing the observation window up to 3 months, model quality increases drastically to F1 scores > 0.8 for most cities. Only relying on user features provides

similar results to using all features; the community features by itself only have a negligible impact on prediction quality. To be clear, this does not imply that the community only has negligible impact on user lifetime or user experience.

Noteworthy, predictions for larger communities tend to be better than for e.g., the Al Jafr community, always being off presumably due to its very few users (only 174). We cannot explain why the predictions for Al Bahah perform better for short time frames.

Findings Although our prediction models have proven very good performance, taking a practical stance by only using timely-windowed features depending on the users' active time reveals that classification and regression (not shown) quality deteriorates for real-world use-cases.

B.6.5 Generalization

We next study how well our models predict the lifetime from other communities to investigate whether there is a model—possibly a community, or the Country-model—providing well-suited allround prediction quality.

Within Figure B.6, using the best-performing model from our previous grid-search (cf. Section B.6.2) each, we provide cross-application (community-model \times predictions-for-community) classification scorings of our different community and the Country model(s). The x-axis describes the used model instance, whereas the y-axis denotes the predictor input-dataset. We provide the macro F1 scores for each combination, colored on the z-axis. Note that we added the same-same community model/application F1 scores from previous results as a baseline (diagonal upper left to bottom right).

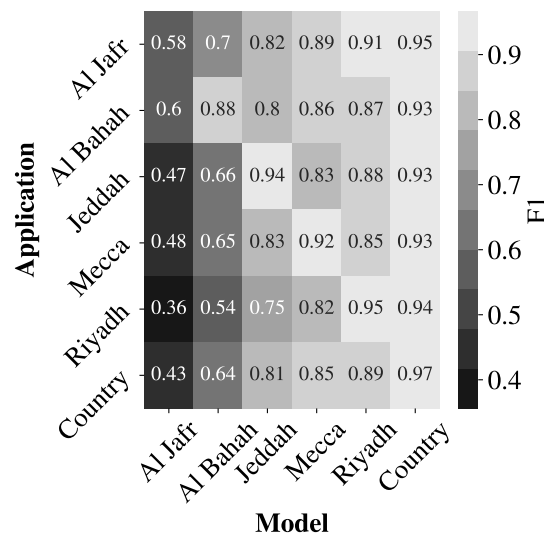


Figure B.6: RF classification model cross application results. We used each created model and performed a prediction for every other dataset. The diagonal same-same instances depict earlier prediction results as a comparison. The country model works best across the board.

Focusing on the *Country*-model first, *i*) we observe that it provides strong generalized performance with F1 scores ≥ 0.93 throughout any community (rightmost column). Taking a closer look into the community cross applications, we find: *ii*) Most individual community models perform very well on their own input dataset; other models significantly improve the Al Jafr community prediction scores. *iii*) The overall best-working community model is Riyadh, falling short in prediction quality for Mecca. The community models from Mecca (Jeddah) still delivers acceptable prediction quality across the board with F1 scores ≥ 0.82 (≥ 0.75). However, the models for the smaller communities do not perform well in a generalized setting. *iii*) By leveraging the model scores as a proxy for community similarly

| #random instance sample size | alike | F1 \pm std | | Δ |
|---------------------------------|----------|---------------------|---------------------|----------|
| | | country | community | |
| 0 | Country | 0.9668 \pm 0.0005 | - | - |
| 1 | Riyadh | 0.9496 \pm 0.0016 | 0.9531 \pm 0.0006 | -0.0035 |
| 2 | Jeddah | 0.9312 \pm 0.0026 | 0.9372 \pm 0.0008 | -0.0060 |
| 3 | Mecca | 0.9136 \pm 0.0024 | 0.9185 \pm 0.0039 | -0.0049 |
| 4 | Al Bahah | 0.8773 \pm 0.0063 | 0.8752 \pm 0.0093 | 0.0021 |
| 5 | Al Jafr | 0.6794 \pm 0.0691 | 0.5807 \pm 0.0594 | 0.0988 |

Table B.6: Data Input Size Ablation Study. RF classification Country model with limited input data corresponding to other city sizes. The country model achieves similar performance to individual models when restricting the input data size. I.e., by reducing the amount of data, performance deteriorates.

w.r.t. user lifetime, we identify that due to shown options for generalization, user lifetime is similar to some extent across the analyzed communities, yet through the models only being implicitly defined. *iv*) For the regression use-case, we find the Country model strictly outperforming all others with an R^2 score of 0.98 in all communities; besides, only the Mecca community model provides consistent strong results applied to other communities at R^2 scores ≥ 0.83 (not shown).

This generalization reveals that the independent communities can be captured well in a single model and behave similar to some extent w.r.t. user lifetime. Note however, the *Country*-model may be skewed in favor of larger communities.

Findings The overall Country model performs very well throughout any tested community for both, regression and classification, with improvements for smaller communities and (for classification) slight deteriorations for the clique of larger cities (Riyadh, Jeddah, Mecca). Yet this model works well and might be used for the whole dataset as a unique predictor, retraining is computationally heavier than selected individual models due to its size.

B.6.6 Country Model in Detail

Our evaluation and cross application showed that the *Country*-model provides all-round performance for regression and classification, while also improving predictions for communities with comparably fewer users. But why is that? Does simply the amount of available data improve the model, or does the country model represent a better cut through the population?

To answer this question, we randomly downscaled all data to the reference values of our other selected communities and ran grid-searches for these new sampled Country models. We present our results in Table B.6 for the classification problem. The alike community column depicts the reference sample size, whereas the F1 country column denotes the model's classification quality. Further, we add the individual city models as an expected upper-bound baseline comparison (column F1 city).

Our evaluation reveals that the model performance remains very strong, but worsens with less data. However, given a statistical significance due to the sheer amount of input data, the observed delta to the individual model arguable remains within margin of error; except for Al Jafr most probably due to its very sparse data.

Findings We observe that the country model performance is similar to the individual models provided it uses the same amount of input data. Still, the country model deals better throughout all communities than any other individual model.

B.6.7 Feature Relevance per Community

We observe the Random Forest to provide good prediction results. Further, we have seen that the communities behave similar w.r.t. user churn to some extent. However, we are missing information which metrics were predominantly used by the models. Do they rely on the very same features or have they learnt differently?

There are different well-known feature importance predictors, such as ReliefF or RFECV. Here, we focus on RF Feature Importance (RFFI), Gini importance/Mean Decrease in Accuracy, depicting the percentage ranking within its decision making. By cross correlating the RFFIs with the ranked Spearman's, we next want to figure out if importance score line up across the different models. We show the results for the classification in Figure B.7, showing a community on both axes, of which the correlation between the cities' importance vector is represented textually on the also colored z-axis. Note that the overall picture remains the same for Regression (not shown).

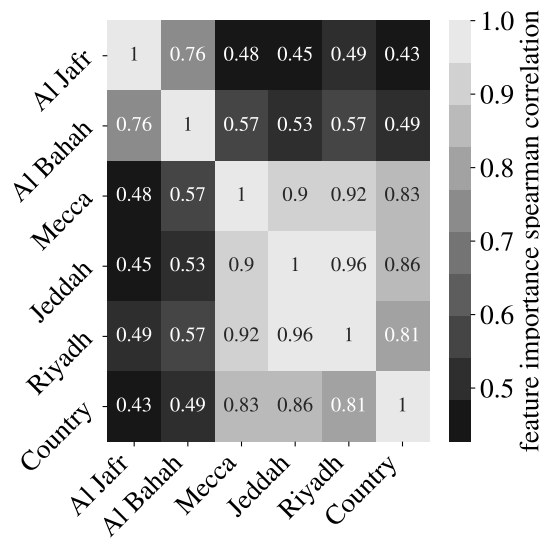


Figure B.7: RF classification features importance spearman correlation. We find that most models strongly correlate in feature importances indicating similar model internals.

We observe that the clique of Mecca-Jeddah-Riyadh line up quite well, whereas the smaller cities fall off. Interestingly, the Country model feature importances also correlate to beforementioned clique indicating that these models work similarly. Digging deeper into the Country model by correlating importances of the sampled versions (not shown), we find that all of these models, despite using less data, are strongly correlating in feature importance and thus, share similar internals; except alike-Al Jafr falling short.

Findings The rather strong correlation among the models for larger cities strengthens our hypothesis of similarly behaving communities due to their model similarities. This also holds true for the overall Country model and its sampled versions.

B.7 Future Work: Empirical Lifetime Study

Having built well-performing predictors for user lifetime and having seen that model feature importances often correlate, these importances figure an important signal of usage within the respective model, which deserve to be fed back into a thorough empirical analysis to better understand model internals.

We argue that used feature values of users subdivided into their lifetime classes represent a projection of its population w.r.t. user lifetime. I.e., by partitioning a feature's population by the optimization target (active time), we discretize the community state available to each

user. Implicitly driven through lifetime predicting RFs, a community state for important single features w.r.t. lifetime might then be given by the distribution in a certain time-slice—which in turn depict qualitative differences between these partitions.

Most likely due to the high number of samples, by applying a Mann-Whitney-U test, we find strong statistical evidence for non-equal distributions in between time slices across all models and most features, which needs to be taken with a grain of salt and leaves us with a lots of data for further qualitative analysis.

However, while punctual information is not sufficient to identify changes, we are more interested in qualitative changes over time within most important features. To provide empirical insights, we first removed upper and lower 1% outliers and then applied MinMax scaling (onto 0 to 1). Then we calculated the set of quantiles 10%, 20%, . . . , 90% and plotted the resulting kernel for each feature over the time subsets (1 day, 3 days, 1 week, 2 week, 1 month, and above) in Table B.7, while the solid line denotes the median. That is, the (second) outmost gray area denotes the population between the quantiles of 10% (20%) and 90% (80%). To better capture qualitative changes, we apply a logarithmic y scale.

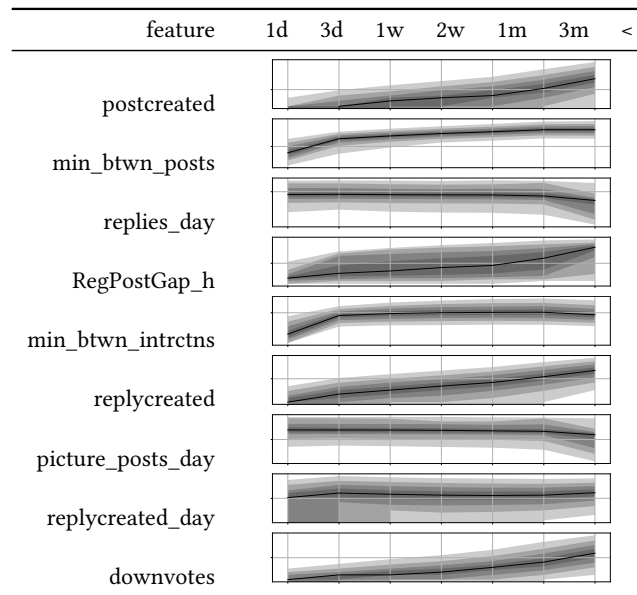


Table B.7: Qualitative population of most important features (country model) subdivided into user lifetime sets. The areas denote quantiles from 10-90%, 20-80% etc. whereas the solid line depicts the median. On a logarithmic y-axis, we observe drastic variations within the metrics over a user's lifetime.

Given new challenges like for some metrics, we can observe clear trends, while others draw an unclear picture, or any counting feature depending on time, we leave next steps for future work. Nonetheless, feature candidates with clear trends might be a valuable object for future research, i.e., learning from model behavior; and ultimately e.g., testing hypotheses with e.g., synthetic tests [Amjad et al., 2019].

B.8 Binary Lifetime Prediction

Although our prediction works quite well overall with classification F1 scores up to 0.95 for Riyadh, having only timely limited user information deteriorates prediction quality significantly (cf. Section B.6.4). In practice, e.g., a network operator usually only asks whether a user is likely to churn in near future. This allows us to reformulate our problem into a supposedly simpler—and easier to answer—binary classification problem: Given an observation time period, will a user's lifetime be longer?

Thus, we ran grid searches for binary predictors for every lifetime class in Table B.2 and all selected communities likewise to the feature subset analysis, which allows us to

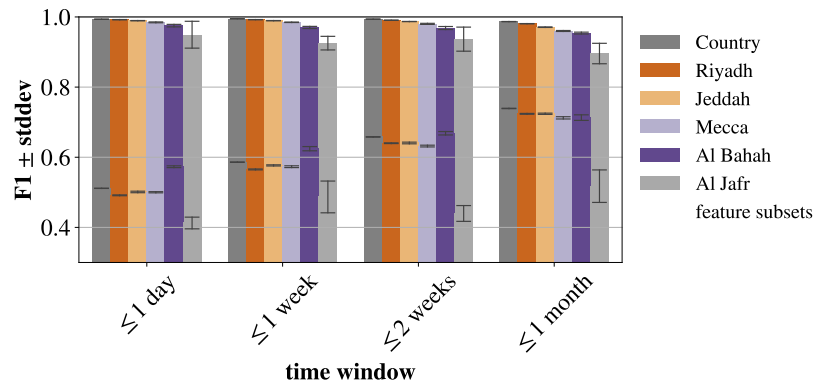


Figure B.8: RF Binary Classification performance results. Regardless of observation time, the binary lifetime prediction works exceedingly well. All time-dependent feature subset comparisons perform by far worse.

generate predictions according to our chosen time-periods of the feature subsets.

We show the results of the best binary classification models in Figure B.8. The x-axis describes the used time window, whereas the y-axis denotes the F1 score. The model results for each independent community are plotted as bars with whiskers indicating the standard deviation across folds. For comparison, we added feature subset analysis results (hatched).

We first observe that the binary classification model works quite well with F1 scores above 0.95 for almost all communities across the time window. The performance delta to the cumulative time-dependent feature subset analysis models only becomes smaller for longer time frames as those models improve. With regards to model complexity, we observe similar results as seen for the other models (cf. B.6.3): That is, e.g., a tree depth of 16 performs far better than 8, but there are diminishing returns beyond this depth and more than 16 estimators. In summary, the binary practical classifier outperforms *any* other presented classification model.

Findings In practice, it is desired to remove complete time-dependency. Thus, the overall models may not reflect real-world performance in windowed feature use-cases. We however showed that classifications on time-dependent window subsets do not perform well. To ease up this problem, we simplified the task to a binary classifier predicting a users' lifetime. This approach achieves better prediction quality than any other presented classifier.

B.9 Conclusions

In this section, we analyze and predict the lifetime of a user in Jodel, a mobile only location-based messaging app. Our results show that Random Forests models provide good prediction results for both, regression and classification tasks across a selection of individual communities of varying sizes throughout the Kingdom of Saudi Arabia. When making models invariant to total observation time, i.e., only relying on timely limited feature sets, prediction results deteriorate substantially. This can be solved by using Random Forrest models to predict a simpler binary classification problem of practical relevance to network operators: Given an observation time period, will the users' lifetime be longer? This approach achieves even better prediction quality than any other presented classifier.

The location-based nature of Jodel yields the creation of disjoint communities throughout the country. When training a single model to the entire data set (i.e., a country-wide model), this model performs well compared to individual community models at similar amounts of input data. That is, while the individual communities are disjoint, users share behavioral pattern. This is further highlighted by the fact that the RF feature importances correlate between most individual and the country model(s). We therefore conclude the

models' internal decision-making processes being similar and hence, also communities sharing alike behavior w.r.t. user lifetime.

Eventually, we argue that the feature importance provides strong hints about model internals and are a good starting point to be fed back into empirical analyses, which we leave for future work.

C Excursus: Data-Driven Long Term Gaming and QoE

A subset of massively multiplayer online games (MMOG) feature long-term game rounds in which players interact for months or even years. The player experience of such long-term games cannot be entirely captured by current study methods, in particular not at scale assessing large player populations. To address this challenge, we posit that long-term, round based games such as Tribal Wars (browser-based) enable a data-driven perspective on long-term game dynamics and experience. In a preliminary study, we monitor and characterize the entire longitudinal game state of a Tribal Wars round that was played by 16k players for 1.5 years, enabling us to investigate behavioral patterns of all active players. We identify features that capture the in-game success and relate to the player experience. We show that only successful players keep up playing.

