Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)

Determining Response-generating Contexts on Microblogging Platforms

Jennifer Fest
RWTH Aachen University
English Linguistics
jennifer.fest@ifaar.rwth-aachen.de

Arndt Heilmann
RWTH Aachen University
English Linguistics
arndt.heilmann@ifaar.rwth-aachen.de

Oliver Hohlfeld
Brandenburg University of Technology
Computer Networks and Communication Systems
oliver.hohlfeld@b-tu.de

Stella Neumann
RWTH Aachen University
English Linguistics
stella.neumann@ifaar.rwth-aachen.de

Jens Helge Reelfs
RWTH Aachen University
Chair of Communication and Distributed Systems
helge.reelfs@comsys.rwth-aachen.de

Marco Schmitt
RWTH Aachen University
Department of Sociology
mschmitt@soziologie-rwth.aachen.de

Alina Vogelgesang
RWTH Aachen University
Department of Sociology
avogelgesang@soziologie.rwth-aachen.de

Abstract

In recent years the study of social media communities has come into the focus of research. One open but central question is which properties stimulate user interaction within communities and thus contribute to community building. In this paper, we provide a first step towards answering this question by identifying features in the *Jodel* microblogging app that trigger user responses as one form of attention. Jodel is a geographically restricted app that allows users to post threads and comments anonymously. The absence of displayed user information on Jodel makes the posted content the only trigger for user interaction, making the language the one and only means for users to gather contextual implications about their discourse partners. This enhanced function of language promises a revealing baseline investigation into linguistic behavior on social media.

To approach this issue, we conducted a sequence of lexico-grammatical analyses and subjected the quantitative results to various statistical tests. While a Principal Component Analysis did not show a significant difference between the grammatical structure of original posts with and without answers, a negative binomial regression model focusing on the interpersonal meta-function yielded significant results. A further analysis of these features correlated to shorter or longer response times showed significant results for the interrogative mood. Additionally, keyword analyses revealed significant differences between posts with answers and without answers. Our study provides a promising first step towards understanding textual features triggering user interaction and thereby community building – an unresolved problem of practical relevance to social network operation.

1 Introduction

Social media in general and microblogging in particular create social spaces that are different from any we encounter in the physical world. In such online spaces, we can communicate with others independently of geographical or social distance. This form of communication is typically enriched with meta information, e.g. user profiles, profile pictures or status messages. The social space takes on yet another form if posted content is not enriched by any data regarding the author – a new type of microblogging.

One such space is created by the app *Jodel*. It provides forums that are based solely on a user's location, i.e., only content in the user's proximity is shown. The users' feeds automatically change as soon as they move to a different location, forming individual local communities. As there is no way to use the app to intentionally get in touch with the same people repeatedly, the purpose cannot be to establish contacts or find specific like-minded users, but rather to participate in a local community/crowd. The only gain from posting on Jodel comes in the form of upvotes/downvotes and responses, i.e. in receiving attention in one way or another, something not every post is able to achieve. Many original posts (*original Jodels*, or *OJs*), pass by unnoticed, some generate only votes, others trigger lively and long-lasting discussions.

The absence of any user-related information makes the communication entirely anonymous and leaves the discourse partners to draw conclusions about each other on the basis of language only. Using Jodel and its content-specific form of communication therefore enables us to study language specific properties that trigger user interaction.

The aim of this study is to find out whether there are linguistic features which decide on or influence the success of a post in terms of it generating answers. We assume that certain topics, keywords and lexico-grammatical features trigger different response behavior patterns and intra-thread references; in some threads, we can mainly find references to the author of the OJ, while others entail active discussions during which participants refer to each other. These response-related patterns create networks that are structured very differently and, in future work, can work as a basis of comparing anonymous communication as on Jodel with non-anonymous discourse on platforms like Twitter or Facebook.

In the following sections, we will briefly review related work in the area of social media and introduce Jodel in some more detail. Section 4 will outline the dataset and methodology, before section 5 will then present the results and conclusions drawn from them. As this is an exploratory study, we will outline future research perspectives, with a focus on transdisciplinary approaches and goals.

2 Research on Microblogging

Social media is subject to several lines of research. Generally, social media and the language that is used on respective platforms present a challenge to research in linguistics and communication science, as the discourse format is fairly recent and unusual and spelling and punctuation are less standardized (Golding et al., 2017). Analyses into this field require interdisciplinary approaches (Bouvier, 2015) as well as part of speech-tagging for highly non-standardized social media texts (Neunerdt, 2016). The most related lines of research can be summarized in four categories.

I) Firstly, the discourse patterns that are used on or caused by social media are a main focus. Conversational analyses have been conducted on Twitter (D'Heer and Verdegem, 2015; Freelon and Karpf, 2015) and Facebook (Androutsopoulos, 2015; Bolander and Locher, 2015), including also considerations on the function of such platforms as news hubs and multipliers (Cataldi et al., 2010). Many analyses in this field focus on individual events, like election campaigns (Enli, 2017), social or political movements (Poell, 2014; Kavada, 2015; Treré, 2015), natural disasters (Liu et al., 2016) or controversial criminal cases (McEnery et al., 2015); in short, events that often increase the use of social media and produce high frequency rates of the production of new content by the users.

