Anonymous Hyperlocal Communities: What do they talk about?

Jens Helge Reelfs
reelfs@b-tu.de
Brandenburg University of
Technology

Oliver Hohlfeld
hohlfeld@b-tu.de
Brandenburg University of
Technology

Niklas Henckell
niklas@jodel.com
The Jodel Venture GmbH

ABSTRACT

In this paper, 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.

CCS CONCEPTS

• **Information systems** → **Social networks**; **Crowdsourcing**; *Location based services*; • **Security and privacy** → *Pseudonymity, anonymity and untraceability*;

1 INTRODUCTION

Anonymity on internet platforms is often controversially discussed between *i)* enabling freedom of speech and *ii)* enabling toxic environments [14]. 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 [5]. The anonymous location-based app YikYak can be characterized with entertaining and informational contents leveraging self-supervised learning [3], while others find evidence for flirting & dating [26] 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 [22].

This opens the question if anonymity yields to a richer content spectrum, especially in more conservative regimes. In the case of Saudi Arabia, [7] report on interviews with KSA women about boundaries and new freedoms, granted through the Internet—rendering anonymous platforms specifically interesting. Alsanea protrays the Saudi life in the 2007's novel *Girls of Riyadh* [1] 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 2017. Recent research empirically shows the Saudi Arabian user base to be active in creating and reacting to content over voting—contrasting a Western counterpart [19]. We take this as motivation to study social media posts in the KSA as intra-country and platform study.

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 [19]. **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 & 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.
Methodology. 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. Our **contributions** are as follows.

- 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
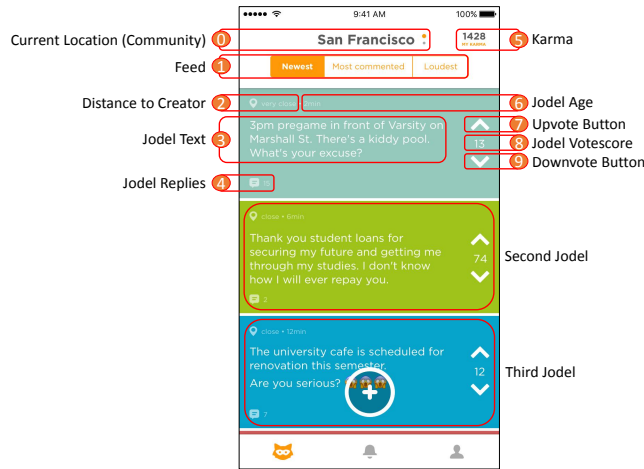
Figure 1: Jodel iOS mobile application.

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 & how users appreciate content through voting.

- We develop a second schema to classify the top 1k hashtags.

2 JODEL - HYPERLOCAL MESSAGING APP

Jodel is a mobile-only messaging application which we show in Figure 1. It is *location-based* and establishes local communities relative to the users' location ❶. Within these communities, users can *anonymously* post both images and textual content of up to 250 characters length ❸ (i.e., microblogging) and reply to posts forming discussion threads ❹. Posted content is referred to as "Jodels", colored randomly ❸. Posts are only displayed to other users within close (up to ≈20km) geographic proximity ❷. Unlike other networks (e.g., Facebook, Twitter, etc.), communication to remote users outside the local community is *not* possible. This biases all posts to having relevance to the local community—a very different mode of communication that opens questions on *what* is being posted and if content differs between communities.

Further, all communication is *anonymous* to other users since the app does not display any handle or user-related information. Only *within* a single discussion thread, Jodel enumerates users to enable referencing other users. The app builds upon three main feeds showing ❶: i) most *recent* threads, ii) most *discussed* threads, and iii) *loudest* showing threads with highest voting scores (likes).

Being a key success parameter for an anonymous messaging app, Jodel employs a community-driven filtering and moderation scheme to avoid harmful or abusive content. In Jodel, content filtering relies on a distributed voting scheme in which every user can increase or decrease a post's vote score by up- (+1) ❼ or downvoting (-1) ❾ (similar to StackOverflow). Posts reaching a cumulative vote score ❽ below a negative threshold (e.g., -5) are no longer displayed. Depending on the number of vote-contributions, this scheme filters out adverse content while also potentially preferring mainstream

content. To increase user engagement w.r.t. posting and voting, Jodel uses lightweight gamification by awarding *Karma* points ❺.

