Journal of Open Humanities Data

The Telegram Chronicles of Online Harm

RESEARCH PAPER

TATJANA SCHEFFLER
VERONIKA SOLOPOVA
MIHAELA POPA-WYATT
Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Harmful language is frequent in social media, in particular in spaces which are considered anonymous and/or allow free participation. In this paper, we analyze the language in a Telegram channel populated by followers of former US President Donald Trump. We seek to identify the ways in which harmful language is used to create a specific narrative in a group of mostly like-minded discussants. Our research has several aims. First, we create an extended taxonomy of potentially harmful language that includes not only hate speech and direct insults (which have been the focus of existing computational methods), but also other forms of harmful speech discussed in the literature. We manually apply this taxonomy to a large portion of the corpus, including the time period leading up to and the aftermath of the January 2021 US Capitol riot. Our data gives empirical evidence for harmful speech, such as in/out-group divisive language and the use of codes within certain communities, that have not often been investigated before. Second, we compare our manual annotations of harmful speech to several automatic methods for classifying hate speech and offensive language, namely list-based and machine-learning-based approaches. We find that the Telegram data sets still pose particular challenges for these automatic methods. Finally, we argue for the value of studying such naturally-occurring, coherent data sets for research on online harm and how to address it in linguistics and philosophy.

CORRESPONDING AUTHOR:
Tatjana Scheffler
German Studies,
Ruhr-Universität Bochum,
Bochum, Germany
tatjana.scheffler@rub.de

KEYWORDS:
online harm; social media; hate speech; Telegram; corpus linguistics; offensive language detection

TO CITE THIS ARTICLE:
Scheffler, T., Solopova, V., & Popa-Wyatt, M. (2021). The Telegram Chronicles of Online Harm. *Journal of Open Humanities Data*, 7: 8, pp. 1–15. DOI: *https://doi.org/10.5334/johd.31*

(1) INTRODUCTION

Digital media can cause harm in different ways. In addition to language that directly harms specific individuals and members of target groups, such as bullying, hate speech attacks, incendiary and dehumanizing language, or trolling, there are many more indirect and implicit avenues for online harm. These can include the spread of propaganda and disinformation, as well as the use of "othering" or divisive rhetoric which seeks to promote a sense of in-group identity at the expense of individuals perceived as part of the out-group. Such language can cause harm even when not directed at or read by target individuals or groups. It does so by poisoning online discourse, in particular by changing the boundaries of acceptability of harmful language and normalizing harmful practices with consequences beyond the online forums where such language is primarily trafficked.[1]

In this paper, we analyze a public channel from the direct messaging platform *Telegram*, which is rife with such indirect forms of harmful language. The channel is a platform which attracts individuals with extreme right-wing views, and thus facilitates not only networking but also potentially recruiting and mobilizing new members as part of a narrative designed to strengthen the in-group identity and sense of belonging to a community of like-minded people. This can arguably create a toxic climate in which users feel emboldened to share and to respond to hateful content both explicitly and implicitly at the expense of target groups as a way of justifying violent actions in the real world. For example, the Telegram users in our data set gradually went from discussing governmental overthrow as a theoretical possibility to planning the January 6, 2021 Capitol riot by sharing information on hotels and transportation in Washington, DC, and finally discussing the aftermath of the event.

Our paper makes the following contributions. First, we propose a taxonomy of harmful speech in online discussions which includes both direct and indirect forms, for categorizing empirical data. Second, we manually annotate a subset of the corpus with our taxonomy. Third, we apply several automatic annotations of hate speech and offensive language to our data, chronicling the prevalence of this language in the Telegram corpus. Finally, we evaluate the currently available automatic methods of offensive language detection by comparing them with our manual annotations of harmful speech. This will provide pointers for future work.[2]