Other studies focus on very concrete features of social or linguistic nature, but independent of any single (type of) event, thereby identifying characteristics that are more generalizable for social media. The topics here are very diverse and range from the creation of virtual friendships and groups (Wang et al., 2009; Chambers, 2013) to the linguistic manifestation of such affiliations (Zappavigna, 2012), dialectology (Eisenstein, 2018; Hovy and Purschke, 2018), humour (Locher and Bolander, 2015), language change and awareness (Dooly, 2018) and expectations towards peers (French and Bazarova, 2017), to name but a few.

II) Related, though not often explicitly connected to this strand are sociological investigations into the emergence of networks on social media platforms. Even more so than studies into linguistic aspects, these analyses are almost exclusively limited to one platform, as their functions are too different to generalize findings or even permit the creation of similar network structures. Some, such as Facebook or Google+, work on the level of individuals as well as groups. They include fan-pages and promote the use of private chats and groups, and primarily make content visible to friends and contacts. Others, like Twitter or Instagram, rely on the profiles of individuals and, despite the function to ‘follow’ users, make content publicly visible. They are therefore used for different purposes, and network structures vary considerably. The dynamics of news production and trends, regardless of the topic, makes this a very interesting field of research.

III) Furthermore, as a special form of this dynamic structure, a number of studies analyses the implications of social media platforms on web 2.0. The vast majority of services does not provide a differentiation between laymen and experts, and all users can provide content and comment on events without being subject to moderation or quality control (Laux and Schmitt, 2017). Political, economic, scientific and mass media elites have discovered social media as a means of self-promotion and, as individuals or members of the organization, are clearly influencing the course of debates. While the tweets of regular users often have no (greater) resonance, well-known personalities attract a lot of attention and can use and increase social capital accumulated in other contexts. Reputation is mirrored in the number of ‘followers’ (Laux and Schmitt, 2017). This holds true for laymen as well, and social media platforms have become a major channel for aspiring models, musicians and other celebrities.

IV) Lastly, every online space has to cope with users that do not follow rules and produce fake news and abusive content. A lot of research is being conducted into creating means of identifying and filtering such content (Waseem and Hovy, 2016; Mondal et al., 2017; Baider, 2018; Ruzaite, 2018), and platform providers are eager for success in this field so as to avoid criticism and face potential consequences in the future. Especially since political parties and stakeholders have taken to social media and Donald Trump’s election in 2016 was accompanied by many accusations regarding “fake news”, this field of research has gained importance and has acquired an ideological dimension as well. From an economic point of view, failing to prevent harmful content can seriously harm a platform and drive away users, as the downfall of the Jodel-alike app *YikYak* demonstrated quite clearly (Safronova, 2017).

In sum, popular platforms such as Twitter and Facebook are the dominant focus in research. On platforms like these, however, users always promote themselves to a certain degree and can be assumed to adapt their content and language to the image they wish to convey of themselves. Jodel, being anonymous and therefore useless for self-promotion, does not require its users to adapt to anything. The discourse here is much rawer and fewer variables interfere with or influence the language, making the analysis of this discourse a sound potential basis for research into other social media channels.

3 Jodel Explained

Jodel is a mobile-only microblogging app. Launched in 2014, it has been widely adopted in several European and the GCC countries. Like Twitter, Jodel enables users to share short posts of up to 250 characters and images (or short videos). Unlike Twitter and other traditional social media platforms, Jodel *a)* does not have user profiles, thereby making user to user communication anonymous (although users are enumerated within a thread, enabling interactions), and *b)* displays content only in the proximity of the user’s location, forming local communities.

Jodel is based on a community-driven filtering and moderation scheme to avoid adverse or abusive content. As stated above, effective moderation is a key parameter for the success of any social network or group messaging app. In Jodel, content filtering relies on a distributed voting scheme in which every user can raise or lower a post’s vote score by upvoting (+1) or downvoting (-1), similar to the mechanisms on other platforms such as StackOverflow or 9gag. OJs reaching a cumulative vote score below a negative threshold (e.g., -5) are not displayed anymore and can therefore not be commented or voted on any longer. Within threads, posts below this threshold are suppressed, but can sill be clicked on to read, should anyone need their content to understand the whole discussion.

Depending on the number of vote contributions, this scheme filters out bad content while also potentially boosting mainstream content. As a second line of defense, Jodel selects community moderators who have the authority to cast a moderator vote on posts that have been flagged (i.e. reported) by other users. On this superordinate level, the decision is again made by cumulative vote scores calculated when several moderators have made their choice. These moderators are supposed to decide on the basis of the app's official guidelines, which forbids insults, sexually explicit pictures, any means of identification and other controversial content that can be clearly defined. Posts that are voted out on this level are automatically blocked entirely.

4 Method

For our analysis, the Jodel network operator provided us with anonymized data of their network. Our corpus (cf. Table 1) contains posts from the city of Aachen, Germany, from April 1 to August 31, 2017, as well as metadata on the individual posts. The focus on a single location and an arguably short time frame enabled us to focus on a more homogeneous group of users in the dimensions of content and time. All analyzed data has been publicly posted and thus been visible to all other Jodel users.