2.1 Data Set Description and Ethics

Meta Data. The Jodel operator provided us with excerpts of *anonymized* content data of their network. The obtained data contains *metadata* for 469 M posts created within the KSA by 1.3 M users. It spans over multiple years from the beginning of the network in 2014 up to August 2017. Due to being captured for operational purposes during the rollout, more recent data is not available to us. To protect user privacy on raw-data level, it is limited to metadata only without textual content, and anonymized user IDs. The structure of this dataset includes 3 categories: interactions, content, and users.

Defining "a" community is not possible on Jodel given that the app always displays content relative to each user's location and thus usually 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.

Smaller sub-dataset with textual posts. For *content* classification, we have been provided with a random sample of 15k threads throughout the country. Furthermore, we have access to used hashtags counts from a #1.12M random subsample. Users generally post content consciously into the public application domain. Albeit against the Jodel Terms of Service that forbid posting personal information, i.e., identifying individuals, posts may still contain such information or even references to other social media platforms directly opening up the possibility for identification of individuals. In liaison with Jodel and our university's Chief Data Officer, our crowdsourcing experiments *cannot* be conducted by any personnel not associated to our university; i.e., to protect users' privacy, we cannot use typical crowdsourcing platforms, such as Amazon Mechanical Turk or Microworkers, but must employ coders under direct governance. We store data strictly encrypted on a dedicated firewalled server; (selected) content is only accessible to authenticated internal users (coders). We generally inform and synchronize with the Jodel operator on analysis we perform on their data.

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) I of a post. *ii)* What topics Θ are presented in a post. For each the intents I and the topics Θ, multiple labels can be attributed by human annotators to a single post.

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).

Abbreviation	Class Information	#Annot	pdf
$\mathcal{I}$ = **Intents**	#Coders=2; $\alpha_M^{\mathcal{I}} = 0.74$, *substantial*		
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
Θ = **Topics**	#Coders=2; $\alpha_M^{\Theta} = 0.64$, *substantial*		
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 1: Annotation Schema & 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 $\alpha_M^{\Theta} = 0.74$.

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 [10] taxonomy, derived from semi-structured interviews with social media users. Table 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.

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 [5, 16], which we iteratively refine and adapt to content shared via Jodel KSA. We show the list of topics Θ in Table 1. We also opted for multi-labelling.

Iterative schema development. We based the initial version of the schema on prior work ([10] for the intents and [5, 16, 26] 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≈1.6% of outvoted (disliked) or otherwise blocked posts and replies. Though this figure is higher for in-app prominently displayed posts with p≈7.3%, only a minor fraction of these posts has been blocked by escalated moderation and flagging with p≈2.3%. This indicates that applied moderation works; especially due to

Figure 2: 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.

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.

Implementation. We realized the crowdsourcing system from scratch as a web application in PHP, shown in Figure 2. 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.

4 CLASSIFICATION CAMPAIGN

4.1 Study Design

We apply our content classification schema (§ 3) to annotate Jodel posts in the KSA (see § 2.1 for the data set description). 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 & 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 [21]. 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.

4.2 Campaign & 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 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.
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* [25] (identified as *seeking interaction* in the later results).
Coder Agreement. We measure our expert classifier interrater-agreement with Krippendorf's alpha [11]. 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 2 using various distance metrics: Binary, Jaccard, and MASI [15].

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 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 [11] suggests

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

Table 2: Classification agreement for multiple iterations on thread starting posts; For intents I & 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 (I) than topics (Θ).

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.
Multi-Labels. We analyzed the amount of classified posts with multiple intents I or topics Θ. While 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. 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* × *FitnessHealth*, *EduWork*, *FashionBeauty*, or *EntertCulture* × *AnimalsNature*.
Overall Confusion. In Figure 4a and Figure 4b, 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* I, in Figure 4a, 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* × *EntObserv*, or *DistRelComp* × *SocVentComp*; yet we must take note of the rest.

As for *topics* Θ, in Figure 4b, 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*.
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 ($\alpha_M^I = 0.75$, $\alpha_M^\Theta = 0.64$) across #posts=1,789 (#Coders=2) and conclude that our proposed schema has sufficient quality for further evaluation.*

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.