C.1 Introduction

While a user's decision to leave a platform for good is a final signal of user experience, the operator usually has interest to uncover reasons. As Jodel does not set any time restrictions for participation, research suggests high degrees of ephemerality. We next present a blueprint on a prime example MMOG implementing long-term round structures spanning years to measure **★ C) Long-Term Gaming Quality of Experience** laying a data-driven foundation to the quality of experience research community for data-driven (pre-)studies that serve as a key enabler boosting the hypothesizing processes that then needs further validation with classical methods—for a better understanding of user experience, including reasons for churn.

Current Gaming Quality of Experience evaluations assume short test stimuli for interactive tests, e.g., in the order of 90-120 seconds for short or 10-15 minutes for long interactive stimuli as recommended by P.809 [ITU-T Recommendation, 2018]. Stimuli of this length are sufficient to assess the player experience for *most* games. The implicit assumption is that the stimuli duration is short enough to avoid fatigue and render interactive tests feasible, yet long enough to represent a typical game situation that enables assessing quality features (e.g., evaluating the effect of interaction delays usually requires short tasks only).

Yet, games exist where single rounds span multiple months or even years. Typical examples are MMOG games such as World of Warcraft or Tribal Wars (TW). While e.g., World of Warcraft implements an endless game world without defined start or end, other MMOG's such as TW establish game rounds having a clearly defined start date and round-ending goal or maximum time. In TW, multiple rounds (game worlds) exist in parallel featuring different game configurations and goals (e.g., casual, regular). Successfully finishing a round requires a player's long-term commitment over years on a regular basis. We posit that this well-defined round structure and the surprising commitment of many players render games such as TW prime candidates to study long-term gaming dynamics and user experience.

While many quality features describing player experience can still be assessed in short interactive tests, they cannot be used to capture the player experience over an entire round that lasts for years. As a result, it remains unclear how the player experience can be assessed longitudinal *at scale*. Interactive tests at this widened timeframe are infeasible and questionnaire-based surveys do not scale to large populations of unknown users. Consequently, many related aspects such as the long-term integration of single or cumulative bad usage experiences remain unknown.

Structure [JH3]

 [C.2: Related Work](#)

 [C.3: Long-Term Gaming and Data Collection](#)

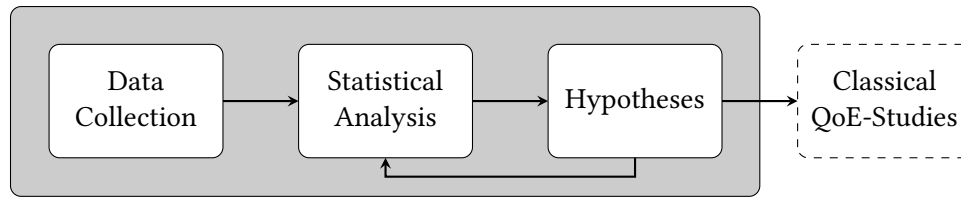


Figure C.1: Data-Driven Long-term QoE and Gaming Dynamics Approach.

♥ C.4: Long-Term Game Dynamics Analysis

🕒 C.4.1: Player Lifetime

⚔️ C.4.2: Inevitable Tension in Expansion

😊 C.4.3: Player Experience Indicators

📦 C.5: Conclusions and Future Directions

C.1.1 Research Questions

We make the case for a longitudinal data-driven perspective to gain insights into long-term game dynamics. By assessing the entire game state, this approach evaluates behavioral patterns of *all active players*—not only few selected players under study. We posit that this perspective is a key enabler in the realm of widened timeframes. Our first step analyzing game metrics is as crucial as attractive for studying games with single round durations lasting months or even years. We showcase argue that this approach can identify behavioral patterns and helps to hypothesize about circumstances impacting the player experience. In a second step (beyond this section), identified hypotheses can then be examined in dedicated tests following traditional study designs, e.g., P.809.

C.1.2 Approach

To enable this study, we have continuously monitored the game state of all 16k players of a single Tribal Wars game round for more than 1.5 years. The data collected from the game’s public API enables us to take an empirical perspective on the dynamics of a long-term game in which rounds last years and thus require a long-term commitment by the active players. We monitor the progression of the game state over time and identify situations that influence the player experience (e.g., causing players to keep playing or to quit).

C.1.3 Results

We envision new data-driven techniques to explore long-term gaming dynamics and user behavior, especially w.r.t. churn as an ultimate Quality of Experience signal. While creating causality cannot easily be achieved with data-driven methods, this approach does not replace classical lab tests. However, we show at the example of Tribal Wars over 1.5 years that data can provide good candidates and hypotheses for further evaluation.

C.2 Related Work

Gaming QoE. The QoE community has focused on studying factors that influence gaming QoE within the recent years. These efforts focused on certain game architectures—e.g., cloud [Schmidt, 2021] and mobile games [Slivar et al., 2015]—or game genres: e.g., MMORPGs [Suznjevic et al., 2013] and First Person Shooters [Vlahovic et al., 2019]—or display technologies such as VR gaming and related cybersickness [Vlahovic et al., 2021]—to name a few. These studies provided an understanding of factors that influence gaming

QoE [Möller et al., 2013], which resulted in standardized study methods for gaming quality (P.809 [ITU-T Recommendation, 2018]) or the opinion model predicting gaming QoE for cloud gaming services (G.1072 [ITU-T Recommendation, 2020]). All of these studies have so far focused on short study durations and did not evaluate game dynamics in situations where the gameplay spans over months or years.

Long-Term QoE Integration. Research on the perceived quality over multiple usage episodes (multi-episodic QoE) emerged in 2011 with multi-day experiments [Möller et al., 2011, Guse and Möller, 2013]. Later, *individual sessions* were studied (i.e., continuous use of the same service with multiple usage episodes) [Guse et al., 2017, Guse et al., 2020]. Despite first findings, the formation process of multi-episodic perceived quality remains far from being understood. We posit that a data-driven perspective on the complete game state complement complex field or lab tests and can thus provide an interesting perspective on the long-term usage of a service.

C.3 Long-Term Gaming and Data Collection

Tribal Wars Game. We base our study on the popular massively multiplayer online game (MMOG) *Tribal Wars*, that was first released in 2003 [InnoGames,]. In *Tribal Wars*, players start off controlling a medieval village that needs to be expanded and protected. Players can team-up in tribes and conquer other players' villages to expand their empire. Any player may join various isolated and dedicated game *rounds*. Such game rounds have a specific goal that needs to be reached and is typically played for multiple years (except speed rounds). A list of currently active rounds can be obtained at twstats.com. Successfully playing and finishing a round thus requires a long-term frequent time commitment by a player.

C.3.1 Dataset and Data Collection

To enable statistics collection (e.g., twstats.com), the game platform provides a public API to collect current snapshots of the entire game state (see [Tri,] for an API documentation). We retrieve the data for a single casual round every hour starting in May 2020 until December 2021 at varying update frequencies by the API provider. It captures the state of all 16k players, 93k villages, 1.4k tribes, and other game-related events and properties of the game world. Note that *casual* rounds enforce heavy restrictions on possible hostile attacks and conquests.

Ethics. The collected game state data is publicly provided by the publisher to enable statistical analysis of the game state. It does not contain personal or otherwise sensitive information.

Public Dataset and Evaluation Pipeline [SD2]. Our work constitutes a proof of concept of how to leverage domain-specific data-driven approaches in time-series analysis for hypothesizing about quality of experience factors of otherwise infeasible timeframe possibly spanning years. To enable reproducibility and encourage future research on this very dataset as a starting point, we decided to open access our dataset and analysis pipeline [SD2].

C.4 Long-Term Game Dynamics Analysis

After discussing the players' active lifetime, we explain emerging game dynamics of which we then derive experience indicators.

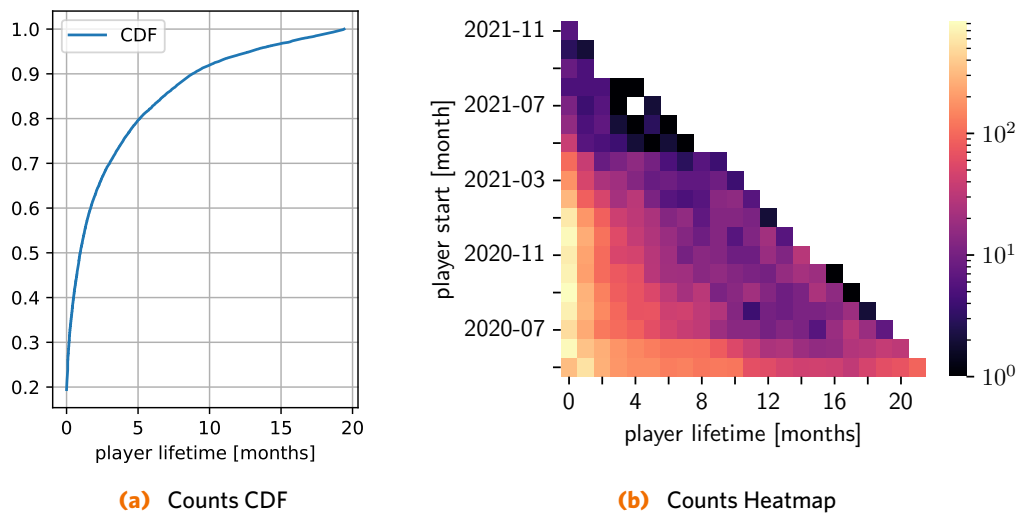


Figure C.2: Player Lifetime. (a) *Left:* Heatmap of Player Lifetime according to starting months. We observe a rich set of players joining the game within the first two months that likely stay in the game much longer. The later a player joins the game, the shorter lifetimes are observed overproportionally, though naturally bound. (b) *Right:* Cumulative Distribution Function (CDF) of Player Lifetime in months. Most Player lose interest in the game, the lifetime distribution is long-tailed.

C.4.1 Player Lifetime

As we focus on player churn as a signal for user experience, we set out to provide insights into player lifetime distributions.

We set out with discussing the active player lifetime as a Cumulative Distribution Function (CDF) in Figure C.2a. The time players keep playing is heavy-tailed as about 50% (75%) only last one (4) month(s), whereas only few players stick to the game until the very end. We add more detail to this observation by also incorporating the starting month of a player (y-axis) in comparison to the active lifetime in months (x-axis) as a heatmap in Figure C.2b. As expected from the lifetime distribution CDF, most players last only very few months. However, this detailed view shows a heavy decay in this distribution over the starting month. Furthermore, we observe a skew within this gradient towards the early game stages, i.e., players starting early within the game round are more likely to play longer.

Next, we take another view of these lifetime distributions—translating them into weekly counts of new and quitting players over time as displayed in Figure C.3a. The *new*-series (blue) show decreasing weekly new players over time as already implicitly observed in the lifetime heatmap (heatmap cumulative columns equal the counts here); The game experiences almost no new players beginning with April 2021, about one year into this round. In comparison to the new player, the *quit*-series (orange) denotes amounts of players leaving the round for good. While the loss in players is steady over the first year, it declines with alongside overall player numbers (shown in Figure C.5, orange). The difference in new to leaving players allows a *net*-player flow measure (green). Only after few months into the game round, i.e., September 2020, actual player numbers are decreasing constantly.

C.4.2 Inevitable Tension in Expansion

We begin by studying game dynamics from data, which is fundamental to derive player experience indicators (C.4.3).

The need for continuous growth. The monitored round is a *casual* round, which lets most players focus on only constructing and expanding their villages rather than on de-

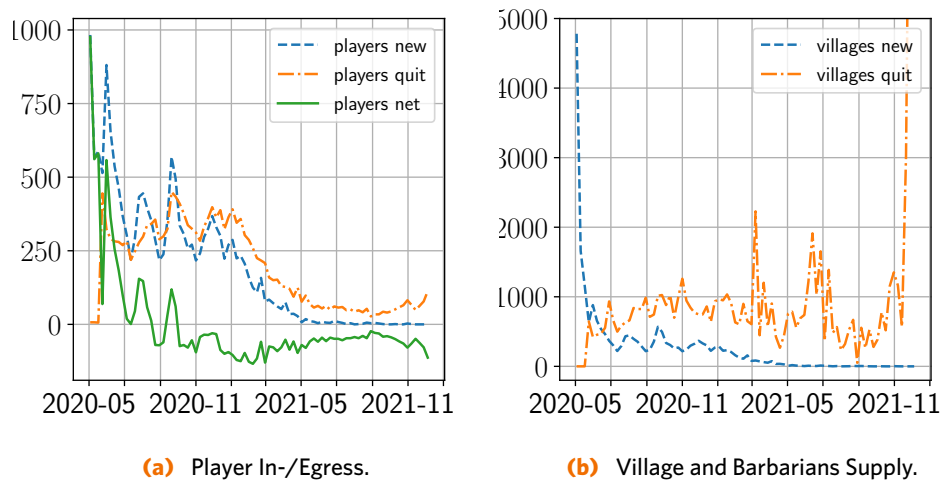


Figure C.3: Village Increase and Fresh Barbarians Supply. (a) *Left:* Plotted series denote the weekly amount of *new* (blue) and *quitting* players (orange), of which the *p net* series describes the difference to the game. We observe declining figures in new players within the early months, which evolves into steady player churn over the first half year into the game. Afterwards, the game world experiences net active player loss. (b) *Right:* The amount of *new* villages (blue) decreases linked to new player numbers (left). Furthermore, villages becoming barbarian due to players *quitting* (orange) remains at a steady level; although less players leave the game, they generally possess more villages in later stages.

fending against or attacking others. Thus, continuous growth and expansion becomes the ultimate goal. Yet, the room for expansion gets smaller over time, creating tension. Consequently, the player experience largely depends on the remaining space.

Acquiring new villages. For each new joining player, the game spawns a new player village and additional unoccupied villages, *barbarians*, concentrically on a 2d world map, i.e., the oldest villages are within the center and players joining later start off at larger distances to the game map center. We begin by discussing the amount of free villages over time by week in Figure C.3b. The series *new* (blue) denotes created villages due to new players joining the game. Note that the start of our observation is a cumulative value from the timeframe before and is thus not representative to other weekly aggregates. There is a clear trend in declining numbers of newly created villages directly correlated to new player figures discussed in Figure C.3a. I.e., world expansion nearly comes to an end in mid 2021. The series *quit* (orange) represents the amount of villages that become unoccupied (barbarians) due to players quitting the game. Throughout the game period, there is a constant churn in players with net decreasing figures, amounts of villages becoming barbarian fluctuate due to individual (large) players leave in later game stages.

Next, we analyze arising expansion pressure in detail by illustrating the village-resource concentration within the game world. In our case we find only ≈ 5.7 villages per unique player, which heavily dissatisfies demands limiting room for expansion thereby creating pressure. To showcase this pressure for expansion, we compute the density of non-barbarian villages for each occupied village individually in a radius of up to 30. We aggregate these values for July 1, 2020, as a heatmap in Figure C.4 in which darker colors represent low pressure areas that still have barbarian villages left to satisfy expansion demands. Naturally, we identify concentric outer regions with available barbarian villages due to having spawned only recently. However, over time the used game world expands with more players and outer low-pressure regions transform rapidly as shown by the circles indicating the expanse one month earlier and later (darkgreen).