Metric	Entries
#Threads	182,413
#Responses	1,275,763
#Users	21,282
#Tokens	19,627,690

Table 1: Corpus

The corpus data was preprocessed with SoMeWeTa (Proisl, 2018), a part of speech-tagger for German social media and web texts. This tagger includes several tags specific to language used in social media texts in addition to the tags from the Stuttgart Tübingen Tag Set (STTS). Using a broad range of lexico-grammatical features as typically done in register analysis (Biber and Conrad, 2009; Neumann, 2014; Fest, 2016) allows us to explore linguistic variation in general, which might reflect differences between posts that begin a thread and receive answers from those that do not. These features range from lexical based measures (e.g., Lexical Density, Nominal Density) to approximations of clausal complexity (e.g., finite verbs per sentence). Frequency counts of these lexico-grammatical features were performed in the Corpus Workbench (Evert and Hardie, 2011) with the help of a query script based on the example of a script developed by Neumann et al. (2017) for English and Dutch. The script exploits on part of speech-annotation in combination with positional information and word lists. Some queries address features specific to social media texts. The latter were queries for some of the additional tags provided by the SoMeWeTa, foreign language material, words specific to social media or the platform Jodel (*OJ, Heimatjodel*, etc.), colloquialisms and the metadata tags (hashtags, mentions, channels) used on Jodel.[1]

In order to explain answer behaviour patterns and determine potential triggers of successfully generating responses, we used exploratory and confirmatory techniques. The first analysis is exploratory and is a global linguistic assessment of posts that start a thread (OJs) to find out if specific linguistic patterns emerge from a multitude of linguistic features that distinguish response-generating OJs from those that remain unanswered. This allows us to find out if successful OJs (OJs with at least one answer) are associated with particular sets of grammatical behavior(s). A total of 68 features was included after discarding collinear features ($r > |0.9|$). Each post is thus represented as a vector in a 68 dimensional space.

To reduce the dimensionality of the data, we first applied a Principal Component Analysis (PCA) to a random sample of the OJs in the Jodel corpus (n=10,000). The frequency results of the queries for the individual features were normalized with respect to appropriate units of measurement. For example, sentences were given per post, lexical measures such as nouns are included as the ratio of nouns to tokens, whereas passives and tense related features are given as the proportion of sentences. We also included measures of lexical density, which comes closest to a lexical indicator of the features considered. All values were standardized as z-scores, bringing the indicators to the same scale. Additionally, the indicators were transformed using a signed logarithm to reduce the skew of some of the variables that may have made it difficult to interpret the PCA. Based on these standardized indicators, a vector was built for each post which assumes a position in the multidimensional space.

In a second step, we shifted from the exploratory analysis to a confirmatory one and used regression

[1]For a list of all queried features, see appendix A.

modelling with a set of linguistic features that we deemed reflective of social interaction. We decided to include first and second person pronouns, each per total amount of pronouns. Unfortunately, the second person plural pronoun *ihr* ('you' pl.) is indistinguishable from *ihr* ('her' sg.) in its third person singular sense, so there might be some overlap between both categories. We also included the number of imperatives, salutations and vocatives as well as the number of interrogatives per sentence. Additionally, we added the numbers of hashtags and emojis per total amount of tokens. Lastly, we controlled for number of words per sentence to account for length effects. Due to overdispersion issues with Poisson regression we used a negative binomial regression model. We hypothesized that it may also be specific content and not only linguistic behaviour per se, that triggers answers from the Jodel community. To this end, we conducted keyword analyses to filter out unusually frequent words in posts that gathered answers as opposed to those that did not.

5 Results

5.1 Grammatical Analysis

For the PCA, we assume that the Euclidean distances between feature vectors are suitable measures of (dis)similarity between data points (the OJs) with respect to the geometric configuration of vectors in multidimensional space and that these distances can be visualized using orthogonal projections to draw conclusions about the data. Clusters of posts that become apparent in the orthogonal projections can be interpreted as posts with different grammatical features. One or more clusters may be associated with grammatical features that attract answers.

Figure 1 visualizes the first three orthogonal dimensions. Most of the variance in the data is explained by the first dimension. Variance explained by the first three dimensions was 0.122, 0.058 and 0.052.

The PCA does not suggest a clear separation of OJs with and without answers. Separate clusters are virtually absent in the first three PCA dimensions and with respect to answers we could not identify a clear location of posts with no answer within the single big cluster.