5.1 Countrywide Perspective on Jodel Content

We begin with analyzing the content classification for the countrywide overall annotations, before we study differences between two cities. First, we discuss Table 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 3. We complement this heatmap by discussing topic distributions across intents next (not shown).

Intents $\mathcal{I}$. The dominant intents are (see Table 1): $\mathcal{I}$-*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 $\mathcal{I}$-*Entobserv $\times$ GenEntert*, or apparent confusions along *Self*.

Topics Θ. Albeit slightly weaker annotator agreement, the discussed topics largely revolve around Θ-*PeopleRelation* accounting for $\Sigma 25\%$ annotations, which is also our most confused category. We find other popular themes in Θ-*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.

Intents $\mathcal{I} \times \Theta$ Topics. We identify specific hotspots of interests by combining $\mathcal{I} \times \Theta$ as a heatmap in Figure 3. Jodel is mostly being used out of the intent of $\mathcal{I}$-*Seek*ing Information & Interaction ($\Sigma 34\%$) for Θ-*PeopleRelation* ($p[\Theta|\mathcal{I}]$=23%) and *EntertCulture* ($p[\Theta|\mathcal{I}]$=12%), closely followed by others. Likewise, finding $\mathcal{I}$-*Self* ($\Sigma 15\%$) Expression across the board, users again focus on Θ-*PeopleRelation* ($p[\Theta|\mathcal{I}]$=23%) and *Self* Expression ($p[\Theta|\mathcal{I}]$=14%). Out of $\mathcal{I}$-*GenEntert* ($\Sigma 10\%$), we want to highlight Θ-*BeliefPol* ($p[\Theta|\mathcal{I}]$=57%). Whereas $\mathcal{I}$-*SocVentComp* ($\Sigma 10\%$) almost naturally goes along with the topic Θ-*PeopleRelation* ($p[\Theta|\mathcal{I}]$=41%), $\mathcal{I}$-*DistRelComp* ($\Sigma 9\%$) aligns with Θ-*Self* ($p[\Theta|\mathcal{I}]$=31%) and *PeopleRelation* ($p[\Theta|\mathcal{I}]$=22%).

Anonymity. From our analysis, it comes apparent that most content posted on Jodel indeed is related to users' intent for $\mathcal{I}$-*Seek*ing Information & Interaction, and *Self* Exression accounting for 50% of all annotations with strong trends towards the topics Θ-*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 $\mathcal{I}$-*SocVentComp* and *DistRelComp* within the same topical regime. Unfortunately, crowdsourcing a well-suited anonymity-sensitivity score relying on many (n=89) coders [23] is not possible in our case. Foreshadowing our categorization on hashtags in (§ 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.

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 $\mathcal{I}$-*Seek*ing Interaction & Information actually refers to local matchmaking, events, local services, or educational institutions. By focusing on the platforms content and driving factors first,

Figure 3: Overall annotation counts, intents $\mathcal{I} \times \Theta$ topics.

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 § 5.2 comparing intents & topics in Jeddah $\times$ Riyadh, and provide deeper insights via hashtags in § 6.

Findings. *($\mathcal{I}$) 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).*

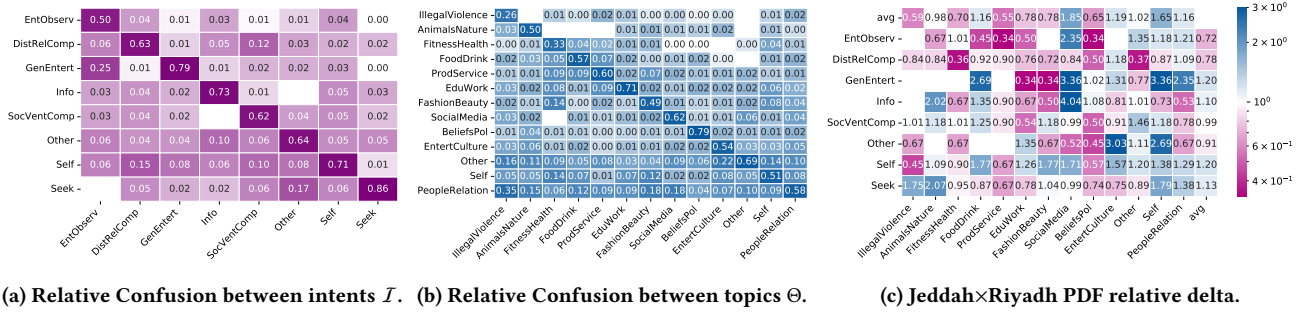## 5.2 City-Level Perspective on Jodel Content

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.