We confirm this trend in Figure C.5, where we plot pressure-related measures and the amount of active players over time. In the beginning of our observation period, global vil-

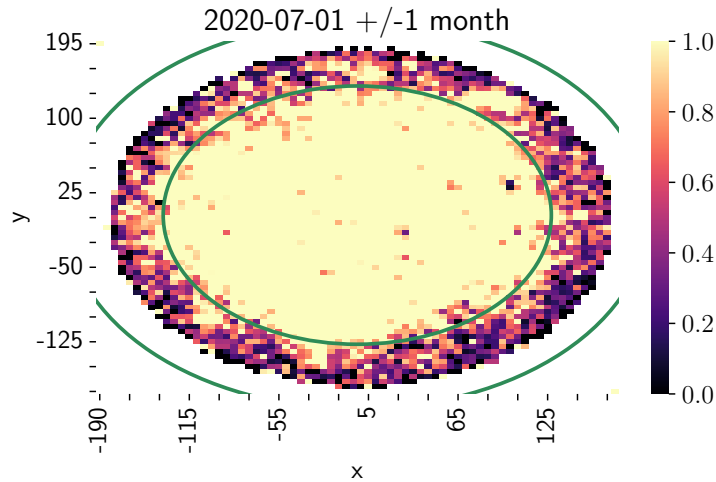


Figure C.4: Village Pressure. Heatmap showing the ratio of available villages in the 2d-game world on July 1 2020. With new players, the game spawns new villages concentric (circles indicate ± 1 month). Due to player expansion, unoccupied villages decrease over time, resulting in a declining village pressure gradient towards the younger outer regime.

lage pressure (blue) is only at a ratio of 0.65, i.e., 65% of all available villages are already occupied. This ratio increases rapidly over 4 months into the game, i.e., September 2020, almost reaching full village occupation (≈ 1.0). Next, we look at the local pressure, which we define as the ratio of available villages in the local neighborhood of all occupied villages. The quantiles (0.25, 0.5, 0.75; black) of local village pressure vary accounting for developments at the outer regimes of the gaming world, but follow the same trend.

Expanding existing villages. Besides acquisition of new villages, construction in possessed villages is equally important; thus, we also show the construction potential proxied by remaining optimal village points (green) across time, which naturally decreases due to limited village resources.

Number of players. We provide the number of active players, but normed by the maximum value (orange) as a comparison. Within early stages, active player figures increase alleviating expansion pressure (active players and local pressure Q25 counteract), while players start dropping out in 2021. As new player figures are flattening, expansion pressure is ever increasing. The vanishing potential to fortify existing villages (green line) correlates with declining numbers of active players. Further ever higher pressure correlates with shorter player lifetimes at later game stages (as shown in [C.4.1 Player Lifetime](#)).

The only *competitive* solution for continuing to play is further expansion by either scanning for nearby abandoned villages, targeting newly spawned villages at possibly very long runtimes, or engaging in hostile actions against other players within allowed casual game restrictions. Either way, obtaining new villages becomes highly competitive the more the game progresses and the more villages are already occupied by other players. We spotlight in-game village conquers over time in [Figure C.6a](#). Focusing the overall conquers (total; blue), we observe a correlation to active player numbers, primarily driven by conquers on barbarian villages (orange) until end of 2020. In later game stages, abandoned villages from leaving players get re-occupied within only few days. In line with observed expansion pressure, players experience a more and more aggressive environment at increasing figures of hostile conquers (green), which largely happens disproportionate in player points (red), i.e., the rich get richer.

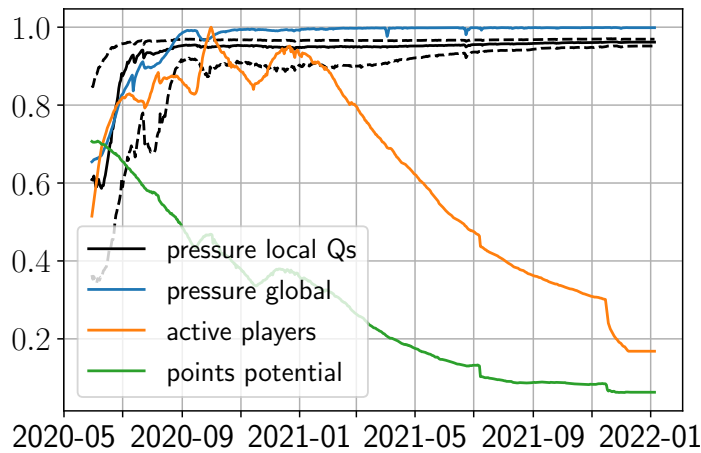


Figure C.5: Pressure, Players and Potential. The village pressure changes drastically within the early phase of our observation, albeit more joining players counteract on short term. Beginning with 2021 however, players lose interest, while potential for further construction of possessed villages decreases.

Further, the CDF of mean player kills over time in Figure C.6b reveals an increasing trajectory in per player kill counts over time confirms this more hostile environment. While *att* denotes killed units via active attacks, *def* describes killed amounts while defending own villages. As the *def* values are strictly larger than the attack kills, most attacks are defended by volume—unsurprisingly as the game implements defender’s advantages. Further the *other* series describes kills via proxy wars that happen when a player decides to defend foreign villages: e.g., barbarians to conquer them in a planned and well defended manner.

In summary, continuing to play requires ever harder expansion due to dwindling potentials and increasing competitiveness.

C.4.3 Player Experience Indicators

Informed by game dynamics shown before, we set out to identify features that relate to player experience. Since we cannot directly measure the player experience, we leverage user churn as a negative experience signal. We align players in-game time enabling a systematic comparison of player metric populations over time. Thus, we first interpolate starting dates of players that joined before our observation period from the conquers log by distance to the game map center. Next, we slice weekly buckets through the relative in-game time assigning each player to a non-churn and churn class depending on whether the user will quit within the next two weeks. We study per-player metrics derived from amounts of villages, point developments, kills, (hostile) conquers, and the village pressure. This allows for identifying significant differences in metric populations according to a MannWhitneyU-test.

In Table C.1, we present positive indicator significance responses (at $p < 0.1$ in over 90% of the weekly buckets) on players’ first four weeks (timeframe for most players), and overall in comparison. We further provide insights about the in-indicator relation between the non-churner and churner group by comparing averages across buckets.

The relation between both groups’ average indicator values immediately point into the success direction to varying degrees (“=”: within the same regime; “>”: single digit relative difference; “>>”: an order of magnitude; “>>>”: two orders of magnitude). While most identified player experience indicators mostly reflect in-game activity and success, global village pressure and amounts of acquired barbarian villages are significant within a player’s first 4 weeks. Heavy user influx and low pressure regions within early game stages explain this

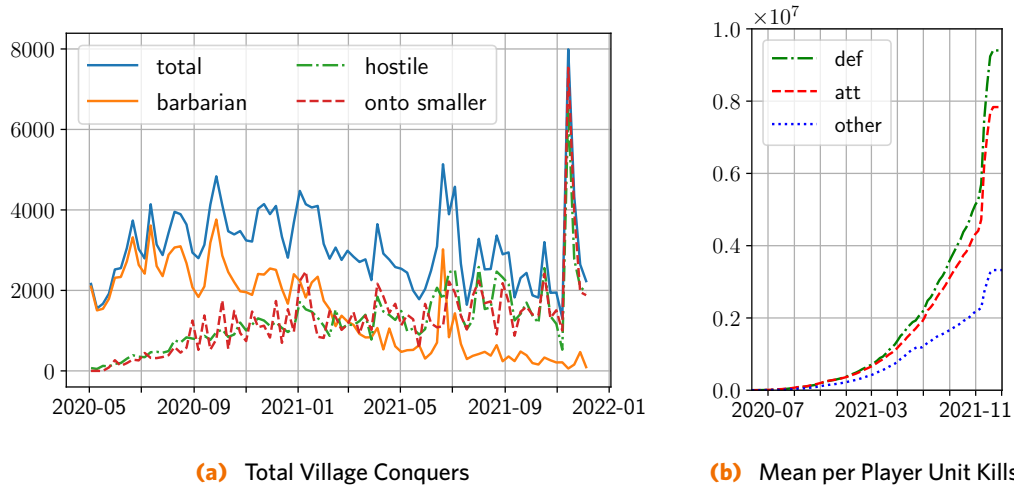


Figure C.6: Hostility: Conquers and Kills (a) Left: Game conquers over time: Total conquers (blue) remain stable. While barbarian conquers (orange) decrease (due to expansion pressure), hostile takeovers (green) increase. Such takeovers usually happen against smaller targets (red). (b) Right: We observe an exponential increase in kills for players defending (green), attacking (red), or proxying wars (blue).

| metric | significance $p < 0.1$ | | avg relation non/churn |
|-------------------------|-------------------------|-----------------|---------------------------|
| | first 4 weeks | overall | |
| player rank | $1.626 \cdot 10^{-7}$ | 0.004248 | < 0.762 |
| player villages | $2.104 \cdot 10^{-7}$ | 0.000003 | > 3.568 |
| player points | $2.866 \cdot 10^{-2}$ | 0.000018 | > 3.629 |
| has ally fraction | $4.209 \cdot 10^{-117}$ | 0.007655 | > 1.574 |
| village pressure global | $1.443 \cdot 10^{-11}$ | 0.230208 | = 0.996 |
| village pressure local | $1.783 \cdot 10^{-4}$ | 0.414349 | = 1.004 |
| villages gain | $1.569 \cdot 10^{-8}$ | 0.067301 | \gg 16.82 |
| villages gain barbarian | $5.366 \cdot 10^{-8}$ | 0.219401 | \gg 16.92 |
| kill all | $2.372 \cdot 10^{-12}$ | 0.073920 | > 2.697 |
| kill att | $2.382 \cdot 10^{-9}$ | 0.070090 | > 2.990 |
| kill def | $7.096 \cdot 10^{-6}$ | 0.144612 | > 2.224 |
| kill other | $4.408 \cdot 10^{-1}$ | 0.067183 | \gg 343.6 |

Table C.1: Significant player experience indicators. We identify significant differences in feature distributions of players churning within the next two weeks compared to others. Some features are only significant in the beginning of a player's career. The features and relation between non-/churners reflect measures of in-game success.

finding providing a hypothesis that apparent rising expansion pressure might be a driving factor for a bad user experience resulting in churn; both measures are indistinguishable at later game stages creating heavy tensions in continuing to expand.

C.5 Conclusions and Future Directions

No gaming QoE study is based on monitoring the state of *all players* that play a dedicated game round *for years*. In this section, we shed light on this still unexplored area of gaming experience. While selecting an appropriate game is challenging, we posit that long-term, round-based games such as Tribal Wars are a prime candidate to study player experience. We reason that well-defined rounds with a shared goal being played for years enable controlled statistical evaluations of player dynamics. This way, we aim to lay the foundation of future work in the direction of multi-episodic gaming QoE.

We identify features capturing in-game success and a hypothesis relating to the player experience. While e.g., questionnaires about specific game events or sequences may provide sentiments for better modeling experience, we argue that such a data-driven perspec-

tive is a necessary first step to identify influence factors that in turn enable their study in classical interactive tests. That is, such a data-driven perspective resulting in significant correlations cannot replace classical studies that identify causality, yet it can enable them by identifying test cases. This way, we aim to pave the way for exploring gaming QoE in a new field: long-term gaming experience.

Chapter Summary

An alive platform with growing, and sustainable communities needs to address various pressing factors. First, a platform must implement mechanisms against adverse content. Due to scaling factors, this usually involves distributed feedback schemes. After having introduced adoption process insights to—among others— Jodel DE & SA, and provided detailed analyzes of the user interactions, we have shifted to an interactional empirical level investigating how **★ A) Content Moderation** allows the operator to regularize the communities. We provide structural empirical evaluations w.r.t. voting and blocking mechanisms, while also modeling abusive content with Masked Language Models. Second, a platform must ensure a healthy influx of new users, preferably at exponentially increasing net-plus rates. Thus, we are next interested in user churn. That is, we investigate churning factors and determine the extent of **★ B) User Lifetime**. We provide insights to the communities in both Jodel countries DE & SA, while also modeling user lifetime value specifically for KSA and discuss applicability and generalizability. While a user's decision to leave a platform for good is a final signal of user experience, the operator usually has interest to uncover reasons. As Jodel does not set any time restrictions for participation, research suggests high degrees of ephemerality. We developed a blueprint on a prime example MMOG implementing long-term round structures spanning years to measure **★ C) Long-Term Gaming Quality of Experience** laying a data-driven foundation to the quality of experience research community for data-driven (pre-)studies that serve as a key enabler boosting the hypothesizing process that then needs further validation with classical methods—for a better understanding of user experience, including reasons for churn.



FUTURE WORK

IF WE HAD 10 MORE YEARS...

The king is dead, long live the king!

Throughout this thesis, we focus on Jodel, a new type of social media application implementing anonymity and hyperlocality. Leveraging ground truth data, we shed light on identified key essential online platform focus areas applying data-driven methods. While we have investigated and shown a broad holistic view on Jodel, much of our work can be deepened. However, many aspects of this thesis might likewise be complemented in more general directions, which we will outline next.

A Community Graph Perspective and Content Diffusion

This thesis provides vast insights into apparent interaction distributions mostly being heavy-tailed. However, unlike much research on social networks, we did not focus on a graph perspective due to prevalent anonymity. Given its anonymity, Jodel does not implement any type of user profile and thus removes possibilities to form specific social ties, or following relationships. We provide evidence of absent social ties within [♦ A.2.4: Anonymous Friends](#), i.e., within the realm of Jodel, specific conversational partners are unlikely to interact frequently, raising the question of social networkedness on anonymous platforms. Indeed, shallow preliminary investigations on the emerging Jodel interaction graphs have shown to indeed providing signals that these structures may better be modeled with random graphs than small-world networks, which are usually prevalent in social networks (it is likely that you know many friends of your friends).

Though we have discussed information diffusion through the Jodel via hashtags [♦ A: Information Spreading along Hashtags](#), we focused on a community meta perspective, rather than the individual participating users. Research has shown that Jodel, like other platforms, create community self-identities and community specific rules, yet the influence of specific participants and the main drivers of happening discourse remains unknown. As such, we argue that modeling content on a per-user basis in combination with a graph perspective might reveal interesting insights how content is appreciated and propagated within single communities. That is, one might use state-of-the-art neural content embeddings, e.g., SentenceTransformer [Reimers and Gurevych, 2019], and apply diffusion through the graph perspective to identify topical user clusters.

B Cross-Platform Perspective

While we focused on the independent analysis of Jodel and implications due to its design properties, adding a cross-platform perspective from other social networks or anonymous platforms will certainly help to understand effects of design decisions. As shown, even the very same ingredients on the very same platform might result in vastly different user behavior. That is, a cross-platform approach can only provide insights in the big picture, but it would nonetheless be of interest how non-anonymous, or global platforms compare to Jodel.

Such comparison studies should also incorporate modeling as suitable, i.e., emerging graph structures of social networks usually evolve into small-world networks that might be modeled e.g., by a Watts-Strogatz [Watts and Strogatz, 1998] Model, while anonymous network may tend to resemble scale free random networks, modeled by e.g., Erdős-Rényi [Erdős et al., 1959] models.

C Content Moderation and Radicalization

As mentioned before with our contributions to **◆ USER MANAGEMENT**, managing a healthy sustained growth has become a major challenge in any of the nowadays online platforms.

Distributed Moderation Mechanisms. Due to the sheer amount of created content, most online platforms implement a distributed moderation mechanism, that would not scale otherwise. Such distributed moderation schemes have been empirically investigated on various platforms, also as shown for Jodel with our contribution. Specifically our cross-country comparison highlights that user bases might appreciate different platform interactions more than others, i.e., the Saudi user base prefers creating content in comparison the German communities, where voting constitutes the most popular interaction.

However, it remains unclear how to measure the effectiveness and success-factors of such moderation schemes: Can we observe self-reinforcing feedback loops across communities w.r.t. voting behavior, and more importantly, which are factors that are crucially important to a working moderation mechanism, such as a sufficiently critical user mass?

Radicalization. Specifically implementing anonymity has shown the possibility to tilt into toxic environments. Within online social platforms, like-minded users might gather and create filter bubbles in a reinforcing process. That is, even with distributed moderation mechanisms at place, users might accept or even enjoy such toxicity or apparent bubbles they are part of. As such, operators usually still need to monitor contents to ensure fighting against such potential harmful excesses. Though we have shown how a classifier might be used as a textual baseline for comparing independent communities w.r.t. abusive content, this topic still needs further attention.

D User Perspective

Discussed future topic focus primarily on the communities. However, inferring more about how users behave from a content perspective is another angle to leverage our dataset. While much work splits its users into producers and consumers, it becomes interesting in which ways the producers setup themselves in the veil of anonymity. Does anonymous usage lead to consistent opinions, or do users ad hoc choose their stance? Other important factors connecting to our work on word-emoji embeddings incorporates further digging into language use and language variation, of which the latter has already qualitatively been shown to appear across geographical regions.