(2) ONLINE HARM AND SOCIAL MEDIA

Online harm has become a central research focus in several fields, including computational linguistics, social and political philosophy, communication science, as well as in discussions about policy and regulation in the context of free speech debates (Brison & Gelber, 2019). One challenge for this work is the lack of a common definition of the relevant categories. A widely studied type of online harm is "hate speech", but Poletto et al. (2020, Figure 1) show that this concept is closely related to other terms as well. Hate speech has a specialized use as a legal concept, referring to speech that is subject to regulation by legal bodies and systems. The boundaries of what counts as hate speech are contested, and as such various terms have been proposed as more specific or more comprehensive: e.g., "dangerous speech", "offensive language", "assaultive language", "poisonous language", "discriminatory verbal harassment", "incitement", etc. (Matsuda et al., 1993; Haraszti, 2012; Benesch et al., 2018; Brison & Gelber, 2019). Furthermore, even data-driven computational approaches do not agree on clear definitions of what counts as "offensive" language (Vidgen & Derczynski, 2020), and where definitions are provided, they often contradict each other. To avoid such contested territories, we shall instead adopt a rather unspecified umbrella term of potentially "harmful speech". Our goal is not conceptual analysis. Instead, we seek to circumscribe a sufficiently comprehensive linguistic tool in order to gather empirical data about different varieties of harmful speech

1 Though not concerned with online communication, arguments to this effect have been made in the philosophical literature; see (McGowan, 2019; Tirrell, 2012, 2018).

2 For illustration of our taxonomy and in order to understand the nature of the platform channel that we analyze, this paper contains examples of harmful language. We cite these examples as sparingly as possible, and all are attested in the corpus.

in online communities. To this end, we shall provide a comprehensive taxonomy of harmful language which will help us detect how it is used in online (in-group) discussions.

One practical approach to analyzing online harm is by studying text corpora containing harmful language; either general corpora from websites, online forums or social media, or specific corpora collected around an event of interest (such as discussions of the European refugee crisis in 2015, which triggered large amounts of xenophobic and racist hate speech on the internet). The public repository hatespeechdata[3] has already collected several dozen text corpora with hate speech in different languages (Vidgen & Derczynski, 2020).

A first problem of this work is the narrow focus on hate speech, while disregarding other types of harmful speech. There is a further challenge with using empirical data for this research, which is the lack of diversity of topics, platforms, and collection methods studied. Most existing corpora of harmful language were collected opportunistically from easily accessible media, mainly Twitter (Poletta et al., 2020; Vidgen & Derczynski, 2020). For practical reasons, many datasets are further restricted to certain specific domains (sexism, anti-immigrant rhetoric). Finally, many corpora and annotations concentrate on largely overt forms of hate speech and offensive language. In this paper, we extend this analysis to more indirect forms of harmful language, which may be critical to creating an in-group community where such speech becomes common currency. In particular, we seek to establish an empirical basis for implicit expressions of online harm in the context of supporters of former US president Donald Trump.

(2.1) TELEGRAM

Telegram is an encrypted mobile messaging platform which is popular as an alternative to less secure messengers, such as WhatsApp, in person-to-person communication. In addition, it offers private and public "channels" for one-to-many interactions. These channels can be created by any user and are often employed to share information or news; but they also serve as discussion forums by allowing responses. Due to the encryption features, Telegram has been used by extremist groups to spread their ideology and recruit users (Prucha, 2016; Yayla & Speckhardt, 2017; Shehabat, Mitew, & Alzoubi, 2017). As an additional feature to protect users, many channels regularly delete all posted content (e.g., performing daily purges) to make them unavailable (Baumgartner et al., 2020). Baumgartner et al. (2020) provide a large snapshot of raw data from public channels on Telegram, collected by bootstrapping from a seed list of channels. However, they do not specifically analyze the language included in this large sample. We are not aware of any previous Telegram corpora addressing offensive or harmful language. Following the recommendations by Vigden & Derczynski (2020), we create a new dataset from a Telegram channel which we expect to contain a significant amount of harmful speech. We now briefly introduce our working definition and taxonomy for the analysis.

(2.2) ONLINE HARM AND HATE SPEECH

In the legal context of regulating discrimination, two key terms are prominently used as a term of art: "hate speech" and "dangerous speech". They are core landmarks in guiding policy for detection and regulation of online harmful content. There are two challenges to this project. One is a definitional problem: we lack a univocal definition of what exact forms of speech count as hate/dangerous speech (Brown, 2017; Benesch et al., 2018; Gelber, 2019). The other challenge is the legal problem of establishing under what conditions the harm achieved is subject to legal protection (Bleich, 2011; Waldron, 2014; Oster, 2015; Heinze, 2016; Howard 2019). The two challenges are compounded in the context of online communication.