According to Halliday et al. (2014), a significant part of the meaning potential of language revolves around interpersonal meanings, that is, around the

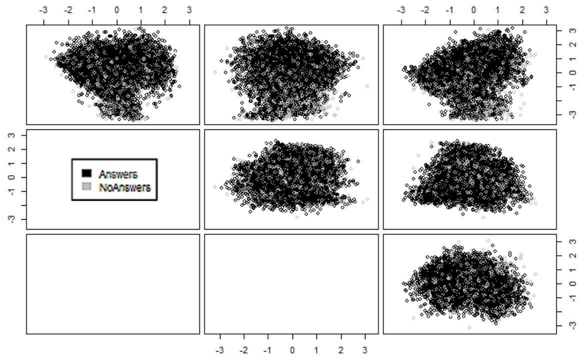

Figure 1: PCA Posts with answers / without answers

ways a writer uses to enact a social relationship with a reader. We hypothesized that features that can be associated with this kind of meaning could allow a more fine-grained understanding of linguistic features concerning the answer behavior of Jodel users. We therefore selected variables that we associated with indicators of dialogic interaction (cf. Table 2) and ran generalized linear regression models. This has the advantage that the effect of each variable on answering activity can be directly measured. We ran the regression models with the same $10,000$ OJs but only a subset of the linguistic features from the prior analysis and predicted the number of answers.

No. of Answers	Estimate	Std. Error	z	p
word_S	0.024	0.002	10.97	<0.001
emoji_T	-0.374	4.174	-0.09	0.929
hashtag_T	-2.728	0.161	-16.979	<0.001
p1pronoun_Pr	0.364	0.048	7.59	<0.001
p2pronoun_Pr	-0.236	0.09	-2.632	<0.01
ihr_Pr	1.048	0.118	8.878	<0.001
imperative_S	-0.166	0.126	-1.315	0.188
salutation_S	0.184	0.129	1.434	0.152
vocatives_S	-0.095	0.211	-0.451	0.652
interrogative_S	0.377	0.043	8.766	<0.001

Table 2: Results for negative binomial regression model for the number of answers

The results show that several conversational features correlated positively and statistically significantly with the number of answers per OJ. These are words per sentence, interrogatives, first person pronouns, second person pronouns – particularly *ihr* and hashtags. Interestingly, an increased use of hashtags led to fewer responses, as did second person pronouns. Because we hold constant the effects for the second person plural *ihr*, this effect likely reflects the second person singular *du* ('you'). Why

Answer Delay	Estimate	Std. Error	z	p
word_S	-0.001	0.004	-0.155	0.877
emoji_T	1.894	9.693	0.195	0.845
hashtag_T	0.95	0.475	2.001	<0.05
p1pronoun_Pr	-0.032	0.097	-0.334	0.738
p2pronoun_Pr	-0.413	0.17	-2.43	<0.05
ihr_Pr	-0.151	0.235	-0.644	0.52
imperative_S	0.555	0.283	1.961	<0.05
salutation_S	-0.307	0.263	-1.169	0.243
vocatives_S	-0.268	0.428	-0.627	0.531
interrogative_S	0.228	0.087	2.614	<0.01

Table 3: Results for binomial logistic regression model for the time passed between OJ and the first answer (> 5 minutes resp. < 5 minutes). A positive sign indicates that a quick response is more likely than a slow response (e.g. interrogative mood makes a quick response more likely.)

an OJ would single out a specific addressee via *du* ('you' sg.) is somewhat interesting because Jodel is anonymous. A closer look at the data revealed that frequently this is used as an indefinite pronoun and often occurs in posts that reproduce dialogues of some kind or in reports of personal experience where the *du* actually refers to the OJ him-/herself and almost has the character of an internet meme:

"*Wenn du morgens aufstehst, dich für die Uni fertig machst und losfährst und dann merkst, dass du anstatt zur Uni zu deiner alten Schule gefahren bist.*" ('When you get up in the morning, get ready for university and start driving and then realize that you've driven to your old school instead of university.')

This kind of *du* thus refers to the OJ in the first instance and to others in a second instance, i.e. Jodel members that have made a similar experience or can relate to OJ. Syntactically it could be replaced with the pronoun *man* ('one' sg.), yet the achieved effect would be less personal. As a further analysis of the answer behavior we tested the time it took for the first person to answer the OJ with the same set of variables using a logistic binomial regression model. We categorized the answer time as "quick" (<5 minutes) and "slow" (>5 minutes) and predicted the likelihood of a quick answer compared to a slow answer. We chose to use a binomial model here because the model residuals were not normally distributed.

The model yielded significant results for interrogative mood, second person pronouns, imperatives and hashtags. The higher the ratio of hashtags per number of tokens, the more quickly a response was issued – probably due to reasons of visibility. Second person pronouns lead to slower responses. In the light of the example above, the use of *du* may not actually invite responses (neither quantitatively nor temporally) and may rather be reflective of the OJ's need to express themselves. Typical conversational features that we also observe in face to face conversation (interrogative mood and imperative mood) were more likely to trigger quick responses. This is noteworthy because this is not face to face conversation but anonymous texting. The findings suggest that a variety of mood aspects elicit answers from the Jodel community and that asking questions and the use of hashtags trigger responses quickly. The latter, however, does not generate many answers at the same time.