CONCLUSIONS

What we have learnt on this journey.

Online platforms have silently shaped our new digital becoming our daily driver for (casual) communication, entertainment, information and news, or countless other specific purposes. A large body of research has characterized, analyzed, and evaluated various aspects of such systems. Most popular platforms are globally accessible and implement some sort of profile, enabling following certain people, befriending with them, or more generically, enabling social ties and social credit. However, a new type of social media has emerged coupling the exact opposite: anonymity and hyperlocality, enabling anonymous communication that spatially links content and restricts access to users in its proximity.

Within this thesis, we set out to showcase the implications of this unique design property combination at the example of Jodel. Leveraging complete ground truth information provided by the operator, we complement existing research via a holistic data-driven view of identified four essential key platform topics.

◆ **USER ADOPTION** . General user adoption processes are inherently hard to observe due to the necessity to include desirably long temporal sequences. We showcase three different platform adoption processes across three platforms, showcasing their very birth, or specific platform changes w.r.t. user adoption. While detailing the early adoption of the ★ **A) Corona-Warn-App, Scheduled for the good [JH6]** , we provide insights to daily usage patterns and volume, while also dissecting interest on a spatial dimension across Germany. With escalated and ongoing Russo-Ukrainian hybrid warfare, we measure the re-purposing and operator reaction to ★ **B) Platform Sidechannels [JH1]** informing Russian citizens about the ongoing war. We showcase two very different adoption patterns of ★ **C) Jodel DE & SA [JH8]** . While the German communities evolve organically, the observed adoption in the Kingdom of Saudi Arabia appears similar to the CWA with a sudden influx of new users.

◆ **USER INTERACTIONS** . Next, we focus on user interactions happening on Jodel, providing a ★ **A) Structural Characterization** comparing the Jodel usage between both countries. We further detail interaction differences in a dedicated ★ **B) Cross-Country Insights [JH5]** , showcasing vast differences and platform implications. That is, we show how the very same platform properties may very different behavioral patterns. The Saudi users prefer discussions, while the German user base loves rather passive engagement. We

finish this chapter by characterizing and modeling Jodel [★ C\) Spotlighting Jodel SA \[JH8\]](#) providing rich insights to many similar, or qualitatively distinct interaction distributions.

◆ **USER CONTENT** . Apart from user interactions, actual contents of platform matter. That is, we explore [★ A\) Information Diffusion \[JH9\]](#) through the landscape of independent communities represented by hashtags across the Jodel platform in Germany. We show how larger communities act as information hubs; while types of used hashtags can be set apart by their temporal and spatial extent. Next, we develop a classification scheme and crowdsource actual message [★ B\) Content SA \[JH4\]](#) . While we find little evidence of toxicity, people enjoy sharing personal stories and beliefs, entertainment, and (local) information. Due to prevalent [★ C\) Emoji \[JH7, JH10\]](#) popularity in casual communication and their added expressiveness beyond words, we provide a differential empirical study for both countries. Shifting our focus to enabling tools w.r.t. content, we detail quantitative and qualitative insights to semantic associations captured in word-emoji embeddings on our Jodel data. Given the success, we provide evidence that our method leveraging semantic differentials can further add interpretability to word-emoji embeddings well in line with human judgement.

◆ **USER MANAGEMENT** . Our last chapter concentrates on mechanisms that keep the platform running. After empirically investigating Jodel's [★ A\) Distributed Moderation \[JH5\]](#) system, and discussing the modeling harmful contents, we exemplify how such a model can provide a content-baseline across independent communities. While abusive content constitutes an immediate threat to the platform, [★ B\) User Lifetime \[JH8, JH2\]](#) and churning factors are likewise of interest in the long term. After empirically investigating user churn, we show how to model a prediction from domain-specific metadata—and discuss model feature importance. Due to model insights appearing rather incomprehensive, we lastly set out to pave the way to generically derive quality of experience factors with a data-driven approach. That is, we feature an [★ C\) Excursus: Long-Term QoE \[JH3\]](#) , displaying dynamics and user churn in a long-term online game, enabling new hypotheses for subsequent classical analysis steps.

REFERENCES

- [ENA,] <https://developer.apple.com/documentation/exposurenotification>.
- [ENG,] <https://www.google.com/covid19/exposurenotifications/>.
- [cwa,] <https://github.com/ohohlfeld/corona-warn-app-monitor>.
- [Tri,] Tribal wars world state api. <https://forum.die-staemme.de/index.php?threads/weltdaten-und-configs.183996/>.
- [cor, 2020] (2020). Corona Warn App. <https://www.coronawarn.app/en/>.
- [CWA, 2020] (2020). Corona warn app faq: My risk status hasn't been updated for over a day. an internet connection was available, what's wrong? <https://www.coronawarn.app/en/faq/>.
- [Cor, 2020] (2020). Corona Warn App Github. <https://github.com/corona-warn-app>.
- [Cwa, 2020] (2020). German corona warn app (cwa) backend infrastructure overview. <https://github.com/corona-warn-app/cwa-documentation/blob/master/backend-infrastructure-architecture.pdf>.
- [DWA, 2020] (2020). Germany's coronavirus tracing app initially 'disabled' on Android smartphones. <https://p.dw.com/p/3frEh>.
- [App, 2020] (2020). statista: Anzahl der downloads der corona-warn-app über den apple app store und den google play store in deutschland im juli 2020. <https://de.statista.com/statistik/daten/studie/1125951/umfrage/downloads-der-corona-warn-app/>.
- [1pi6i, 2018] 1pi6i (2018). Instagram profile.
- [3w1_4, 2018] 3w1_4 (2018). Instagram profile.
- [5vmd, 2018] 5vmd (2018). Instagram profile.
- [8. Lee et al., 2019] 8. Lee, J., Li, J., and Mina, A. X. (2019). Hanmoji: what chinese characters and emoji reveal about each other. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 459--460.
- [Abbar et al., 2018] Abbar, S., Castillo, C., and Sanfilippo, A. (2018). To post or not to post: Using online trends to predict popularity of offline content. In *Proceedings of the 29th on Hypertext and Social Media*, pages 215--219.
- [Abnar and Zuidema, 2020] Abnar, S. and Zuidema, W. (2020). Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190--4197.

- [Ai et al., 2017] Ai, W., Lu, X., Liu, X., Wang, N., Huang, G., and Mei, Q. (2017). Untangling emoji popularity through semantic embeddings. In *Eleventh International AAAI Conference on Web and Social Media*.
- [Albert, 2015] Albert, G. (2015). Semiotik und syntax von emoticons. *Zeitschrift für angewandte Linguistik*, 2015(62):3--22.
- [Alsanea, 2007] Alsanea, R. (2007). *Girls of Riyadh*. Fig Tree.
- [Amjad et al., 2019] Amjad, M., Misra, V., Shah, D., and Shen, D. (2019). mrsc: Multi-dimensional robust synthetic control. *Measurement and Analysis of Computing Systems*.
- [Anderson et al., 2012] Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. (2012). Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *SIGKDD*. ACM.
- [Anderson et al., 2013] Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. (2013). Steering user behavior with badges. In *WWW*.
- [Antoun et al., 2020] Antoun, W., Baly, F., and Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. In *Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*.
- [Arora et al., 2015] Arora, P., Ganguly, D., and Jones, G. J. (2015). The good, the bad and their kins: Identifying questions with negative scores in stackoverflow. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1232--1239. IEEE.
- [Baele et al., 2021] Baele, S. J., Brace, L., and Coan, T. G. (2021). Variations on a theme? comparing 4chan, 8kun, and other chans' far-right "/pol" boards. *Perspectives on Terrorism*, 15(1):65--80.
- [Baltadzhieva and Chrupala, 2015] Baltadzhieva, A. and Chrupala, G. (2015). Predicting the quality of questions on stackoverflow. In *Proceedings of the international conference recent advances in natural language processing*, pages 32--40.
- [Barbieri et al., 2017] Barbieri, F., Ballesteros, M., and Saggion, H. (2017). Are Emojis Predictable? In *ACL*.
- [Barbieri and Camacho-Collados, 2018] Barbieri, F. and Camacho-Collados, J. (2018). How gender and skin tone modifiers affect emoji semantics in twitter. The Association for Computational Linguistics.
- [Barbieri et al., 2016] Barbieri, F., Ronzano, F., and Saggion, H. (2016). What does this Emoji Mean? A Vector Space Skip-Gram Model for Twitter Emojis. In *LREC*.
- [Basile et al., 2019] Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54--63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- [Baumgärtner et al., 2020] Baumgärtner, L., Dmitrienko, A., Freisleben, B., Gruler, A., Höchst, J., Kühlberg, J., Mezini, M., Miettinen, M., Muhamedagic, A., Nguyen, T. D., Penning, A., Pustelnik, D. F., Roos, F., Sadeghi, A.-R., Schwarz, M., and Uhl, C. (2020). Mind the gap: Security & privacy risks of contact tracing apps. arXiv cs.CR/2006.05914 <https://arxiv.org/abs/2006.05914>.

- [Bayne et al., 2019] Bayne, S., Connelly, L., Grover, C., Osborn, N., Tobin, R., Beswick, E., and Rouhanif, L. (2019). The social value of anonymity on campus: a study of the decline of yik yak. In *The Datafication of Education*. Taylor & Francis.
- [(BBC), 2022] (BBC), V. S. (2022). Ukraine war: Protester exposes cracks in kremlin's war message.
- [bduc_, 2018] bduc_ (2018). Instagram profile.
- [Becker et al.,] Becker, H., Naaman, M., and Gravano, L. Beyond Trending Topics: Real-World Event Identification on Twitter. ICWSM'11.
- [Beierle et al., 2021] Beierle, F., Dhakal, U., Corder, C., Eicher, S., and Pryss, R. (2021). Public perception of the german covid-19 contact-tracing app corona-warn-app. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 342--347. IEEE.
- [Benevenuto et al., 2009] Benevenuto, F., Rodrigues, T., Cha, M., and Almeida, V. (2009). Characterizing user behavior in online social networks. In *SIGCOMM*.
- [Benítez-Andrades et al., 2022] Benítez-Andrades, J. A., González-Jiménez, Á., López-Brea, Á., Aveleira-Mata, J., Alija-Pérez, J.-M., and García-Ordás, M. T. (2022). Detecting racism and xenophobia using deep learning models on twitter data: Cnn, lstm and bert. *PeerJ Computer Science*, 8:e906.
- [Berenguères and Castro, 2017] Berenguères, J. and Castro, D. (2017). Differences in emoji sentiment perception between readers and writers. In *BIGDATA*. IEEE.
- [Bergau, 2021] Bergau, M. (2021). Automatic humor detection on jodel. In *of the 8th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-Corpora 2021)*, page 11.
- [Bernstein et al., 2011] Bernstein, M. S., Monroy-Hernández, A., Harry, D., André, P., Panovich, K., and Vargas, G. G. (2011). 4chan and/b: An analysis of anonymity and ephemerality in a large online community. In *ICWSM*.
- [Bielenberg et al., 2012] Bielenberg, A., Helm, L., Gentilucci, A., Stefanescu, D., and Zhang, H. (2012). The growth of diaspora—a decentralized online social network in the wild. In *INFOCOM*. IEEE.
- [Birnholtz et al., 2015] Birnholtz, J., Merola, N. A. R., and Paul, A. (2015). Is it weird to still be a virgin: Anonymous, locally targeted questions on facebook confession boards. In *SIGCHI*. ACM.
- [Black et al., 2016] Black, E. W., Mezzina, K., and Thompson, L. A. (2016). Anonymous social media—understanding the content and context of yik yak. In *SIGCHI*. Elsevier.
- [Bock et al., 2020] Bock, K., Fax, Y., Reese, K., Singh, J., and Levin, D. (2020). Detecting and evading {Censorship-in-Depth}: A case study of {Iran's} protocol whitelister. In *10th USENIX Workshop on Free and Open Communications on the Internet (FOCI 20)*.
- [Bock et al., 2019] Bock, K., Hughey, G., Qiang, X., and Levin, D. (2019). Geneva: Evolving censorship evasion strategies. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 2199--2214.
- [Bojanowski et al., 2017a] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017a). Enriching word vectors with subword information.

- [Bojanowski et al., 2017b] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017b). Enriching word vectors with subword information. *Transactions of the ACL*.
- [Bollen et al., 2011] Bollen, J., Mao, H., and Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM*.
- [Bosu et al., 2013] Bosu, A., Corley, C. S., Heaton, D., Chatterji, D., Carver, J. C., and Kraft, N. A. (2013). Building reputation in stackoverflow: an empirical investigation. In *2013 10th Working conference on mining software repositories (MSR)*, pages 89--92. IEEE.
- [Breiman et al., 1984] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Routledge.
- [Brodersen et al., 2012] Brodersen, A., Scellato, S., and Wattenhofer, M. (2012). Youtube around the world: geographic popularity of videos. In *WWW*, pages 241--250.
- [Bronstein et al., 2021] Bronstein, M. M., Bruna, J., Cohen, T., and Velickovic, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*.
- [Buntain and Golbeck, 2014] Buntain, C. and Golbeck, J. (2014). Identifying social roles in reddit using network structure. In *Proceedings of the 23rd international conference on world wide web*, pages 615--620.
- [Cannarella and Spechler, 2014] Cannarella, J. and Spechler, J. A. (2014). Epidemiological modeling of online social network dynamics. *CoRR*.
- [Cao et al., 2012] Cao, N., Lin, Y.-R., Sun, X., Lazer, D., Liu, S., and Qu, H. (2012). Whisper: Tracing the spatiotemporal process of information diffusion in real time. *IEEE transactions on visualization and computer graphics*, 18(12):2649--2658.
- [Carmel et al., 2012] Carmel, D., Roitman, H., and Yom-Tov, E. (2012). On the relationship between novelty and popularity of user-generated content. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):1--19.
- [Caselli et al., 2020] Caselli, T., Basile, V., Mitrovic, J., and Granitzer, M. (2020). Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.
- [Catanese et al., 2011] Catanese, S. A., De Meo, P., Ferrara, E., Fiumara, G., and Provetti, A. (2011). Crawling facebook for social network analysis purposes. In *WIMS*, page 52.
- [Cavusoglu et al., 2015] Cavusoglu, H., Li, Z., and Huang, K.-W. (2015). Can gamification motivate voluntary contributions? the case of stackoverflow q&a community. In *Proceedings of the 18th ACM conference companion on computer supported cooperative work & social computing*, pages 171--174.
- [Cha et al., 2009] Cha, M., Mislove, A., and Gummadi, K. P. (2009). A measurement-driven analysis of information propagation in the flickr social network. In *WWW*.
- [Chaabane et al., 2014] Chaabane, A., Chen, T., Cunche, M., De Cristofaro, E., Friedman, A., and Kaafar, M. A. (2014). Censorship in the wild: Analyzing internet filtering in syria. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, pages 285--298.
- [Chamberlain et al., 2017] Chamberlain, B. P., Cardoso, Â., Liu, C. H. B., Pagliari, R., and Deisenroth, M. P. (2017). Customer lifetime value prediction using embeddings. In *SIGKDD*.