Jeddah $\times$ 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, α_M^{Θ}=0.62, Riyadh: $\alpha_M^{\mathcal{I}}$=0.75, α_M^{Θ}=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={Jeddah-Riyadh: 0.75, Jeddah-random: 0.44, 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 4c. 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.

(a) Relative Confusion between intents $\mathcal{I}$.   (b) Relative Confusion between topics Θ.   (c) Jeddah×Riyadh PDF relative delta.

Figure 4: (Left, middle:) Relative Confusion, column wise normalization, sums to 100%. De-Biased multiset join by weighting factor $(n \times m)^{-1}$. Log color scale. (Right:) Community differences. 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.

Content which is more popular in Jeddah (blue). In Jeddah, we find more $\mathcal{I}$-*Self* Expression ▹(1.20; Σ18%), *GenEntert*ainment ▹(1.20; Σ10%) and individuals *Seek*ing Information & Interactions ▹(1.23; Σ49%) in 77% $\mathcal{I}$-annotations. Increased topic figures are rather personal and casual around Θ-*SocialMedia* ▹(1.85; Σ9%), *Self* (1.65; Σ13%), *EnterCultule* ▹(1.19; Σ11%), and *PeopleRelation* ▹(1.16; Σ29%) 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; Σ3%) and *DistRelComp* ▹(0.78; Σ14%), 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; Σ6%), *IllegalViolence* ▹(0.59; Σ2%), *BeliefsPol* ▹(0.65; Σ11%), *FitnessHealth* ▹(0.70; Σ4%), *EduWork* ▹(0.78; Σ6%), and *FashionBeatuy* ▹(0.78; Σ7%). They account for another 36% of overall Θ-annotations.

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 topics, whereas the Jeddah community focuses more on $\mathcal{I}$ -(Self Expression, Entertainment) along Θ-(Social Media, People & Relationships).*

5.3 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- & downvotes, both accumulated called Karma, and posts & repliers per annotated thread. From these figures, we derive two scores: *i) Conversationness* [19] 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 (#uvotes, #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 & votes counts in Figure 5a; whereas we show CDFs of both metrics *i)* & *ii)* in Figure 5. To be brief, we only highlight and discuss distributional outliers.

Replies. The amount of replies is generally heavy-tailed as shown in Figure 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 ↑25 for 97%), whereas others already reach up to ↑25 replies for between 80% to 90%. Within Θ-topic distributions, the least replies can be expected for *BeliefsPol* (up to ↑25 for already 90%), appearing far less discussed than *IllegalViolence*, *PeopleRelation* & *FashionBeauty* (↑25 ≈70-80%).

Coversationness. 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 (↑x=1.0) as overall shown in Figure 5b. Most heterogeneous discussions appear for $\mathcal{I}$-*Seek Info* and *SocVentComp* (cut-offs at ↑0.75 to ↑0.8), versus more homogeneous conversations in *GenEntert* & *EntertObs* (↑0.43, ↑0.58). Θ-*BeliefsPol* (↑0.52) remains most homogeneously discussed at two replies per participant on average.

Votes. Similar to replies, vote count distributions for Intents and topics remain heavy-tailed as shown overall series in Figure 5a.

Karma. Karma describes the accumulated vote score between up- and downvotes. As can be seen in Figure 5a, Jodel Karma is long-tailed to positive votes, whereas disliked posts naturally fall off around the post-remove-threshold [19]. We identify higher scores, indicating appreciation, in $\mathcal{I}$-*GenEntert* (↑10 for only up to 60%) and Θ-*BeliefsPol* (↑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.