5.2 Keyword Analysis

The last step of our investigation was keyword analyses to examine the content level and determine which topics are likely to generate answers on Jodel. Table 4 shows the keywords and their significance for OJs that generated answers in contrast to those that did not.

word	raw freq.	keyness[2]	meaning/used as
ihr	18,099	907.91	pl. pronoun *you*
jhj	7,504	762.87	hashtag, *jodler helps jodler*
oder	11,949	462.89	*or*
kann	15,524	444.16	*can*, 3. p. sing.
jemand	13,124	373.00	*someone/anyone*
habe	11,410	295.32	*have*, 1. p. sing.
wo	6,041	256.10	*where*
und	53,899	254.31	*and*
habt	3,191	245.02	*have*, 2. p. pl.
tipps	1,296	242.77	*advice*

Table 4: Top 10 Keywords for Jodels with answers (vs Jodels without answers)

As can be seen, the hashtag *#jhj*, which signals a request for help from others, is one of the highest ranking keywords. The list furthermore includes *kann* in the 3rd person singular, which often collocates with *jemand*, indicating interrogatives. *Habt*, in 2nd person plural, is a direct address, and with *wo*, a direct question word is sixth in the list. The last item of the top ten keywords explicitly refers to advice.

Of those OJs which generate answers, we further examined the keywords for those that received the first answer within 5 minutes, in contrast to

[2]The keyness was calculated using Log-Likelihood and a threshold of p<0.05

word	raw freq.	keyness	meaning/used as
jhj	6,512	886.35	hashtag, *jodler helps jodler*
was	14,293	150.73	*what*
freundin	3,548	125.64	*(girl-)friend*
freund	3,066	117.36	*(boy-)friend*
frage	2,138	103.53	*question*
balloon	231	103.26	emoji
jodel	4,109	90.45	
er	5,932	83.27	*he*
warum	2,600	81.38	*why*
nicht	18,776	81.34	*not*

Table 5: Top 10 Keywords for Jodels with an answer delay < 5 min. (vs Jodels with an answer delay > 5 min.)

those that had to wait longer. As Table 5 shows, direct questions appear even stronger here, with the interrogative pronouns *was*, *warum* and the noun *Frage* being key. Again, *#jhj* is strong, meaning that not only does this hashtag trigger responses as such, but fast ones, too. This is particularly interesting as hashtags in general, as was described above, have proven to be counterproductive as a discussion starter; *#jhj* appears to be a significant exception.

The results confirm the assumption that Jodel contains a certain service function showing itself in the active answering of questions other users might have.

Another interesting finding that can be gathered from the quick response keywords is that one topic shows to be particularly popular, which is relationships. *Freund* and *Freundin* are the only nouns apart from *Frage* in the ten strongest keywords. To further investigate this assumption, we split the OJs that had received answers into nine subgroups, based on how many answers they received (cf. Table 6).

OJs	42,419	27,367	30,648	23,845
#ANS	1-2	3-4	5-8	9-16
11,925	4,388	1,137	248	47
17-32	33-64	65-128	129-256	257-∞

Table 6: OJs with specific numbers of answers

As expected, the majority of OJs received only few answers; 42,419 (i.e. 30%) were answered only once or twice. Others were followed by long discussions of over 100 contributions. As an exploratory test, we contrasted the 1-2 answer-group with the other groups and then gradually moved the border of contrast upwards. For a contrast of posts with up to 64 answers to those with 65 and more, only 19 keywords were significant at all for the long threads. 6 of these are directly linked to relationships and sex: *Männer* ('men'), *Beziehung* ('relationship'), *Kerle* ('guys'), *w* (for *weiblich*, 'female'), and *treu*, ('honest'/'faithful'). Another 6 are personal pronouns, and the emoji of the colourful rainbow, referring to LGBT communities, also features in the list.

5.3 Discussion

The results give some clear indications as to the formats and contents which most likely generate answers and discussions on Jodel for the Aachen community from April 2017 to August 2017, and also offer interesting insights into the particularities of the dynamics within an anonymous network. In contrast to other social media channels, where statements are made to produce reactions and users promote themselves and compete for followers, Jodel seems to be driven to a considerable degree by posting questions and getting answers. Regarding the popular topic of relationship and sex, it is likely that the anonymity on the platform is used to treat personal and sensitive issues which users would not discuss quite as openly elsewhere.

Because of its anonymity, the platform does not function as a public accumulation of status or social capital, in the form of likes or retweets, and the success of a post is not connected to a person or individual. Neither, of course, is failure of a post or disagreement to the posted content, which makes Jodel attractive to publish also controversial questions and opinions. The image of a person, which is at the center of interest on platforms like Twitter and Facebook, is of no consequence on Jodel and is therefore never at stake should anything not be received positively by the community.

Nevertheless, it can be seen that the Aachen Jodel users within our corpus indeed protect an image, if not their own. Questions are answered – and very often quickly – and especially calls for help generate fast responses. *#jhj* is such an example, and yet the answers to it are not always positive. The hashtag is key for posts that receive downvotes, particularly when the questions are very obviously either self-explanatory or could have been answered with little effort by the OJ. "Why don't you just google it?" or "Learn how to google!" are common answers in such instances, showing that users take the time to reprimand someone although they have no immediate gain from that action, solely for the purpose of enforcing the community's rules and keeping up its identity. At this point, therefore, the attempt is made to maintain the image of the platform, part of which is that the questions asked should not be too trivial and that the content can be controversial, but should definitely be original and exciting.