- [Chandra et al.,] Chandra, S., Khan, L., and Muhaya, F. B. Estimating Twitter User Location Using Social Interactions--A Content Based Approach. *SocialCom/PASSAT'11*.
- [Chang et al., 2015] Chang, Y.-T., Yu, H., and Lu, H.-P. (2015). Persuasive messages, popularity cohesion, and message diffusion in social media marketing. *Journal of Business Research*, 68(4):777--782.
- [Chen, 2012] Chen, C. (2012). The creation and meaning of internet memes in 4chan: Popular internet culture in the age of online digital reproduction.
- [Chen et al., 2019] Chen, J., Liang, D., Zhu, Z., Zhou, X., Ye, Z., and Mo, X. (2019). Social media popularity prediction based on visual-textual features with xgboost. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2692--2696.
- [Chen et al., 2018] Chen, P. P., Guitart, A., del Río, A. F., and Perriáñez, . (2018). Customer lifetime value in video games using deep learning and parametric models. In *Big Data*.
- [Chen et al., 2015] Chen, X., Geng, R., and Cai, S. (2015). Predicting microblog users' lifetime activities - a user-based analysis. *Electron. Commer. Rec. Appl.*
- [Chen et al., 2017] Chen, Y., Hoffman, M. W., Colmenarejo, S. G., Denil, M., Lillicrap, T. P., Botvinick, M., and Freitas, N. (2017). Learning to learn without gradient descent by gradient descent. In *International Conference on Machine Learning*, pages 748--756. PMLR.
- [Chiu and Nichols, 2016] Chiu, J. P. and Nichols, E. (2016). Named entity recognition with bidirectional lstm-cnns. *Transactions of the association for computational linguistics*, 4:357--370.
- [Cho et al., 2011] Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082--1090.
- [Choromanski et al., 2020] Choromanski, K. M., Likhoshesterov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J. Q., Mohiuddin, A., Kaiser, L., et al. (2020). Rethinking attention with performers. In *International Conference on Learning Representations*.
- [Clark et al., 2019] Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276--286, Florence, Italy. Association for Computational Linguistics.
- [Clark-Gordon et al., 2017] Clark-Gordon, C. V., Workman, K. E., and Linvill, D. L. (2017). College students and yik yak: An exploratory mixed-methods study. *Social Media+ Society*, 3(2):2056305117715696.
- [Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37--46.
- [Collins and Kent, 2022] Collins, B. and Kent, J. L. (2022). Facebook, twitter remove disinformation accounts targeting ukrainians. NBC News.
- [Conneau et al., 2020] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440--8451.

- [Correa et al., 2015] Correa, D., Silva, L. A., Mondal, M., Benevenuto, F., and Gummadi, K. P. (2015). The many shades of anonymity: Characterizing anonymous social media content. In *ICWSM*.
- [Correa and Sureka, 2014] Correa, D. and Sureka, A. (2014). Chaff from the wheat: characterization and modeling of deleted questions on stack overflow. In *WWW*, pages 631--642.
- [Crandall et al., 2007] Crandall, J. R., Zinn, D., Byrd, M., Barr, E. T., and East, R. (2007). Conceptdoppler: a weather tracker for internet censorship. *CCS*, 7:352--365.
- [Crawford and Gillespie, 2016] Crawford, K. and Gillespie, T. (2016). What is a flag for? social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3):410--428.
- [Danescu-Niculescu-Mizil et al., 2013] Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., and Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. In *WWW*.
- [De Boom et al., 2016] De Boom, C., Van Canneyt, S., Demeester, T., and Dhoedt, B. (2016). Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*.
- [De Candia et al., 2022] De Candia, S., De Francisci Morales, G., Monti, C., and Bonchi, F. (2022). Social norms on reddit: A demographic analysis. In *14th ACM Web Science Conference 2022*, pages 139--147.
- [De Francisci Morales et al., 2021] De Francisci Morales, G., Monti, C., and Starnini, M. (2021). No echo in the chambers of political interactions on reddit. *Scientific reports*, 11(1):1--12.
- [deepset AI, 2019] deepset AI (2019). German bert.
- [Delort et al., 2011] Delort, J.-Y., Arunasalam, B., and Paris, C. (2011). Automatic moderation of online discussion sites. *International Journal of Electronic Commerce*, 15(3):9--30.
- [Dennis et al., 2016] Dennis, E. E., Martin, J. D., and Wood, R. (2016). Media use in the middle east, 2016: A six-nation survey.
- [Deutsche Welle, 2020a] Deutsche Welle (2020a). Berlin-Neukölln: Low income and migrant families in coronavirus lockdown. <https://p.dw.com/p/3e0Cd>.
- [Deutsche Welle, 2020b] Deutsche Welle (2020b). Germany maps out coronavirus regulations on domestic travel. <https://p.dw.com/p/3eQL4>.
- [Deutsche Welle, 2020c] Deutsche Welle (2020c). Germany puts two western districts on lockdown. <https://p.dw.com/p/3eEdt>.
- [Devlin et al., 2019a] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019a). Bert: Pre-training of deep bidirectional transformers for language understanding.
- [Devlin et al., 2019b] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019b). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171--4186.
- [Ding et al., 2019] Ding, K., Wang, R., and Wang, S. (2019). Social media popularity prediction: A multiple feature fusion approach with deep neural networks. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2682--2686.

- [Dong et al., 2021] Dong, C., Bharambe, S., and Bick, M. (2021). Why do people not install corona-warn-app? evidence from social media. In *European, Mediterranean, and Middle Eastern Conference on Information Systems*, pages 305--318. Springer.
- [Dow et al., 2013] Dow, P. A., Adamic, L. A., and Friggeri, A. (2013). The anatomy of large facebook cascades. In *ICWSM*.
- [Dror et al., 2012] Dror, G., Pelleg, D., Rokhlenko, O., and Szpektor, I. (2012). Churn prediction in new users of yahoo! answers. In *WWW-CQA*.
- [Eisner et al., 2016a] Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M., and Riedel, S. (2016a). emoji2vec: Learning Emoji Representations from their Description. In *SocialNLP*.
- [Eisner et al., 2016b] Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M., and Riedel, S. (2016b). emoji2vec: Learning emoji representations from their description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, Austin, TX, USA. Association for Computational Linguistics.
- [Elson et al., 2020] Elson, S. B., Kelty, R., Paulson, K., Bornmann, J., and De Angelis, K. K. (2020). Social media use at a us military academy: Perceived implications for performance and behavior. In *Social Media and the Armed Forces*, pages 31--49. Springer.
- [Engler et al., 2023] Engler, J., Sikdar, S., Lutz, M., and Strohmaier, M. (2023). Sensepoplar: Word sense aware interpretability for pre-trained contextual word embeddings. In *EMNLP*.
- [Erdos et al., 1959] Erdos, P., Rényi, A., et al. (1959). On the evolution of random graphs.
- [Fader et al., 2005a] Fader, P. S., Hardie, B. G., and Lee, K. L. (2005a). Rfm and clv: Using iso-value curves for customer base analysis. *Journal Marketing Research*.
- [Fader et al., 2005b] Fader, P. S., Hardie, B. G., and Lee, K. L. (2005b). "counting your customers" the easy way: An alternative to the pareto/nbd model. *Marketing science*.
- [Fair and Wesslen, 2019] Fair, G. and Wesslen, R. (2019). Shouting into the void: A database of the alternative social media platform gab. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 608--610.
- [Fang,] Fang, D. Chinese propaganda and cyber-nationalism under the russia--ukraine war.
- [Farrell et al., 2019] Farrell, T., Fernandez, M., Novotny, J., and Alani, H. (2019). Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM Conference on Web Science*, pages 87--96.
- [Felbo et al., 2017a] Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., and Lehmann, S. (2017a). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615--1625.
- [Felbo et al., 2017b] Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., and Lehmann, S. (2017b). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *EMNLP*.
- [Felbo et al., 2017c] Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., and Lehmann, S. (2017c). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm, emoji,emojis. In *EMNLP*.

- [Feldmann et al., 2021] Feldmann, A., Gasser, O., Lichtblau, F., Pujol, E., Poese, I., Dietzel, C., Wagner, D., Wichtlhuber, M., Tapiador, J., Vallina-Rodriguez, N., et al. (2021). Implications of the covid-19 pandemic on the internet traffic. In *Broadband Coverage in Germany; 15th ITG-Symposium*. VDE.
- [Ferragina et al., 2015] Ferragina, P., Piccinno, F., and Santoro, R. (2015). On analyzing hashtags in twitter. In *ICWSM*.
- [Ferrara et al.,] Ferrara, E., Varol, O., Menczer, F., and Flammini, A. Traveling Trends: Social Butterflies or Frequent Fliers? *COSN'13*.
- [Ferraz Costa et al., 2015] Ferraz Costa, A., Yamaguchi, Y., Juci Machado Traina, A., Traina Jr, C., and Faloutsos, C. (2015). Rsc: Mining and modeling temporal activity in social media. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 269--278.
- [Ferrer et al., 2021] Ferrer, X., van Nuenen, T., Such, J. M., and Criado, N. (2021). Discovering and categorising language biases in reddit. In *ICWSM*, pages 140--151.
- [Ferretti et al., 2020] Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L., Parker, M., Bonsall, D., and Fraser, C. (2020). Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing. *Science*, 368(6491).
- [Fiesler et al., 2018] Fiesler, C., McCann, J., Frye, K., Brubaker, J. R., et al. (2018). Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*.
- [Fifield et al., 2012] Fifield, D., Hardison, N., Ellithorpe, J., Stark, E., Boneh, D., Dingleline, R., and Porras, P. (2012). Evading censorship with browser-based proxies. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 239--258. Springer.
- [Fink et al., 2016] Fink, C., Schmidt, A., Barash, V., Kelly, J., Cameron, C., and Macy, M. (2016). Investigating the observability of complex contagion in empirical social networks. In *ICWSM*.
- [Freeman, 2004] Freeman, L. (2004). The development of social network analysis. *A Study in the Sociology of Science*.
- [Frith and Saker, 2017] Frith, J. and Saker, M. (2017). Understanding yik yak: Location-based sociability and the communication of place. *First Monday*, 22(10).
- [Fung and Ji, 2022] Fung, Y. R. and Ji, H. (2022). A weibo dataset for the 2022 russo-ukrainian crisis. *arXiv preprint arXiv:2203.05967*.
- [Funke and Reips, 2012] Funke, F. and Reips, U.-D. (2012). Why semantic differentials in web-based research should be made from visual analogue scales and not from 5-point scales. *Field Methods*, 24(3).
- [Gaudette et al., 2021] Gaudette, T., Scrivens, R., Davies, G., and Frank, R. (2021). Upvoting extremism: Collective identity formation and the extreme right on reddit. *New Media & Society*, 23(12):3491--3508.
- [Gers et al., 2000] Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451--2471.
- [Ghosh and Ghosh, 2019] Ghosh, S. and Ghosh, S. (2019). Exploring the ideal depth of neural network when predicting question deletion on community question answering. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, pages 52--55.

- [Ghosh et al., 2016] Ghosh, S., Vinyals, O., Strope, B., Roy, S., Dean, T., and Heck, L. (2016). Contextual lstm (clstm) models for large scale nlp tasks. *arXiv preprint arXiv:1602.06291*.
- [Gilbert, 2013] Gilbert, E. (2013). Widespread underprovision on reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 803--808.
- [Golovchenko, 2022] Golovchenko, Y. (2022). Fighting propaganda with censorship: A study of the ukrainian ban on russian social media. *The Journal of Politics*, 84(2):000--000.
- [Gong et al., 2012] Gong, N. Z., Xu, W., Huang, L., Mittal, P., Stefanov, E., Sekar, V., and Song, D. (2012). +. In *IMC*.
- [Gonzalez-Ibanez et al., 2011] Gonzalez-Ibanez, R., Muresan, S., and Wacholder, N. (2011). Identif. sarcasm in Twitter: a closer look. In *NAACL-HLT*.
- [Goodrich and de Mooij, 2014] Goodrich, K. and de Mooij, M. (2014). How 'social' are social media? a cross-cultural comparison of online and offline purchase decision influences. *Journal of Marketing Communications*, 20(1-2):103--116.
- [Google, 2020] Google (2020). Google trends ``yudel", arabic. <https://trends.google.com/trends/explore?q=%D9%8A%D9%88%D8%AF%D9%84>.
- [Granik and Mesyura, 2017] Granik, M. and Mesyura, V. (2017). Fake news detection using naive bayes classifier. In *2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON)*, pages 900--903. IEEE.
- [Grant and Betts, 2013] Grant, S. and Betts, B. (2013). Encouraging user behaviour with achievements: an empirical study. In *Mining Software Repositories*.
- [Grootendorst, 2022] Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- [Grover and Leskovec, 2016] Grover, A. and Leskovec, J. (2016). Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, page 855-864, New York, NY, USA. Association for Computing Machinery.
- [Guhr et al., 2020] Guhr, O., Schumann, A.-K., Bahrmann, F., and Böhme, H. J. (2020). Training a broad-coverage german sentiment classification model for dialog systems. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1627--1632.
- [Guibon et al., 2018] Guibon, G., Ochs, M., and Bellot, P. (2018). Emoji recommendation in private instant messages. In *SAC*.
- [Guntuku et al., 2019] Guntuku, S. C., Li, M., Tay, L., and Ungar, L. H. (2019). Studying cultural differences in emoji usage across the east and the west. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 226--235.
- [Gupta et al., 2021] Gupta, M., Torrico, D. D., Hepworth, G., Gras, S. L., Ong, L., Cottrell, J. J., and Dunshea, F. R. (2021). Differences in hedonic responses, facial expressions and self-reported emotions of consumers using commercial yogurts: A cross-cultural study. *Foods*, 10(6):1237.
- [Guse et al., 2020] Guse, D., Hohlfeld, O., Wunderlich, A., Weiss, B., and Möller, S. (2020). Multi-episodic perceived quality of an audio-on-demand service. In *QoMEX*.

- [Guse and Möller, 2013] Guse, D. and Möller, S. (2013). Macro-temporal Development of QoE: Impact of Varying Performance on QoE over Multiple Interactions. In *Proceedings of AIA-DAGA Conference on Acoustics*.
- [Guse et al., 2017] Guse, D., Weiss, B., Haase, F., Wunderlich, A., and Möller, S. (2017). Multi-episodic perceived quality for one session of consecutive usage episodes with a speech telephony service. *Quality and User Experience*, 2(1).
- [Guta and Karolak, 2015] Guta, H. and Karolak, M. (2015). Veiling and blogging: Social media as sites of identity negotiation and expression among saudi women. *Journal of International Women's Studies*.
- [Hamilton, 2022] Hamilton, I. A. (2022). Google and tripadvisor disable restaurant reviews in russia after they were flooded with protests against the ukraine invasion. *Business Insider*.
- [Hamilton et al., 2018] Hamilton, W., Bajaj, P., Zitnik, M., Jurafsky, D., and Leskovec, J. (2018). Embedding logical queries on knowledge graphs. *Advances in neural information processing systems*, 31.
- [Hamp and Feldweg, 1997] Hamp, B. and Feldweg, H. (1997). GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.
- [Hanley et al., 2022] Hanley, H. W., Kumar, D., and Durumeric, Z. (2022). Happenstance: Utilizing semantic search to track russian state media narratives about the russo-ukrainian war on reddit. *arXiv preprint arXiv:2205.14484*.
- [Haq et al., 2022a] Haq, E.-U., Tyson, G., Braud, T., and Hui, P. (2022a). Weaponising social media for information divide and warfare. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, page 259-262, New York, NY, USA. Association for Computing Machinery.
- [Haq et al., 2022b] Haq, E.-U., Tyson, G., Lee, L.-H., Braud, T., and Hui, P. (2022b). Twitter dataset for 2022 russo-ukrainian crisis. *arXiv preprint arXiv:2203.02955*.
- [Häring et al., 2021] Häring, M., Gerlitz, E., Tiefenau, C., Smith, M., Wermke, D., Fahl, S., and Acar, Y. (2021). Never ever or no matter what: Investigating adoption intentions and misconceptions about the {Corona-Warn-App} in germany. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 77--98.
- [Help, 2022] Help, G. (2022). Maps user-generated content policy.
- [Heston and Birnholtz, 2016a] Heston, M. and Birnholtz, J. (2016a). (in) visible cities: an exploration of social identity, anonymity and location-based filtering on yik yak. *IConference 2016 Proceedings*.
- [Heston and Birnholtz, 2016b] Heston, M. and Birnholtz, J. (2016b). (in)visible cities: An exploration of social identity, anonymity and location-based filtering on yik yak. In *IConference*. iSchools.
- [Heuman, 2020a] Heuman, A. (2020a). Negotiations of language ideology on the jodel app: Language policy in everyday online interaction. *Discourse, Context & Media*, 33:100353.
- [Heuman, 2020b] Heuman, A. (2020b). Negotiations of language ideology on the jodel app: Language policy in everyday online interaction. *Discourse, Context & Media*, 33.