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 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 (≈15% of all posts have score 1.0). Whereas for most $\mathcal{I}$ Intents cut off the S-shape between ↓0.02 to ↓0.1 on the lower end, they reach between ↑0.61 to ↑0.79 at the upper end; $\mathcal{I}$-(*Other*, *Info*) being outliers at (0.82, 0.88). This general observation

(a) Replies, Votes & Karma   **(b) consensus & Conversat.**
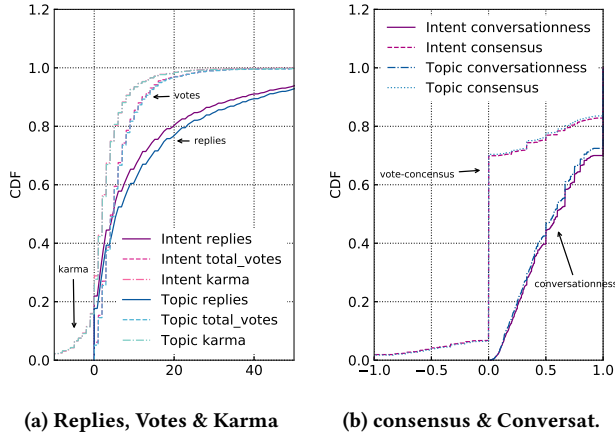
Figure 5: Received Platform Reactions. Vote-Related: (left) Votes, Karma, (right) Vote-consensus. Reply-Related: (left) Replies, (right) Conversationness.

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.
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 $\triangle := (\mathcal{I}$-GenEntert $\times$ Θ-BeliefsPol) As for conversationness, we find most homogeneous discussions especially within $\triangle \cup \mathcal{I}$-EntertObs. The community often appreciates $\triangle$-content with high confidence. In conclusion, $\triangle$ 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).*

5.4 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 [2], 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 [13, 24]. Thus, we believe that more data may improve results—as has our

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 3: Classification of the top 1015 hashtags.

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.

6 HASHTAGS

After a deep dive into quantitative insights and empirically peeking into community reactions upon content, we identified driving factors intents & discussed topics; yet we realized that it was missing crucial qualitative aspects of our annotators' experience.

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 [6]. 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 [23] 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 3, 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 matchmaking, *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 & Relations*, or combinations of both topics with *Distress Release* and *Social Venting*. Albeit being very broad, most defined categories may be sensitive to anonymity giving *Debates & Opinions* the 18+ mark; except *Upvote Campaigns*[1] and *Other*. We largely observe posts under such hashtags sharing personal experiences or confessions; other *DatingFlag*ged topics often discuss love, sex, marriage, playful matchmaking, or games.

---

[1]Invitation for gathering *Karma*, a lightweight in-app gamification.

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.*

6.2 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 neglection, social pressure, and others out of manifold reasons.

The Story of Dating & 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* accouting 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.

Findings. *Our observations are no exception to shown results on Whisper [23]. 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 & opinions—one main reason using the application as concluded from interviews [10].*

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.

7 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 [14]. 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 [4], or Twitter [27]. The latter was used to model information diffusion [9], also conducted on Jodel itself [17]. 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 [5, 23], and very similar app YikYak of which we highlight a few examples. A wide range in methodology can be found in empirical [20], but also qualitative studies [12].

To quantify discussed topics, works leveraged (self)-supervised models [3] finding platform content to be rather ephemeral in an impersonal environment, which is confirmed via survey study in [12, 22]. Nonetheless, another work revealed "Personal Admission", "Observation" or "Information/Advice" resembling popular content [8]. Others conducted crowdsourcing at large scale [26] 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. While the Jodel app's usage has surged in the Middle East (KSA), establishing a loyal user base [18], it is characterized by a hugely different user behavior compared to a Western counterpart [19]. Within this context, 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.

8 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 [5]. 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" [1, pp. 313-314].