Cultural and local identity is also mirrored in the topics that are received with most enthusiasm. Many general topics are discussed on the platform – from politics, university and job life, the city and parties to the weather and current events – and all of them receive various amounts of answers, but nothing sparks so much interest as relationships and sex. In Aachen, a city with a technical university and student population that is 68% male (RWTH Aachen University, 2018), how to get a partner, a one-night-stand or any other contact with the opposite sex are ever-present questions that are at times so dominant that the city's Jodel community has coined the term *geiern* ('to vulture') for males that hit on women all too obviously and persistently.

Taking all this into account, we can conclude that the anonymous character of Jodel leads to the creation of conventions – within the framework of the general user guidelines – at the center of which is not the identity of the user, but the image of a local community on the platform. This identity is cherished and protected, which in turn triggers an entirely unexpected and hard to define type of abusive content for a specific community, namely insulting anyone who does not abide by the self-imposed laws of the community.

6 Outlook and Future Work

As this last point already shows, an analysis into abusive content and hate speech is certainly promising on an anonymous platform. The community's identity is built and under constant negotiation, and users that offend the community's sense of self in some way have to expect to be discriminated against, even if they do not post anything objectively offensive. Finding and automatically filtering such instances is very difficult, yet we can assume that we can find similar mechanisms on other social media platforms as well. In future work, therefore, the understanding obtained by taking this focused perspective on the social network of Jodel can be extended to a broader set of locations and time frames.

Some next steps within the present line of research are very apparent; by widening the dataset in a spatial and temporal dimension for a broader analysis, more complex models and classifiers as a predictor for a community's responsiveness can be developed. Also, we plan to extend our feature set to include metadata variables and examine more closely the role of hashtags and emojis.

As yet a different angle, a comparative approach to platforms like Facebook and Twitter seems promising. The differences in the contexts provided to the users by the platforms allow for the emergence of different social spaces, which in turn affects the purpose of using the platform as well as its communities and networks. This is reflected for instance in the use of hashtags. While a first look at hashtag usage on Jodel has shown that in many cases, such as the mentioned *#jhj*, hashtags define content of a category as they do on Twitter, many other hashtags on Jodel are means of content rather than of categorization (Fest et al., 2019; Reelfs et al., 2019). The negative effect of the hashtags on the likelihood of receiving answers indicates that this method, although frequently used, might not be too popular with the community.

On Twitter on the other hand, the function of hashtags is of a more exclusively categorizing nature, including the possibility to feature one's tweets in discussions on topics they are not related to simply by including the most trending hashtags of the moment. On both platforms, however, hashtags are a context marker – and the differences in usage therefore pose a window on just how context can be perceived and created.

Acknowledgement

This work has been funded by the Excellence Initiative of the German federal and state governments.

References

Jannis Androutsopoulos. 2015. Networked multilingualism: Some language practices on Facebook and their implications. *International Journal of Bilingualism*, 19(2):185–205.

Fabienne Baider. 2018. "Go to hell fucking faggots, may you die!" framing the LGBT subject in online comments. *Lodz Papers in Pragmatics*, 14(1):69–92.

Douglas Biber and Susan Conrad. 2009. *Register,* *Genre and Style*. Cambridge University Press, Cambridge, UK.

Brook Bolander and Miriam A. Locher. 2015. "Peter is a dumb nut": Status updates and reactions to them as 'acts of positioning' in Facebook. *Pragmatics*, 25(1):99–122.

Gwen Bouvier. 2015. What is a discourse approach to twitter, facebook, youtube and other social media: connecting with other academic fields? *Journal of Multicultural Discourses*, 10(2):149–162.

Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. 2010. Emerging topic detection on Twitter based on temporal and social terms evaluation. In *Proceedings of the 10th International Workshop on Multimedia Data Mining*, page ch. 4. ACM, New York, NY.

Deborah Chambers. 2013. *Social media and personal relationships: Online intimacies and networked friendship*. Palgrave Macmillan, Basingstoke.

Evelien D'Heer and Pieter Verdegem. 2015. What social media data mean for audience studies: a multidimensional investigation of twitter use during a current affairs tv programme. *Information, Communication & Society*, 18(2):221–234.

Melinda Dooly. 2018. "i h8 txt msgs". how social media has had an impact on language awareness. In Peter Garrett and Josep M. Cots, editors, *The Routledge Handbook of Language Awareness*, pages 306–322. Routledge, New York and London.

Jacob Eisenstein. 2018. Identifying regional dialects in online social media. In Charles Boberg, Dominic James Landon Watt, and John A. Nerbonne, editors, *The handbook of Dialectology*, pages 368–383. John Wiley & Sons Inc, Hoboken, NJ.

Gunn Enli. 2017. Twitter as arena for the authentic outsider: exploring the social media campaigns of trump and clinton in the 2016 us presidential election. *European Journal of Communication*, 32(1):50–61.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics Conference 2011, University of Birmingham, UK, 20-22 July 2011*, Birmingham, UK. University of Birmingham.

