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;

ACM Reference Format:

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 or 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 portrays 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).
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) (7) or downvoting (-1) (9) (similar to StackOverflow). Posts reaching a cumulative vote score (8) 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 (5).

2.1 Data Set Description and Ethics
Meta Data. The Jodel operator provided us with excerpts of anonymously mixed 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 normally 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).

**Intent \( I \) 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 \( I \), e.g., if a user is sharing information or seeking interaction. 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 \( \Theta \) 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 \( \Theta \) 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 in Table 1. We also opted for multi-labelling.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Class Information</th>
<th>#Annot</th>
<th>pdf</th>
</tr>
</thead>
<tbody>
<tr>
<td>EntObserve</td>
<td>Entertaining Observation</td>
<td>122 0.019</td>
<td></td>
</tr>
<tr>
<td>DistBelComp</td>
<td>Distress Release &amp; Complain. ab. Self</td>
<td>580 0.091</td>
<td></td>
</tr>
<tr>
<td>GenEntert</td>
<td>General Entertainment</td>
<td>618 0.097</td>
<td></td>
</tr>
<tr>
<td>Info</td>
<td>Information Sharing</td>
<td>638 0.100</td>
<td></td>
</tr>
<tr>
<td>SocVentComp</td>
<td>Social Venting &amp; Complain. ab. Others</td>
<td>644 0.101</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>Other</td>
<td>691 0.108</td>
<td></td>
</tr>
<tr>
<td>Self</td>
<td>Self Expression</td>
<td>927 0.146</td>
<td></td>
</tr>
<tr>
<td>Seek</td>
<td>Seeking Interaction/Information</td>
<td>2,149 0.337</td>
<td></td>
</tr>
<tr>
<td>(</td>
<td>\Theta</td>
<td>8 )</td>
<td>6,191 1.0</td>
</tr>
</tbody>
</table>

Table 1: Annotation Schema & Crowdsourcing Results. We code intents \( I \) catching the individual incentives, and topics \( \Theta \) 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, \( \alpha_M = 0.74 \) and \( \alpha_M = 0.74 \).

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.

**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

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.
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 Krippendorff’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 $\Theta$ 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 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 $\Theta$. While $I$ intents almost accidentally only were assigned a multi-label twice, we found 1,917 multi-labels across our coders for $\Theta$ topics. While $I$ intents almost accidentally only were assigned a multi-label twice, we found 1,917 multi-labels across our coders for $\Theta$ 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 $a_B$ scores across the board; lower $a_M$ values in comparison to $a_J$ 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 $\Theta$ 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 $\Theta$-hotspots are worth a look: Some may raise self-explainable confusion: e.g., $\Theta$-*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* × *SoeVentComp*; yet we must take note of the rest.

As for *topics* $\Theta$, 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 *PeopleRelation*, against *FitnessHealth*, and naturally along $\Theta$-*Other*. Noteworthy, we also identify several hotspots along $\Theta$-*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 ($a_M^j = 0.75, a_M^j = 0.64$) across #posts=1,789 (#Coders=2) and conclude that our proposed schema has sufficient quality for further evaluation.

<table>
<thead>
<tr>
<th>No</th>
<th>#Posts</th>
<th>#Coders&gt;1</th>
<th>$I \Theta$</th>
<th>$a_B$</th>
<th>$a_J$</th>
<th>$a_M$</th>
<th>Agreement $M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>733</td>
<td>0</td>
<td>$I \Theta$</td>
<td>0.66</td>
<td>0.66</td>
<td>0.66</td>
<td>substantial</td>
</tr>
<tr>
<td>2</td>
<td>1,999</td>
<td>0</td>
<td>$I \Theta$</td>
<td>0.44</td>
<td>0.57</td>
<td>0.52</td>
<td>moderate</td>
</tr>
<tr>
<td>3</td>
<td>400</td>
<td>398</td>
<td>$I \Theta$</td>
<td>0.69</td>
<td>0.69</td>
<td>0.69</td>
<td>substantial</td>
</tr>
<tr>
<td>4</td>
<td>1,000</td>
<td>993</td>
<td>$I \Theta$</td>
<td>0.45</td>
<td>0.60</td>
<td>0.55</td>
<td>moderate</td>
</tr>
<tr>
<td>5</td>
<td>400</td>
<td>398</td>
<td>$I \Theta$</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>substantial</td>
</tr>
<tr>
<td>all</td>
<td>4,532</td>
<td>1,789</td>
<td>$I \Theta$</td>
<td>0.57</td>
<td>0.62</td>
<td>0.61</td>
<td>substantial</td>
</tr>
</tbody>
</table>

Table 2: Classification agreement for multiple iterations on thread starting posts; For intents $I$ & topics $\Theta$ we show classifier agreement by $a_B$ Binary, $a_J$ Jaccard, and $a_M$ MASI distance. There is a substantial agreement between our two coders. Coders agree better on intents ($I$) than topics ($\Theta$).
5 WHAT JODELP 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 I by annotation counts. Since topics and intents are intertwined, we also show the combination of $I \times Θ$ as a heatmap in Figure 3. We complement this heatmap by discussing topic distributions across intents next (not shown).

**Intents I.** The dominant intents are (see Table 1): $I$ - Seek (Σ34%) and Self (Σ15%), followed by SocVent & DistRel (Σ19%), Info (Σ10%), and Entertainment (Σ12%). We observe only little disagreement between classifiers, explained by a possible ambiguity within $I$ - EntObserv × GenEntert, or apparent confusions along Self.