- [Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735--1780.
- [Hosseinmardi et al., 2014] Hosseinmardi, H., Ghasemianlangroodi, A., Han, R., Lv, Q., and Mishra, S. (2014). Towards understanding cyberbullying behavior in a semi-anonymous social network. In *ASONAM*.
- [Hoßfeld et al., 2013] Hoßfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., and Tran-Gia, P. (2013). Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing. *IEEE Transactions on Multimedia*, 16(2):541--558.
- [Hovy and Purschke, 2018] Hovy, D. and Purschke, C. (2018). Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383--4394.
- [Howard and Ruder, 2018] Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.
- [Hu et al., 2017a] Hu, T., Guo, H., Sun, H., Nguyen, T. T., and Luo, J. (2017a). Spice up your chat: the intentions and sentiment effects of using emojis. In *ICWSM*.
- [Hu et al., 2017b] Hu, T., Guo, H., Sun, H., Nguyen, T. T., and Luo, J. (2017b). Spice up your chat: the intentions and sentiment effects of using emojis. In *ICWSM*.
- [Huang et al., 2018] Huang, F., Chen, J., Lin, Z., Kang, P., and Yang, Z. (2018). Random forest exploiting post-related and user-related features for social media popularity prediction. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 2013--2017.
- [Huang et al., 2015] Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- [iiim7mdz, 2018] iiim7mdz (2018). Twitter profile.
- [Illendula and Yedulla, 2018] Illendula, A. and Yedulla, M. R. (2018). Learning emoji embeddings using emoji co-occurrence network graph. *Emoji*.
- [InnoGames,] InnoGames. Tribal wars. <https://www.innogames.com/games/tribal-wars/>.
- [ITU-T Recommendation, 2018] ITU-T Recommendation (2018). P.809: Subjective evaluation methods for gaming quality.
- [ITU-T Recommendation, 2020] ITU-T Recommendation (2020). G.1072: Opinion model predicting gaming quality of experience for cloud gaming services.
- [Jaderberg et al., 2017] Jaderberg, M., Czarnecki, W. M., Osindero, S., Vinyals, O., Graves, A., Silver, D., and Kavukcuoglu, K. (2017). Decoupled neural interfaces using synthetic gradients. In *ICML*.
- [Jain and Kasbe, 2018] Jain, A. and Kasbe, A. (2018). Fake news detection. In *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pages 1--5. IEEE.
- [Jhaver et al., 2019] Jhaver, S., Bruckman, A., and Gilbert, E. (2019). Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1--27.

- [Jiang et al., 2014] Jiang, A. H., Bischof, Z. S., and Bustamante, F. E. (2014). A cliq of content curators. In *Proceedings of the 2014 ACM conference on SIGCOMM*, pages 139--140.
- [Jiang et al., 2013] Jiang, J., Wilson, C., Wang, X., Sha, W., Huang, P., Dai, Y., and Zhao, B. Y. (2013). Understanding latent interactions in online social networks. *TWEB*.
- [jodel sa, 2018] jodel sa (2018). Twitter profile.
- [Jüttner et al., 2021] Jüttner, K., Nowak, P., Scheibe, K., Zimmer, F., and Fietkiewicz, K. J. (2021). The faceless vicinity: Who uses location-based anonymous social networks like jodel and why? In *International Conference on Human-Computer Interaction*, pages 54--73. Springer.
- [Kairam et al., 2012] Kairam, S., Brzozowski, M., Huffaker, D., and Chi, E. (2012). Talking in circles: selective sharing in google+. In *SIGCHI*.
- [Kamarudin et al., 2018] Kamarudin, N. S., Rakesh, V., Beigi, G., Manikouda, L., and Liu, H. (2018). A study of reddit-user's response to rape. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 591--592. IEEE.
- [Kamath et al., 2013] Kamath, K. Y., Caverlee, J., Lee, K., and et al. (2013). Spatio-temporal dynamics of online memes: a study of geo-tagged tweets. In *WWW*.
- [Kang et al., 2019] Kang, P., Lin, Z., Teng, S., Zhang, G., Guo, L., and Zhang, W. (2019). Catboost-based framework with additional user information for social media popularity prediction. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2677--2681.
- [Kang et al., 2016] Kang, R., Dabbish, L., and Sutton, K. (2016). Strangers on your phone: Why people use anonymous communication applications. In *CSCW*.
- [Karimi et al., 2018] Karimi, H., Roy, P., Saba-Sadiya, S., and Tang, J. (2018). Multi-source multi-class fake news detection. In *Proceedings of the 27th international conference on computational linguistics*, pages 1546--1557.
- [Karimi and Tang, 2019] Karimi, H. and Tang, J. (2019). Learning hierarchical discourse-level structure for fake news detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3432--3442.
- [Karumur et al., 2016] Karumur, R. P., Nguyen, T. T., and Konstan, J. A. (2016). Early activity diversity: Assessing newcomer retention from first-session activity. In *CSCW*.
- [Kasakowskij et al., 2018] Kasakowskij, R., Fietkiewicz, K. L., Friedrich, N., and Stock, W. G. (2018). Anonymous and non-anonymous user behavior on social media: A case study of jodel and instagram. *Journal of Information Science Theory and Practice*, 6(3):25--36.
- [Kaufer, 2022] Kaufer, S. (2022). An open letter on ukraine from tripadvisor's steve kaufer.
- [Kayes et al., 2015] Kayes, I., Kourtellis, N., Quercia, D., Iamnitchi, A., and Bonchi, F. (2015). The social world of content abusers in community question answering. In *WWW*.
- [Keneshloo et al., 2016] Keneshloo, Y., Wang, S., Han, E.-H., and Ramakrishnan, N. (2016). Predicting the popularity of news articles. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 441--449. SIAM.

- [Kenter et al., 2016] Kenter, T., Borisov, A., and De Rijke, M. (2016). Siamese CBOW: Optimizing Word Embeddings for Sentence Representations. *arXiv*.
- [Kimura and Katsurai, 2017] Kimura, M. and Katsurai, M. (2017). Automatic construction of an emoji sentiment lexicon. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 1033--1036.
- [Kimura-Thollander and Kumar, 2019] Kimura-Thollander, P. and Kumar, N. (2019). Examining the "global" language of emojis: Designing for cultural representation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1--14.
- [Kircher, 2016] Kircher, M. M. (2016). College gossip app you forgot about ditches anonymity. *nymag.com*.
- [Kotler, 2016] Kotler, P. (2016). *A framework for marketing management*. Pearson Education.
- [Kotsakos et al.,] Kotsakos, D., Sakkos, P., Katakis, I., and Gunopulos, D. #Tag: Meme or Event? ASONAM '14.
- [Kouloumpis et al., 2011] Kouloumpis, E., Wilson, T., and Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg! In *ICWSM*.
- [Kozyreva et al., 2021] Kozyreva, A., Lorenz-Spreen, P., Lewandowsky, S., Garrett, P. M., Herzog, S. M., Pachur, T., and Hertwig, R. (2021). Psychological factors shaping public responses to covid-19 digital contact tracing technologies in germany. *Scientific reports*, 11(1):1--19.
- [Kraut et al., 2012] Kraut, R. E., Resnick, P., Kiesler, S., Ren, Y., Chen, Y., Burke, M., Kittur, N., Riedl, J., and Konstan, J. (2012). *Building Successful Online Communities: Evidence-Based Social Design*.
- [Kriehn, 2021] Kriehn, L. (2021). Case study: the corona contact tracing app in germany. In *Digital responses to Covid-19*, pages 37--54. Springer.
- [Krippendorff, 2011] Krippendorff, K. (2011). Computing krippendorff's alpha-reliability.
- [Krippendorff, 2012] Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Sage publications.
- [Kumar et al., 2010] Kumar, R., Novak, J., and Tomkins, A. (2010). Structure and evolution of online social networks. In *KDD*. ACM.
- [Kusmierczyk and Gomez-Rodriguez, 2018] Kusmierczyk, T. and Gomez-Rodriguez, M. (2018). On the causal effect of badges. In *Proceedings of the 2018 world wide web conference*, pages 659--668.
- [La Cava et al., 2022] La Cava, L., Greco, S., and Tagarelli, A. (2022). Network analysis of the information consumption-production dichotomy in mastodon user behaviors. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1378--1382.
- [Laaksonen and Rantasila, 2021] Laaksonen, S.-M. and Rantasila, A. (2021). Rocketing sheep: Affective discipline in anonymous mobile social media jodel during the covid-19 pandemic. *AoIR Selected Papers of Internet Research*.
- [Lab and Others, 2014] Lab, C. and Others (2014). Url testing lists intended for discovering website censorship. <https://github.com/citizenlab/test-lists>.

- [Ladygina, 2022] Ladygina, Y. V. (2022). Cyborgs vs. vatniks: Hybridity, weaponized information, and mediatized reality in recent ukrainian war films. *East/West: Journal of Ukrainian Studies*, 9(1):105--138.
- [Lampe and Johnston, 2005] Lampe, C. and Johnston, E. (2005). Follow the (slash) dot: effects of feedback on new members in an online community. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 11--20.
- [Lampe and Resnick, 2004] Lampe, C. and Resnick, P. (2004). Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 543--550.
- [Lampe et al., 2014] Lampe, C., Zube, P., Lee, J., Park, C. H., and Johnston, E. (2014). Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly*, 31(2):317--326.
- [Lampe et al., 2007] Lampe, C. A., Johnston, E., and Resnick, P. (2007). Follow the reader: filtering comments on slashdot. In *SIGCHI*.
- [LaRose et al., 2014] LaRose, R., Connolly, R., Lee, H., Li, K., and Hales, K. D. (2014). Connection overload? a cross cultural study of the consequences of social media connection. *Information Systems Management*, 31(1):59--73.
- [Larsen, 2020] Larsen, M. C. (2020). Social media insecurities in everyday life among young adults--an ethnography of anonymous jodel disclosures. *AoIR Selected Papers of Internet Research*.
- [Leavitt, 2015] Leavitt, A. (2015). This is a throwaway account: Temporary technical identities and perceptions of anonymity in a massive online community. In *CSCW*.
- [Lee et al., 2017] Lee, J.-S., Yang, S., Munson, A. L., and Donzo, L. (2017). What people do on yik yak: Analyzing anonymous microblogging user behaviors. In *SCSM*, pages 416--428. Springer.
- [Lee-Thorp et al., 2021] Lee-Thorp, J., Ainslie, J., Eckstein, I., and Ontanon, S. (2021). Fnet: Mixing tokens with fourier transforms. *arXiv e-prints*, pages arXiv--2105.
- [Lemmens and Croux, 2006] Lemmens, A. and Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal Marketing Research*.
- [Leshno et al., 1993] Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861--867.
- [Leskovec et al.,] Leskovec, J., Backstrom, L., and Kleinberg, J. Meme-tracking and the Dynamics of the News Cycle. *KDD '09*.
- [Leskovec and Horvitz, 2008] Leskovec, J. and Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. In *WWW*.
- [Lewis et al., 2008] Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., and Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using facebook. com. *Social networks*, 30(4):330--342.
- [Lewter and Profit, 2018] Lewter, B. and Profit, S. (2018). Talk back with yik yak? exploring user engagement via anonymous social media in an academic library. *Journal of Web Librarianship*, 12(2):107--120.

- [Li et al., 2022] Li, H., Hecht, B., and Chancellor, S. (2022). All that’s happening behind the scenes: Putting the spotlight on volunteer moderator labor in reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 584--595.
- [Li et al., 2019] Li, M., Guntuku, S., Jakhetiya, V., and Ungar, L. (2019). Exploring (dis-) similarities in emoji-emotion assoc. on twitter and weibo. In *WWW*.
- [Li and Literat, 2017] Li, Q. and Literat, I. (2017). Misuse or misdesign? yik yak on college campuses and the moral dimensions of technology design. *First Monday*.
- [Li et al.,] Li, R., Lei, K. H., Khadiwala, R., and Chang, K. C. TEDAS: A Twitter-based Event Detection and Analysis System. *ICDE'12*.
- [Lipovetsky and Conklin, 2018] Lipovetsky, S. and Conklin, M. (2018). Decreasing respondent heterogeneity by likert scales adjustment via multipoles. *Stats*, 1(1):169--175.
- [Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- [Ljubešić and Fišer, 2016] Ljubešić, N. and Fišer, D. (2016). A global analysis of emoji usage. In *Web as Corpus Workshop*.
- [Loh, 2014] Loh, W.-Y. (2014). Fifty years of classification and regression trees. *International Statistical Review*, 82(3):329--348.
- [Lu et al., 2016] Lu, X., Ai, W., Liu, X., Li, Q., Wang, N., Huang, G., and Mei, Q. (2016). Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users. In *UBICOMP*.
- [Ludemann, 2018] Ludemann, D. (2018). /pol/emics: Ambiguity, scales, and digital discourse on 4chan. *Discourse, context & media*, 24:92--98.
- [Luo et al., 2021] Luo, Q., Liu, L., Lin, Y., and Zhang, W. (2021). Don’t miss the labels: Label-semantic augmented meta-learner for few-shot text classification. In *ACL-IJCNLP*.
- [Maaten and Hinton, 2008] Maaten, L. v. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*.
- [Magno et al., 2012] Magno, G., Comarela, G., Saez-Trumper, D., Cha, M., and Almeida, V. (2012). New kid on the block: Exploring the google+ social graph. In *IMC*.
- [Mahler, 2015] Mahler, J. (2015). Who spewed that abuse? yik yak isn’t telling.
- [Maier et al., 2009] Maier, G., Feldmann, A., Paxson, V., and Allman, M. (2009). On dominant characteristics of residential broadband internet traffic. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, pages 90--102.
- [Manku et al., 2004] Manku, G. S., Naor, M., and Wieder, U. (2004). Know thy neighbor’s neighbor: the power of lookahead in randomized p2p networks. In *STOC*.
- [Manning et al., 2010] Manning, C., Raghavan, P., and Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, 16(1):100--103.
- [Mathew et al., 2020] Mathew, B., Sikdar, S., Lemmerich, F., and Strohmaier, M. (2020). The polar framework: Polar opposites enable interpretability of pre-trained word embeddings. In *WWW*, pages 1548--1558.

- [Matsubara et al., 2017] Matsubara, Y., Sakurai, Y., Prakash, B. A., Li, L., and Faloutsos, C. (2017). Nonlinear Dynamics of Information Diffusion in Social Networks. *ACM Trans. Web.*
- [Mazloom et al., 2017] Mazloom, M., Hendriks, B., and Worring, M. (2017). Multimodal context-aware recommender for post popularity prediction in social media. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 236--244.
- [Mazloom et al., 2018] Mazloom, M., Pappi, I., and Worring, M. (2018). Category specific post popularity prediction. In *International Conference on Multimedia Modeling*, pages 594--607. Springer.
- [McCormick, 2016] McCormick, C. (2016). Word2vec tutorial - the skip-gram model.
- [Mcilroy et al., 2016] Mcilroy, S., Ali, N., and Hassan, A. E. (2016). Fresh apps: an empirical study of frequently-updated mobile apps in the google play store. *Empirical Software Engineering*, 21(3):1346--1370.
- [McKenzie et al., 2015] McKenzie, G., Adams, B., and Janowicz, K. (2015). Of oxen and birds: Is yik yak a useful new data source in the geosocial zoo or just another twitter? In *SIGSPATIAL*.
- [Meaker, 2022a] Meaker, M. (2022a). Russia blocks facebook and twitter in a propaganda standoff. *WIRED*.
- [Meaker, 2022b] Meaker, M. (2022b). Russia blocks facebook and twitter in a propaganda standoff. *WIRED*.
- [Mejova and Kourtellis, 2021] Mejova, Y. and Kourtellis, N. (2021). Youtubing at home: Media sharing behavior change as proxy for mobility around covid-19 lockdowns. In *13th ACM Web Science Conference 2021*, pages 272--281.
- [Meyer et al., 2021] Meyer, S., Windisch, T., Perl, A., Dzibela, D., Marzilger, R., Witt, N., Benzler, J., Kirchner, G., Feigl, T., and Mutschler, C. (2021). Contact tracing with the exposure notification framework in the german corona-warn-app. In *2021 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pages 1--8. IEEE.
- [Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.
- [Mikolov et al., 2013b] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013b). Efficient estimation of word representations in vector space. *arXiv*.
- [Mikolov et al., 2013c] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, Red Hook, NY, USA*. Curran Associates Inc.
- [Miller et al., 2016a] Miller, H., Thebault-Spieker, J., Chang, S., Johnson, I., Terveen, L., and Hecht, B. (2016a). "blissfully happy" or "ready to fight": Varying interpretations of emoji.
- [Miller et al., 2016b] Miller, H. J., Thebault-Spieker, J., Chang, S., Johnson, I., Terveen, L., and Hecht, B. (2016b). "Blissfully Happy" or "Ready toFight": Varying Interpretat. of Emoji. In *ICWSM*.