Jennifer Fest, Stella Neumann, Arndt Heilmann, Oliver Hohlfeld, Jens Helge Reelfs, and Marco Schmitt. 2019. Hate speech dynamics in anonymous microblogging. In *29th European Systemic Functional Linguistics Conference*.

Jennifer Fest. 2016. *News in the context of regional and functional variation: a corpus-based analysis of newspaper domains across varieties of English*. Phd thesis, RWTH Aachen University, Aachen, Germany.

Deen Freelon and David Karpf. 2015. Of big birds and bayonets: Hybrid Twitter interactivity in the 2012 Presidential debates. *Information, Communication & Society*, 18(4):390–406.

Megan French and Natalya N. Bazarova. 2017. Is Anybody Out There?: Understanding Masspersonal Communication Through Expectations for Response Across Social Media Platforms. *Journal of Computer-Mediated Communication*, 22(6):303–319.

Peter Golding, Karin Raeymaeckers, and Helena Sousa. 2017. Social media – new challenges and approaches for communications research. *European Journal of Communication*, 32(1):3–5.

Michael Alexander Kirkwood Halliday, Christian Matthiessen, and Michael Halliday. 2014. *An introduction to functional grammar*. Routledge.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium, October-November. Association for Computational Linguistics.

Anastasia Kavada. 2015. Creating the collective: social media, the occupy movement and its constitution as a collective actor. *Information, Communication & Society*, 18(8):872–886.

Henning Laux and Marco Schmitt. 2017. Der Fall Bautzen: Eine Netzwerkanalyse zur Entstehung digitaler Öffentlichkeiten. *Berliner Journal für Soziologie*, 27(3-4):485–520.

Brooke Fisher Liu, Julia Daisy Fraustino, and Yan Jin. 2016. Social media use during disasters: How information form and source influence intended behavioral responses. *Communication Research*, 43(5):626–646.

Miriam A. Locher and Brook Bolander. 2015. Humour in microblogging: Exploiting linguistic humour strategies for identity construction in two facebook focus groups. In Marta Dynel and Jan Chovanec, editors, *Participation in Public and Social Media Interactions*, volume 256 of *Pragmatics & Beyond New Series*, pages 135–155. John Benjamins, Amsterdam.

Tony McEnery, Mark McGlashan, and Robbie Love. 2015. Press and social media reaction to ideologically inspired murder: The case of lee rigby. *Discourse & Communication*, 9(2):237–259.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A Measurement Study of Hate Speech in Social Media. In Peter Dolog, Peter Vojtas, Francesco Bonchi, and Denis Helic, editors, *Proceedings of the 28th ACM Conference on Hypertext and Social Media - HT '17*, pages 85–94, New York. ACM Press.

Stella Neumann, Gert De Sutter, and Stefan Evert. 2017. Register-specific interference in translation. In *Conference booklet of the 39th Annual Conference of the German Linguistic Society*, Saarbrücken, Germany, March.

Stella Neumann. 2014. *Contrastive Register Variation: A Quantitative Approach to the Comparison of English and German*. De Gruyter Mouton, Berlin.

Melanie Neunerdt. 2016. *Part-of-Speech tagging and detection of social media texts*. Apprimus, Aachen, Germany.

Thomas Poell. 2014. Social media and the transformation of activist communication: exploring the social media ecology of the 2010 toronto g20 protests. *Information, Communication & Society*, 17(6):716–731.

Thomas Proisl. 2018. SoMeWeTa: A Part-of-Speech Tagger for German Social Media and Web Texts. In Calzolari N, Choukri K, Cieri C, Declerck T, Goggi S, Hasida K, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, Tokunaga T, editor, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 665–670, Miyazaki. European Language Resources Association.

Helge Reelfs, Timon Mohaupt, Oliver Hohlfeld, and Niklas Henckell. 2019. Hashtag usage in a geographically-local microblogging app. In *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019.*, pages 919–927.

Jurate Ruzaite. 2018. In search of hate speech in Lithuanian public discourse: A corpus-assisted analysis of online comments. *Lodz Papers in Pragmatics*, 14(1):93–116.

RWTH Aachen University. 2018. Facts and figures.

Valeriya Safronova. 2017. The Rise and Fall of Yik Yak, the Anonymous Messaging App. *The New York Times*.

Emiliano Treré. 2015. Reclaiming, proclaiming, and maintaining collective identity in the #yosoy132 movement in mexico: an examination of digital frontstage and backstage activism through social media and instant messaging platforms. *Information, Communication & Society*, 18(8):901–915.

Zuoming Wang, Joseph B. Walther, and Jeffrey T. Hancock. 2009. Social identification and interpersonal communication in computer-mediated communication: What you do versus who you are in virtual groups. *Human Communication Research*, 35(1):59–85.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 88–93. ACL, Stroudsburg, PA.

Michele Zappavigna. 2012. *The discourse of Twitter and social media*. Continuum, London and New York.