**Topics Θ.** Albeit slightly weaker annotator agreement, the discussed topics largely revolve around Θ-PeopleRelation accounting for Σ25% annotations, which is also our most confused category. We find other popular themes in Θ-Self (Σ13%), Other (Σ13%), followed by EntertCulture (Σ10%), and BeliefsPol (Σ8%). IllegalViolence (Σ2%), AnimalsNature (Σ3%), and FitnessHealth (Σ3%) are least popular.

**Intents $I \times Θ$ Topics.** We identify specific hotspots of interests by combining $I \times Θ$ as a heatmap in Figure 3. Jodel is mostly being used out of the intent of $I$ - Seeking Information & Interaction (Σ34%) for Θ-PeopleRelation (p[Θ|I]=23%) and EntertCulture (p[Θ|I]=12%), closely followed by others. Likewise, finding $I$ - Self (Σ15%) Expression across the board, users again focus on Θ-PeopleRelation (p[Θ|I]=23%) and Self Expression (p[Θ|I]=14%). Out of $I$ - GenEntert (Σ10%), we want to highlight Θ-BeliefPol (p[Θ|I]=57%). Whereas $I$ - SocVentComp (Σ10%) almost naturally goes along with the topic Θ-PeopleRelation (p[Θ|I]=41%), $I$ - DistRelComp (Σ9%) aligns with Θ-Self (p[Θ|I]=31%) and PeopleRelation (p[Θ|I]=22%).

**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 $Θ$ - PeopleRelation, EntertCulture, and Self totaling for p[Θ|I]=(37%, 16%, 13%, Σ66%) within these intents. Further, another Σ19% of posts are driven by $I$ - SocVentComp and DistRelComp within the same topical regime. Unfortunately, crowdsourcing a well-suited anonymity-sensitivity score relying on many (n=89) coders 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 $I \times Θ$ 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 Interaction & Information actually refers to local matchmaking, events, local services, or educational institutions. By focusing on the platforms content and driving factors first,
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 topic distribution, the least replies can be expected for topics with a low conversational 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 I-SelfExpression & Entertainment along Θ-SocialMedia, People & Relationships.

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 I-SelfExpression & Entertainment (cut-offs at Θ.75 to 0.8), versus more homogeneous conversations in GenEntertainment & EntertainmentObs (Θ.43, 0.58). ©-BeliefsPol (Θ.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.

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 value consistency (≈15% of all posts have score 1.0). Whereas for most I-Topics cut off the S-shape between Θ.02 to 0.1 on the lower end, they reach between Θ.61 to 0.79 at the upper end; I-(Other, Info) being outliers at (0.82, 0.88). This general observation
likewise holds true for Θ topics. Nonetheless, we find most contro-
versially down-voted posts within I-Other (0.10, 0.75), whereas
Θ-BeliefsPol (0.03, 0.61) are most controversial within upvotes.

Findings. From overall distributions across the board, conversation-
ness 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 I × Θ
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 Δ:=I-
GenEntert × Θ-BeliefsPol As for conversationness, we find most
homogeneous discussions especially within Δ ∪ I-EntertObs. The
community often appreciates Δ-content with high confidence. In con-
clusion, 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, I × (Seek, DistRelComp) × Θ-
(ProdService, Self, FitnessHealth) receive the fewest votes. While
the latter topics are most controversially dis-/liked, the same holds
true for I-Info. Lastly, the communities enjoy replying to I × (Self,
Other) × 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 increas-
ing our sample size to create automatic classifications at scale. How-
ever, 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
data set 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 fac-
tors 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 pro-
vide 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 rec-
ognized a central recurring motif of vivid matchmaking, dating and
flirting; thus, we added a DatingFlag correlating with this theme.

<table>
<thead>
<tr>
<th>Type</th>
<th>#Hashtags</th>
<th>#Jodels PDF</th>
<th>DatingFlag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal Information Sharing</td>
<td>501</td>
<td>6.36%</td>
<td>25%</td>
</tr>
<tr>
<td>Confessions</td>
<td>76</td>
<td>1.17%</td>
<td>-</td>
</tr>
<tr>
<td>18+</td>
<td>70</td>
<td>1.11%</td>
<td>74%</td>
</tr>
<tr>
<td>Matchmaking</td>
<td>43</td>
<td>1.02%</td>
<td>100%</td>
</tr>
<tr>
<td>Debates &amp; Opinion</td>
<td>92</td>
<td>0.86%</td>
<td>52%</td>
</tr>
<tr>
<td>Upvote Campaigns</td>
<td>14</td>
<td>0.55%</td>
<td>-</td>
</tr>
<tr>
<td>Other</td>
<td>219</td>
<td>2.85%</td>
<td>3%</td>
</tr>
</tbody>
</table>

Table 3: Classification of the top 1015 hashtags.

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=0.54%),
- Questioning forced wear of the Niqab (p=0.54%),
- Sparkling discussions about women driving (p=0.45%),
- Controversial discussions and questions about homosexuality and corresponding what-if scenarios (p=0.36%),
- Words of racism against foreigners (p=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 negligence, 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 accounting for 45k Jodels (p≈0.4%); 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 ephemeral 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 Flicker 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 Jed-dah, 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 sights of bored men sitting alone in cafes, 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].

ACKNOWLEDGMENTS

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.

REFERENCES

[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.
[27] Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. 2011. Geographical topic discovery and comparison. In WWW.