REFERENCES

[1] Rajaa Alsanea. 2007. *Girls of Riyadh.* Fig Tree.
[2] Wissam Antoun, Fady Baly, and Hazem Hajj. 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.*
[3] Erik W Black, Kelsey Mezzina, and Lindsay A Thompson. 2016. Anonymous social media–Understanding the content and context of Yik Yak. In *SIGCHI*. Elsevier.
[4] Meeyoung Cha, Alan Mislove, and Krishna P Gummadi. 2009. A measurement-driven analysis of information propagation in the flickr social network. In *WWW*.
[5] Denzil Correa, Leandro Araújo Silva, Mainack Mondal, Fabrício Benevenuto, and Krishna P Gummadi. 2015. The Many Shades of Anonymity: Characterizing Anonymous Social Media Content. In *ICWSM*.
[6] Paolo Ferragina, Francesco Piccinno, and Roberto Santoro. 2015. On analyzing hashtags in twitter. In *ICWSM*.
[7] Hala Guta and Magdalena Karolak. 2015. Veiling and blogging: Social media as sites of identity negotiation and expression among Saudi women. *Journal of International Women's Studies* (2015).
[8] Matthew Heston and Jeremy Birnholtz. 2016. (In)visible Cities: An Exploration of Social Identity, Anonymity and Location-Based Filtering on Yik Yak. In *iConference*. iSchools.
[9] Krishna Y Kamath, James Caverlee, Kyumin Lee, and et al. 2013. Spatio-temporal dynamics of online memes: a study of geo-tagged tweets. In *WWW*.
[10] Ruogu Kang, Laura Dabbish, and Katherine Sutton. 2016. Strangers on Your Phone: Why People Use Anonymous Communication Applications. In *CSCW*.
[11] Klaus Krippendorff. 2012. *Content analysis: An introduction to its methodology.* Sage publications.
[12] Joon-Suk Lee, Seungwon Yang, Amanda L Munson, and Lusene Donzo. 2017. What People Do on Yik Yak: Analyzing Anonymous Microblogging User Behaviors. In *SCSM*. Springer, 416–428.
[13] Qiaoyang Luo, Lingqiao Liu, Yuhao Lin, and Wei Zhang. 2021. Don't Miss the Labels: Label-semantic Augmented Meta-Learner for Few-Shot Text Classification. In *ACL-IJCNLP*.
[14] Antonis Papasavva, Savvas Zannettou, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2020. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. In *ICWSM*.
[15] Rebecca J. Passonneau. 2004. Computing Reliability for Coreference Annotation. In *International Conference on Language Resources and Evaluation (LREC)*.
[16] Sharoda A Paul, Lichan Hong, and Ed H Chi. 2011. What is a question? Crowd-sourcing tweet categorization. *CHI 2011* (2011).
[17] Helge Reelfs, Timon Mohaupt, Oliver Hohlfeld, and Niklas Henckell. 2019. Hashtag usage in a geographically-local microblogging app. In *LocWeb*. 919–927.
[18] Jens Helge Reelfs, Max Bergmann, Oliver Hohlfeld, and Niklas Henckell. 2021. *Understanding & Predicting User Lifetime with Machine Learning in an Anonymous Location-Based Social Network.*
[19] Jens Helge Reelfs, Oliver Hohlfeld, and Niklas Henckell. 2022. Geographic Differences in Social Media Interactions Exist Between Western and Middle-East Countries. In *Passive and Active Measurement Conference.*
[20] Martin Saveski, Sophie Chou, and Deb Roy. 2016. Tracking the Yak: An Empirical Study of Yik Yak.. In *ICWSM*.
[21] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *EMNLP*.
[22] J. Mitchell Vaterlaus. 2017. Yik Yak: An Exploratory Study of College Student Uses and Gratifications. *Bulletin of Science, Technology & Society* (2017).
[23] Gang Wang, Bolun Wang, Tianyi Wang, Ana Nika, Haitao Zheng, and et al. 2014. Whispers in the dark: analysis of an anonymous social network. In *IMC*.
[24] Jixuan Wang, Kuan-Chieh Wang, Frank Rudzicz, and Michael Brudno. 2021. Grad2Task: Improved Few-shot Text Classification Using Gradients for Task Representation. In *NeurIPS*.
[25] Wikipedia contributors. 2022. Mensch ärgere Dich nicht — Wikipedia, The Free Encyclopedia. (2022). https://en.wikipedia.org/w/index.php?title=Mensch_%C3%A4rgere_Dich_nicht&oldid=1065322692
[26] Yanqiu Wu, Tehila Minkus, and Keith W. Ross. 2017. Taking the Pulse of US College Campuses with Location-Based Anonymous Mobile Apps. In *TIST*. ACM.
[27] Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. 2011. Geographical topic discovery and comparison. In *WWW*.