Appendix A. Queried linguistic features

Feature	Details
lexical_density	Number of lexical words divided by the number of tokens
nn_T	Number of nouns divided by the number of tokens
ne_T	Number of proper nouns divided by the number of tokens
nominal_T	Number of nominalizations divided by the number of tokens
neoclass_T	Number of neoclassical compounds divided by the number of tokens
poss_T	Number of possesive pronouns divided by the number of tokens
pronouns_T	Number of pronouns divided by the number of tokens
p1pronoun_Pr	Number of 1st person personal pronouns divided by the number of pronouns
p2pronoun_Pr	Number of 2nd person personal pronouns divided the number of by pronouns
p3pronoun_Pr	Number of 3rd person personal pronouns divided the number of by pronouns
ihr_Pr	Number of instances of the pronoun *ihr* divided by the number of pronouns
es_Pr	Number of instances of the pronoun *es* divided the number of pronouns
pospers1_Pr	Number of all 1st person pronouns divided by the number of pronouns
pospers2_Pr	Number of all 2nd person pronouns divided by the number of pronouns
pospers3_Pr	Number of all 3rd person pronouns divided by the number of pronouns
adv_T	Number of adverbs divided by the number of tokens
adj_T	Number of adjectives divided by the number of tokens
atadj_T	Number of attributive adjectives divided by the number of tokens
prep_T	Number of prepositions divided by the number of tokens
finite_S	Number of finite verbs divided by the number of sentences
finite_V	Number of finite verbs divided by the number of verbs
pasttense_F	Number of past tense verbs divided by the number of finite verbs
perfect_F	Number of perfect verbs divided by the number of finite verbs
plusquamperfect_F	Number of instances of past perfect divided by the number of finite verbs
will_F	Number of instances of the modal verb *werden* used to signal future divided by the number of finite verbs
modalverb_V	Number of modal verbs divided by the number of verbs
verb_T	Number of verbs divided by the number of verbs
infinite verbs_F	Number of infinitives with *zu* divided by the number of sentences
passive_S	Number of instances of passive voice divided by the number of sentences
coordination_S	Number of coordinating conjunctions divided by the number of sentences
subordination_S	Number of subordinating conjunctions divided by the number of sentences
interrogative_S	Number of instances of interrogative mood divided by the number of sentences
imperative_S	Number of instances of imperative mood divided by the number of sentences
politeimperative_S	Number of polite imperatives divided by the number of sentences
subjunctive_S	Number of modal verbs in subjunctive mood divided by the number of sentences
title_T	Number of titles divided by the number of tokens
salutation_S	Number of salutations and greetings (eg. *Hallo*, *Tschüss*, *Viele Grüße*) divided by the number of sentences
placeadv_T	Number of adverbs of place divided by the number of tokens
timeadv_T	Number of adverbs of time divided by the number of tokens
vocatives_S	Number of vocatives divided by the number of sentences
nptheme_S	Number of nominal elements in theme position divided by the number of sentences
numbertheme_S	Number of numbers in theme position divided by the number of sentences
pptheme_S	Number of prepositions in theme position divided by the number of sentences
advtheme_S	Number of adverbs in theme position divided by the number of sentences
textheme_S	Number of conjunctions in theme position divided by the number of sentences
whtheme_S	Number of wh-elements in theme position divided by the number of sentences
disctheme_S	Number of discourse markers in theme position divided by the number of sentences
nonfinite verbstheme_S	Number of infinitives with *zu* in theme position divided by the number of sentences
subordconjtheme_S	Number of subordinating conjunctions in theme position divided by the number of sentences
verbtheme_S	Number of verbs in theme position divided by the number of sentences
incompletesentences_S	Number of incomplete sentences divided by the number of sentences
cohesiveadverbs_T	Number of cohesive adverbs divided by the number of tokens
emoji_T	Number of emojis divided by the number of tokens
hashtag_T	Number of hashtags divided by the number of tokens
jodelwords_T	Number of instances of Jodel specific words (eg. *jodel*, *karma*) divided by the number of tokens
socialmediawords_T	Number of instances of social media specific words (eg. *upvote*) divided by the number of tokens
colloqialisms_T	Number of colloquialisms divided by the number of tokens
fm_T	Number of foreign language material divided by the number of tokens
emojitheme_S	Number of emojis in theme position divided by the number of sentences
emojionly_S	Number of sentences consisting only of emojis divided by the number of sentences
hashtagtheme_S	Number of hashtags in theme position divided by the number of sentences
hashtagonly_S	Number of sentences consisting only of hashtags divided by the number of sentences

reftheme_S	Number of references to another user in theme position divided by the number of sentences
correctiontheme_S	Number of corrections in theme position divided by the number of sentences
personaldetails_P	Number of personal details (e.g. gender and age) divided by the number of sentences
modalpart_T	Number of modal particles divided by the number of tokens
focuspart_T	Number of focus particles divided by the number of tokens
multipart_T	Number of multi-word particles divided by the number of tokens
contrverbpron_T	Number of contractions with a verb and a pronoun (eg. *gehts*, *habs*) divided by the number of tokens
contrpronpron_T	Number of contractions with two pronouns (eg. *ers*, *sies*) divided by the number of tokens
contrkoupron_T	Number of contractions with a conjunction and a pronoun (eg. *weils*, *obs*) divided by the number of tokens
contrpreart_T	Number of contractions with a preposition and an article (eg. *beim*, *am*) divided by the number of tokens