- [Mislove et al., 2008] Mislove, A., Koppula, H. S., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2008). Growth of the flickr social network. In *WOSN*.
- [Mislove et al., 2007] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and et al. (2007). Measurement and analysis of online social networks. In *IMC*.
- [Mittos et al., 2020] Mittos, A., Zannettou, S., Blackburn, J., and Cristofaro, E. D. (2020). Analyzing genetic testing discourse on the web through the lens of twitter, reddit, and 4chan. *ACM Transactions on the Web (TWEB)*, 14(4):1--38.
- [Mondal et al., 2020] Mondal, M., Correa, D., and Benevenuto, F. (2020). Anonymity effects: A large-scale dataset from an anonymous social media platform. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 69--74.
- [Monti et al., 2019] Monti, F., Frasca, F., Eynard, D., Mannion, D., and Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.
- [Morgans, 2017] Morgans, M. J. (2017). Freedom of speech, the war on terror, and what's youtube got to do with it: American censorship during times of military conflict. *Fed. Comm. LJ*, 69:145.
- [Movshovitz-Attias et al., 2013] Movshovitz-Attias, D., Movshovitz-Attias, Y., Steenkiste, P., and Faloutsos, C. (2013). Analysis of the reputation system and user contributions on a question answering website: Stackoverflow. In *2013 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2013)*, pages 886--893. IEEE.
- [Möller et al., 2011] Möller, S., Bang, C., Tamme, T., Vaalgamaa, M., and Weiss, B. (2011). From single-call to multi-call quality: a study on long-term quality integration in audio-visual speech communication. In *INTERSPEECH*.
- [Möller et al., 2013] Möller, S., Schmidt, S., and Beyer, J. (2013). Gaming taxonomy: An overview of concepts and evaluation methods for computer gaming QoE. In *QoMEX*.
- [Nazir et al., 2008] Nazir, A., Raza, S., and Chuah, C.-N. (2008). Unveiling facebook: a measurement study of social network based applications. In *IMC*.
- [Newton, 2019] Newton, C. (2019). The Trauma Floor - The secret lives of Facebook moderators in America.
- [@NIRVANA101, 2022] @NIRVANA101 (2022). Russian gbsb.
- [Novak et al., 2015] Novak, P. ., Smailovic, J., Sluban, B., and Mozetic, I. (2015). Sentiment of emojis. *PLoS one*.
- [Nowak et al., 2018] Nowak, P., Jüttner, K., and Baran, K. S. (2018). Posting content, collecting points, staying anonymous: An evaluation of jodel. In *International Conference on Social Computing and Social Media*, pages 67--86. Springer.
- [of Network Interference, 2022] of Network Interference, O. O. (2022). Data ooni. <https://ooni.org/data/>. [Online. Last accessed: 18-05-2022].
- [of the EU, 2022] of the EU, C. (2022). Eu imposes sanctions on state-owned outlets rt/russia today and sputnik's broadcasting in the eu. *Press Release*.
- [Osgood et al., 1957] Osgood, C., Suci, G., and Tannenbaum, P. (1957). *The Measurement of Meaning*. Illini Books, IB47. University of Illinois Press.

- [Oshikawa et al., 2018] Oshikawa, R., Qian, J., and Wang, W. Y. (2018). A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.
- [Otte and Rousseau, 2002] Otte, E. and Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science*.
- [P., 2021] P., V. (2021). Word2vec explained.
- [Panchenko et al., 2012] Panchenko, A., Lanze, F., and Engel, T. (2012). Improving performance and anonymity in the tor network. In *2012 IEEE 31st International Performance Computing and Communications Conference (IPCCC)*, pages 1--10. IEEE.
- [Panchenko et al., 2017] Panchenko, A., Mitseva, A., Henze, M., Lanze, F., Wehrle, K., and Engel, T. (2017). Analysis of fingerprinting techniques for tor hidden services. In *Proceedings of the 2017 on Workshop on Privacy in the Electronic Society*, pages 165--175.
- [Panigrahi et al., 2019] Panigrahi, A., Simhadri, H. V., and Bhattacharyya, C. (2019). Word2sense: sparse interpretable word embeddings. In *ACL*.
- [Papasavva et al., 2020] Papasavva, A., Zannettou, S., De Cristofaro, E., Stringhini, G., and Blackburn, J. (2020). Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. In *ICWSM*.
- [Pape et al., 2021] Pape, S., Harborth, D., and Kröger, J. L. (2021). Privacy concerns go hand in hand with lack of knowledge: The case of the german corona-warn-app. In *IFIP International Conference on ICT Systems Security and Privacy Protection*, pages 256--269. Springer.
- [Parekh et al., 2020] Parekh, D., Margolin, D., and Ruths, D. (2020). Comparing audience appreciation to fact-checking across political communities on reddit. In *12th ACM Conference on Web Science*, pages 144--154.
- [Park et al., 2022] Park, C. Y., Mendelsohn, J., Field, A., and Tsvetkov, Y. (2022). Voynaslov: A data set of russian social media activity during the 2022 ukraine-russia war. *arXiv preprint arXiv:2205.12382*.
- [Park et al., 2013] Park, J., Barash, V., Fink, C., and Cha, M. (2013). Emoticon style: Interpreting differences in emoticons across cultures. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 466--475.
- [Passonneau, 2004] Passonneau, R. J. (2004). Computing reliability for coreference annotation. In *International Conference on Language Resources and Evaluation (LREC)*.
- [Paul et al., 2011] Paul, S. A., Hong, L., and Chi, E. H. (2011). What is a question? crowd-sourcing tweet categorization. *CHI 2011*.
- [Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
- [Pérez-Rosas et al., 2018] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391--3401.
- [Peters et al., 2018] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of*

- the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227--2237, New Orleans, Louisiana. Association for Computational Linguistics.
- [Poese et al., 2011] Poese, I., Uhlig, S., Kaafar, M. A., Donnet, B., and Gueye, B. (2011). Ip geolocation databases: Unreliable? *SIGCOMM Comput. Commun. Rev.*, 41(2):53–56.
- [Price, 1942] Price, B. (1942). Governmental censorship in war-time. *American Political Science Review*, 36(5):837--849.
- [Price, 2018] Price, E. (2018). Should we yak back? information seeking among yik yak users on a university campus. *College & Research Libraries*, 79(2):200.
- [Pudipeddi et al., 2014] Pudipeddi, J. S., Akoglu, L., and Tong, H. (2014). User churn in focused question answering sites: characterizations and prediction. In *WWW*.
- [Purschke and Hovy, 2019] Purschke, C. and Hovy, D. (2019). Lörres, möppes, and the swiss.(re) discovering regional patterns in anonymous social media data. *Journal of Linguistic Geography*, 7(2):113--134.
- [Radford et al., 2018] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding with unsupervised learning.
- [Radford et al., 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.
- [Raman et al., 2019] Raman, A., Joglekar, S., Cristofaro, E. D., Sastry, N., and Tyson, G. (2019). Challenges in the decentralised web: The mastodon case. In *Proceedings of the Internet Measurement Conference*, pages 217--229.
- [Ramesh et al., 2020] Ramesh, R., Raman, R. S., Bernhard, M., Ongkowijaya, V., Evdokimov, L., Edmundson, A., Sprecher, S., Ikram, M., and Ensafi, R. (2020). Decentralized control: A case study of russia. In *Network and Distributed Systems Security (NDSS) Symposium 2020*.
- [Read, 2005] Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL student research workshop*, pages 43--48.
- [Reimers and Gurevych, 2019] Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- [Reis et al., 2019] Reis, J. C., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76--81.
- [Ren et al., 2020] Ren, H., Hu, W., and Leskovec, J. (2020). Query2box: Reasoning over knowledge graphs in vector space using box embeddings. *ICLR*.
- [Ren and Leskovec, 2020] Ren, H. and Leskovec, J. (2020). Beta embeddings for multi-hop logical reasoning in knowledge graphs. *Advances in Neural Information Processing Systems*, 33:19716--19726.
- [Reyaee and Ahmed, 2015] Reyaee, S. and Ahmed, A. (2015). Growth pattern of social media usage in arab gulf states: An analytical study. *Social Networking*.

- [Rieger et al., 2021] Rieger, D., Kümpel, A. S., Wich, M., Kiening, T., and Groh, G. (2021). Assessing the extent and types of hate speech in fringe communities: a case study of alt-right communities on 8chan, 4chan, and reddit. *Social Media+ Society*, 7(4):20563051211052906.
- [Robertson et al., 2021a] Robertson, A., Liza, F. F., Nguyen, D., McGillivray, B., and Hale, S. A. (2021a). Semantic journeys: Quantifying change in emoji meaning from 2012-2018. In *4th International Workshop on Emoji Understanding and Applications in Social Media 2021*, volume abs/2105.00846.
- [Robertson et al., 2018] Robertson, A., Magdy, W., and Goldwater, S. (2018). Self-representation on twitter using emoji skin color modifiers. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.
- [Robertson et al., 2020] Robertson, A., Magdy, W., and Goldwater, S. (2020). Emoji skin tone modifiers: Analyzing variation in usage on social media. *ACM Transactions on Social Computing*, 3(2):1--25.
- [Robertson et al., 2021b] Robertson, A., Magdy, W., and Goldwater, S. (2021b). Black or white but never neutral: How readers perceive identity from yellow or skin-toned emoji. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1--23.
- [Robison and Connell, 2017] Robison, M. and Connell, R. S. (2017). Harnessing yik yak for good: A study of students' anonymous library feedback. *Journal of Web Librarianship*, 11(1):35--55.
- [Romero et al.,] Romero, D. M., Meeder, B., and Kleinberg, J. Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter. WWW '11.
- [Rossi et al., 2001] Rossi, P. E., Gilula, Z., and Allenby, G. M. (2001). Overcoming scale usage heterogeneity. *Journal of the American Statistical Association*, 96(453).
- [Runge et al., 2014] Runge, J., Gao, P., Garcin, F., and Faltings, B. (2014). Churn prediction for high-value players in casual social games. In *IEEE Computational Intelligence and Games*.
- [Sakaki et al.,] Sakaki, T., Okazaki, M., and Matsuo, Y. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. WWW'10.
- [Sanlı and Lambiotte, 2015] Sanlı, C. and Lambiotte, R. (2015). Local variation of hashtag spike trains and popularity in twitter. *PLOS ONE*.
- [Saveski et al., 2016] Saveski, M., Chou, S., and Roy, D. (2016). Tracking the yak: An empirical study of yik yak. In *ICWSM*.
- [Scheible et al., 2020] Scheible, R., Thomczyk, F., Tippmann, P., Jaravine, V., and Boeker, M. (2020). GottBERT: a pure german language model.
- [Scheitle et al., 2018] Scheitle, Q., Hohlfeld, O., Gamba, J., Jelten, J., Zimmermann, T., Strowes, S. D., and Vallina-Rodriguez, N. (2018). A long way to the top: Significance, structure, and stability of internet top lists. In *IMC*, page 478-493.
- [Schiöberg et al., 2012] Schiöberg, D., Schneider, F., Schiöberg, H., Schmid, S., Uhlig, S., and Feldmann, A. (2012). Tracing the birth of an osn: Social graph and profile analysis in google+. In *WebSci*.

- [Schmidt, 2021] Schmidt, S. (2021). *Assessing the Quality of Experience of Cloud Gaming Services*. PhD thesis, TU Berlin.
- [Seidenschnur, 2021] Seidenschnur, T. (2021). A typology of social characters and various means of control: an analysis of communication during the early stages of the corona pandemic in germany. *European Societies*, 23(sup1):S923--S941.
- [Senel et al., 2018] Senel, L. K., Utlu, I., Yucesoy, V., Koc, A., and Cukur, T. (2018). Semantic structure and interpretability of word embeddings. In *IEEE/ACM TASLP*, number 10.
- [Shevtsov et al., 2022] Shevtsov, A., Tzagkarakis, C., Antonakaki, D., Pratikakis, P., and Ioannidis, S. (2022). Twitter dataset on the russo-ukrainian war. *arXiv preprint arXiv:2204.08530*.
- [Shu et al., 2019] Shu, K., Cui, L., Wang, S., Lee, D., and Liu, H. (2019). defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395--405.
- [Shu et al., 2018] Shu, K., Wang, S., and Liu, H. (2018). Understanding user profiles on social media for fake news detection. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 430--435. IEEE.
- [Shuvalova, 2022] Shuvalova, I. (2022). "moskal's," "separs," and "vatniks": The many faces of the enemy in the ukrainian satirical songs of the war in the donbas. *East/West: Journal of Ukrainian Studies*, 9(1):177--200.
- [Shwartz et al., 2017] Shwartz, V., Santus, E., and Schlechtweg, D. (2017). Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *EACL*, pages 65--75.
- [Silina, 2022] Silina, M. (2022). Russia's feminists are protesting the war and its propaganda with stickers, posters, performance and graffiti.
- [Silva et al., 2019] Silva, T. H., Viana, A. C., Benevenuto, F., Villas, L., Salles, J., Loureiro, A., and Quercia, D. (2019). Urban computing leveraging location-based social network data: a survey. *ACM Computing Surveys (CSUR)*, 52(1):1--39.
- [Silverman and Kao, 2022] Silverman, C. and Kao, J. (2022). Infamous russian troll farm appears to be source of anti-ukraine propaganda. *ProPublica*.
- [Simms et al., 2019] Simms, L., Zelazny, K., Williams, T., and Bernstein, L. (2019). Does the number of response options matter? psychometric perspectives using personality questionnaire data. *Psychological Assessment*, 31.
- [Simon and Rieder, 2021] Simon, J. and Rieder, G. (2021). Trusting the corona-warn-app? contemplations on trust and trustworthiness at the intersection of technology, politics and public debate. *European Journal of Communication*, 36(4):334--348.
- [Singer et al., 2014] Singer, P., Flöck, F., Meinhart, C., Zeitfogel, E., and Strohmaier, M. (2014). Evolution of reddit: from the front page of the internet to a self-referential community? In *Proceedings of the 23rd international conference on world wide web*, pages 517--522.
- [Singh et al., 2018] Singh, R., Dunna, A., and Gill, P. (2018). Characterizing the deployment and performance of multi-cdns. In *Proceedings of the Internet Measurement Conference 2018*, pages 168--174.

- [Singhal et al., 2019] Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., and Satoh, S. (2019). Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39--47. IEEE.
- [Slivar et al., 2015] Slivar, I., Dragozet, Z., and Skorin-Kapov, L. (2015). User customization of the gaminganywhere android mobile client interface. In *NetGames*.
- [Smolyak et al., 2021] Smolyak, A., Bonaccorsi, G., Flori, A., Pammolli, F., and Havlin, S. (2021). Effects of mobility restrictions during covid19 in italy. *Scientific reports*, 11(1):1--15.
- [Snow et al., 2008] Snow, R., O’connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast--but is it good? evaluating non-expert annotations for natural language tasks. In *EMNLP*.
- [Soliman et al., 2019] Soliman, A., Hafer, J., and Lemmerich, F. (2019). A characterization of political communities on reddit. In *Proceedings of the 30th ACM conference on hypertext and Social Media*, pages 259--263.
- [Stoddard, 2015] Stoddard, G. (2015). Popularity dynamics and intrinsic quality in reddit and hacker news. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 416--425.
- [Stutzman et al., 2013] Stutzman, F., Gross, R., and Acquisti, A. (2013). Silent listeners: The evolution of privacy and disclosure on facebook. *J. of privacy and confidentiality*.
- [Subramanian et al., 2018a] Subramanian, A., Pruthi, D., Jhamtani, H., Berg-Kirkpatrick, T., and Hovy, E. (2018a). Spine: Sparse interpretable neural embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
- [Subramanian et al., 2018b] Subramanian, A., Pruthi, D., Jhamtani, H., Berg-Kirkpatrick, T., and Hovy, E. (2018b). Spine: Sparse interpretable neural embeddings. In *AAAI*, volume 32.
- [Suznjevic et al., 2013] Suznjevic, M., Skorin-Kapov, L., and Matijasevic, M. (2013). The impact of user, system, and context factors on gaming qoe: A case study involving mmorpgs. In *QoMEX*.
- [Takahashi et al., 2017] Takahashi, K., Oishi, T., and Shimada, M. (2017). Is :-) smiling? Cross-cultural study on recognition of emoticon’s emotion. *Journal of Cross-Cultural Psychology*.
- [Tang et al., 2009] Tang, J., Sun, J., Wang, C., and Yang, Z. (2009). Social influence analysis in large-scale networks. In *SIGKDD*.
- [Tausczik and Pennebaker, 2010] Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24--54.
- [Tay et al., 2021] Tay, Y., Bahri, D., Metzler, D., Juan, D.-C., Zhao, Z., and Zheng, C. (2021). Synthesizer: Rethinking self-attention for transformer models. In *International conference on machine learning*, pages 10183--10192. PMLR.
- [Tenney et al., 2019] Tenney, I., Das, D., and Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593--4601, Florence, Italy. Association for Computational Linguistics.

- [Thomas et al., 2012] Thomas, K., Grier, C., and Paxson, V. (2012). Adapting social spam infrastructure for political censorship. In *5th USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET 12)*.
- [Thukral et al., 2018] Thukral, S., Meisheri, H., Kataria, T., Agarwal, A., Verma, I., Chatterjee, A., and Dey, L. (2018). Analyzing behavioral trends in community driven discussion platforms like reddit. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 662--669. IEEE.
- [Tóth et al., 2020] Tóth, L., Nagy, B., Gyimóthy, T., and Vidács, L. (2020). Why will my question be closed? nlp-based pre-submission predictions of question closing reasons on stack overflow. In *2020 IEEE/ACM 42nd International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*, pages 45--48. IEEE.
- [Tourani et al., 2015] Tourani, R., Misra, S., Kliewer, J., Orteguel, S., and Mick, T. (2015). Catch me if you can: A practical framework to evade censorship in information-centric networks. In *Proceedings of the 2nd ACM Conference on Information-Centric Networking*, pages 167--176.
- [Troianovski, 2022] Troianovski, A. (2022). Russia takes censorship to new extremes, stifling war coverage.
- [Troncoso et al., 2020] Troncoso, C., Payer, M., Hubaux, J.-P., Salathé, M., Larus, J., Bugnion, E., Lueks, W., Stadler, T., Pyrgelis, A., Antonioli, D., Barman, L., Chatel, S., Paterson, K., apkun, S., Basin, D., Beutel, J., Jackson, D., Roeschlin, M., Leu, P., Preneel, B., Smart, N., Abidin, A., Gürses, S., Veale, M., Cremers, C., Backes, M., Tippenhauer, N. O., Binns, R., Cattuto, C., Barrat, A., Fiore, D., Barbosa, M., Oliveira, R., and Pereira, J. (2020). Decentralized privacy-preserving proximity tracing. arXiv cs.CR/2005.12273 <https://arxiv.org/abs/2005.12273>.
- [Trujillo and Cresci, 2022] Trujillo, A. and Cresci, S. (2022). Make reddit great again: assessing community effects of moderation interventions on r/the_donald. *arXiv preprint arXiv:2201.06455*.
- [Tullis and Albert, 2008] Tullis, T. and Albert, W. (2008). *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- [Ugander et al., 2011] Ugander, J., Karrer, B., Backstrom, L., and Marlow, C. (2011). The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*.
- [Unicode Consortium, 2016] Unicode Consortium, T. (2016). Emoji 4.0.
- [Unwala and Ghori, 2016] Unwala, A. and Ghori, S. (2016). Brandishing the cybered bear: Information war and the russia-ukraine conflict. *Military Cyber Affairs*, 1(1):7.
- [Van Mieghem et al., 2011] Van Mieghem, P., Blenn, N., and Doerr, C. (2011). Lognormal distribution in the digg online social network. *The European Physical Journal B*, 83(2):251.
- [Vargo and Hopp, 2018] Vargo, C. and Hopp, T. (2018). Is yik yak a platform for political communication?: Exploring college students' communication on an emergent social media platform. In *Digital Discussions*, pages 134--155. Routledge.
- [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

- [Vaterlaus, 2017] Vaterlaus, J. M. (2017). Yik yak: An exploratory study of college student uses and gratifications. *Bulletin of Science, Technology & Society*, 37(1):23--33.
- [Vaterlaus et al., 2016] Vaterlaus, J. M., Barnett, K., Roche, C., and Young, J. A. (2016). "snapchat is more personal": An expl. study on snapchat behaviors and young adult interspersed relationships. *Computers in Human Behavior*.
- [Verkamp and Gupta, 2012] Verkamp, J.-P. and Gupta, M. (2012). Inferring mechanics of web censorship around the world. In *FOCI*.
- [Viennot et al., 2014] Viennot, N., Garcia, E., and Nieh, J. (2014). A measurement study of google play. In *The 2014 ACM international conference on Measurement and modeling of computer systems*, pages 221--233.
- [Vlahovic et al., 2021] Vlahovic, S., Suznjevic, M., Pavlin-Bernardic, N., and Skorin-Kapov, L. (2021). The effect of vr gaming on discomfort, cybersickness, and reaction time. In *QoMEX*.
- [Vlahovic et al., 2019] Vlahovic, S., Suznjevic, M., and Skorin-Kapov, L. (2019). The impact of network latency on gaming qoe for an fps vr game. In *QoMEX*.
- [Volkova and Bell, 2016] Volkova, S. and Bell, E. (2016). Account deletion prediction on runet: A case study of suspicious twitter accounts active during the russian-ukrainian crisis. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 1--6.
- [Wai-Ho Au et al., 2003] Wai-Ho Au, Chan, K. C. C., and Xin Yao (2003). A novel evolutionary data mining algorithm with applications to churn prediction.
- [Walther and Kaisser,] Walther, M. and Kaisser, M. Geo-spatial Event Detection in the Twitter Stream. *Advances in Information Retrieval'13*.
- [Wang et al., 2014] Wang, G., Wang, B., Wang, T., Nika, A., Zheng, H., and et al. (2014). Whispers in the dark: analysis of an anonymous social network. In *IMC*.
- [Wang et al., 2016] Wang, G., Zhang, X., Tang, S., Zheng, H., and Zhao, B. Y. (2016). Un-supervised clickstream clustering for user behavior analysis. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 225--236.
- [Wang et al., 2021] Wang, J., Wang, K.-C., Rudzicz, F., and Brudno, M. (2021). Grad2task: Improved few-shot text classification using gradients for task representation. In *NeurIPS*.
- [Wang et al., 2020] Wang, S., Li, B. Z., Khabsa, M., Fang, H., and Ma, H. (2020). Linformer: Self-attention with linear complexity. *arXiv e-prints*, pages arXiv--2006.
- [Wang et al., 2013] Wang, S., Lo, D., and Jiang, L. (2013). An empirical study on developer interactions in stackoverflow. In *SAC*.
- [Wani et al., 2021] Wani, A., Joshi, I., Khandve, S., Wagh, V., and Joshi, R. (2021). Evaluating deep learning approaches for covid19 fake news detection. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*, pages 153--163. Springer.
- [Wasserman et al., 1994] Wasserman, S., Faust, K., et al. (1994). *Social network analysis: Methods and applications*.
- [Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *nature*, 393(6684):440--442.

- [Weinzierl et al., 2021] Weinzierl, M. A., Hopfer, S., and Harabagiu, S. M. (2021). Misinformation adoption or rejection in the era of covid-19. In *ICWSM*, pages 787--795.
- [Weng and Lee, 2011] Weng, J. and Lee, B.-S. (2011). Event Detection in Twitter.
- [Weninger et al., 2015] Weninger, T., Johnston, T. J., and Glenski, M. (2015). Random voting effects in social-digital spaces: A case study of reddit post submissions. In *Proceedings of the 26th ACM conference on hypertext & social media*, pages 293--297.
- [West, 2017] West, S. (2017). Yik yak and the knowledge community. *Communication Design Quarterly Review*, 4(2b):11--21.
- [Whittaker and Kowalski, 2015] Whittaker, E. and Kowalski, R. M. (2015). Cyberbullying via social media. *Journal of School Violence*.
- [Wijeratne et al., 2017a] Wijeratne, S., Balasuriya, L., Sheth, A., and Doran, D. (2017a). Emojinet: An open service and api for emoji sense discovery. In *ICWSM*.
- [Wijeratne et al., 2017b] Wijeratne, S., Balasuriya, L., Sheth, A., and Doran, D. (2017b). A semantics-based measure of emoji similarity. In *WIMS*.
- [Wikipedia contributors, 2022] Wikipedia contributors (2022). Mensch ärgere dich nicht --- Wikipedia, the free encyclopedia.
- [Williams and Mahmoud, 2018] Williams, G. and Mahmoud, A. (2018). Modeling user concerns in the app store: A case study on the rise and fall of yik yak. In *2018 IEEE 26th international requirements engineering conference (rE)*, pages 64--75. IEEE.
- [Wilson et al., 2009] Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P., and Zhao, B. Y. (2009). User interactions in social networks and their implications. In *EuroSys*.
- [Wiseman and Gould, 2018] Wiseman, S. and Gould, S. J. (2018). Repurposing emoji for personalised communication: Why (pizzaemoji) means "i love you". In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1--10.
- [Woo and Chen, 2016] Woo, J. and Chen, H. (2016). Epidemic model for information diffusion in web forums: experiments in marketing exchange and political dialog. *Springer-Plus*.
- [(WSJ), 2022] (WSJ), K. D. (2022). Tripadvisor, google maps suspend reviews of some russian listings.
- [Wu et al., 2017] Wu, Y., Minkus, T., and Ross, K. W. (2017). Taking the pulse of us college campuses with location-based anonymous mobile apps. In *TIST*. ACM.
- [Xu et al.,] Xu, W. W., Park, J. Y., Kim, J. Y., and Park, H. W. Networked Cultural Diffusion and Creation on YouTube: An Analysis of YouTube Memes.
- [Xue et al., 2021] Xue, D., Ramesh, R., Evdokimov, L., Viktorov, A., Jain, A., Wustrow, E., Basso, S., and Ensafi, R. (2021). Throttling twitter: an emerging censorship technique in russia. In *Proceedings of the 21st ACM Internet Measurement Conference*, pages 435--443.
- [Xue et al., 2016] Xue, M., Ballard, C., Liu, K., Nemelka, C., Wu, Y., Ross, K., and Qian, H. (2016). You can yak but you can't hide: Localizing anonymous social network users. In *Proceedings of the 2016 Internet Measurement Conference, IMC '16*, page 25-31, New York, NY, USA. Association for Computing Machinery.

- [Xynou and Filastò, 2022] Xynou, M. and Filastò, A. (2022). New blocks emerge in russia amid war in ukraine: An ooni network measurement analysis. *OONI*.
- [Yan et al., 2013] Yan, Q., Wu, L., Liu, C., and Li, X. (2013). Information propagation in online social network based on human dynamics.
- [Yang et al., 2018] Yang, C., Shi, X., Jie, L., and Han, J. (2018). I know you'll be back: Interpretable new user clustering and churn prediction on a mobile social application. In *SIGKDD*.
- [Yang et al., 2017] Yang, D., Kraut, R., and Levine, J. M. (2017). Commitment of newcomers and old-timers to online health support communities. In *CHI*.
- [Yang et al., 2014] Yang, J., Tao, K., Bozzon, A., and Houben, G.-J. (2014). Sparrows and owls: Characterisation of expert behaviour in stackoverflow. In *International conference on user modeling, adaptation, and personalization*, pages 266--277. Springer.
- [Yin et al., 2011] Yin, Z., Cao, L., Han, J., Zhai, C., and Huang, T. (2011). Geographical topic discovery and comparison. In *WWW*.
- [@YourAnonNews, 2022a] @YourAnonNews (2022a). Go to google maps. go to russia. find a restaurant or business and write a review. when you write the review explain what is happening in ukraine. idea via @konrad03249040.
- [@YourAnonNews, 2022b] @YourAnonNews (2022b). Roskomnadzor blocked access to google news.
- [Yu et al., 2015] Yu, L., Li, X., Tang, L., Zhang, Z., and Kou, G. (2015). Social credit: a comprehensive literature review. *Financial Innovation*, 1(1):1--18.
- [Zaki, 2018] Zaki, M. (2018). *Gemeinschaftsbildung trotz Anonymität?* PhD thesis.
- [Zampieri et al., 2019] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75--86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- [Zannettou et al., 2018] Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G., and Suarez-Tangil, G. (2018). On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, pages 188--202.
- [Zannettou et al., 2017] Zannettou, S., Caulfield, T., De Cristofaro, E., and et al. (2017). The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *IMC*.
- [Zelenkauskaitė et al., 2021] Zelenkauskaitė, A., Toivanen, P., Huhtamäki, J., and Valaskivi, K. (2021). Shades of hatred online: 4chan duplicate circulation surge during hybrid media events. *First Monday*.
- [Zhang et al., 2017] Zhang, J., Hamilton, W. L., Danescu-Niculescu-Mizil, C., Jurafsky, D., and Leskovec, J. (2017). Community identity and user engagement in a multi-community landscape. *CoRR*.
- [Zhang et al., 2020] Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., and Artzi, Y. (2020). Revisiting few-sample bert fine-tuning. *arXiv preprint arXiv:2006.05987*.

- [Zhang et al., 2022] Zhang, Y., Herring, S., and Gan, S. (2022). Graphicon evolution on the chinese social media platform bilibili. In *Proceedings of the The Fifth International Workshop on Emoji Understanding and Applications in Social Media*, pages 75--85.
- [Zhang et al., 2018] Zhang, Z., Chen, T., Zhou, Z., Li, J., and Luo, J. (2018). How to become instagram famous: Post popularity prediction with dual-attention. In *2018 IEEE international conference on big data (big data)*, pages 2383--2392. IEEE.
- [Zhao et al., 2018] Zhao, P., Jia, J., An, Y., Liang, J., Xie, L., and Luo, J. (2018). Analyzing and predicting emoji usages in social media. In *WWW*.
- [Zhou et al., 2016] Zhou, L., Wang, W., and Chen, K. (2016). Tweet properly: Analyzing deleted tweets to understand and identify regrettable ones. In *Proceedings of the 25th International Conference on World Wide Web*, pages 603--612. International World Wide Web Conferences Steering Committee.
- [Zhou and Zafarani, 2020] Zhou, X. and Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1--40.
- [Zhu et al., 2022] Zhu, Y., Haq, E.-u., Lee, L.-H., Tyson, G., and Hui, P. (2022). A reddit dataset for the russo-ukrainian conflict in 2022. *arXiv preprint arXiv:2206.05107*.
- [Zhu et al., 2021] Zhu, Z., He, Y., Zhao, X., Zhang, Y., Wang, J., and Caverlee, J. (2021). Popularity-opportunity bias in collaborative filtering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 85--93.
- [Zignani et al., 2018] Zignani, M., Gaito, S., and Rossi, G. P. (2018). Follow the "mastodon": Structure and evolution of a decentralized online social network. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, pages 541--550.
- [Zignani et al., 2019] Zignani, M., Quadri, C., Gaito, S., Cherifi, H., and Rossi, G. P. (2019). The footprints of a "mastodon": How a decentralized architecture influences online social relationships. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 472--477. IEEE.
- [Zohourian et al., 2018] Zohourian, A., Sajedi, H., and Yavary, A. (2018). Popularity prediction of images and videos on instagram. In *2018 4th International Conference on Web Research (ICWR)*, pages 111--117. IEEE.
- [África Periañez et al., 2016] África Periañez, Saas, A., Guitart, A., and Magne, C. (2016). Churn prediction in mobile social games: Towards a complete assessment using survival ensembles. In *IEEE DSAA*.
- [Óskarsdóttir et al., 2020] Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., and et al. (2020). Social network analytics for churn prediction in telco: Model building, evaluation and network architecture.