In the context of dissemination through the media, the European Commission against Racism and Intolerance has updated an internationally adopted definition of "hate speech":

> "Hate speech entails the advocacy, promotion or incitement to denigration, hatred or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatization or threat to such persons on the basis of a non-exhaustive list of personal characteristics or status that includes race, colour, language, religion or belief, nationality or national or ethnic origin, as well as descent, age, disability, sex, gender, gender identity and sexual orientation." (ECRI, 2016)

3 *http://ckan.hatespeechdata.com/*.

Even with an internationally adopted definition on the table, there is no guarantee of a conceptual consensus of what falls within the category of "hate speech". Since our goal is not conceptual analysis, we take the philosophical and legal assumptions as a starting basis, rather than arguing for them here. We follow Brown (2017, 2019) in thinking of "hate speech" as an "opaque idiom signifying a heterogenous collection of expressive phenomena" (Brown, 2019: 208). He thus suggests it might be more useful to think of "hate speech" as a "family resemblance" rather than a concept unified by a single essential feature common to all discursive/expressive phenomena typically labelled as such.

To avoid such conceptual controversies, here we adopt the folk, ordinary concept of *harmful speech*. We intend this as a more encompassing category that includes "hate speech" as a sub-type,[4] but also more subtle types of language use that may lead to harmful consequences. Since we are not concerned with providing a definition of harmful speech, we instead focus our discussion on a set of key features which are often discussed in the philosophical literature as hallmarks of "hate speech" and harmful speech more generally—namely, that it is a form of speech which:

(1) subordinates and disproportionately harms vulnerable target groups, e.g., (historically) oppressed, disadvantaged, marginalized, minority groups identified by marked characteristics (e.g., race, religion, ethnicity, sexual orientation, gender identity or disability, etc.). (Matsuda et al., 1993, Langton, 2012; Maitra, 2012; McGowan, 2019; Popa-Wyatt & Wyatt, 2018).

(2) justifies, glorifies, and incites to violence and hatred against target groups; serves to provoke, stir up hatred, harass, threaten and advocate discrimination, vilify, intimidate, defame (Matsuda et al., 1993; Tirrell, 2012; Oster, 2015).

(3) recruits and enables by-standers through propaganda espousing the inferiority of socially marked groups; promotes racial discrimination, hatred and persecution (Langton, 2012; Tirrell, 2012; Stanley, 2015).

(4) is socially divisive, reinforcing in-group vs. out-group views, which over time may lead to conflict and genocide (Brown, 2017; Tirrell, 2012).

(5) undermines the reputation, social standing, and assurance of target groups that they deserve to be treated as equal citizens (Waldron, 2014).

We take these features as guiding our discussion of harmful speech. Our goal is to establish an empirical basis for the varieties of harmful speech in online communication rather than arguing about their specific effects. To this end, we seek to provide a comprehensive taxonomy of various forms of expressions used to cause harm as outlined under (1)–(5). The difficulty with operationalizing such a taxonomy is that various forms of speech may exhibit more than one of the above features. Thus, determining which category they belong to will sometimes be a contextual matter.

To classify the data in our corpus, we draw on a very comprehensive classification of the varieties of pejoratives, recently proposed by Jeshion (2021). Jeshion distinguishes pejorative lexical items (which are pejorative in virtue of their conventional meaning), on the one hand, from *pejorative uses of words* (e.g., "boy" used to refer to an African-American man), and on the other hand from *pejorative speech acts*, which may occur independently of an utterance of pejorative lexical items or any words used pejoratively at all. In working with the data in our corpus, we also consider forms of speech that may be qualified as assaultive speech and incitement to violence (which are typically the hallmark of the legal concept of "hate speech"). In addition, we include indirect forms of harmful speech such as divisive rhetoric and code words which serve to establish and reinforce an in-group identity and community. Our proposed taxonomy has thus the advantage of expanding the scope of harmful speech from pejoratives and offensive language (which are usually the focus of existing datasets) to include both direct and indirect forms of harmful speech. Overall, we identify five major categories. The full taxonomy, including subcategories, is listed in ***Table 1***.

Category I includes expressions of extreme or dangerous speech, assaultive speech, and language glorifying or inciting to violence (e.g., "just burn in the sun", "DEATH TO CHINA", "violence is 100000% justified now", "Perhaps if trump is installed as dictator, we should send libtards to concentration camps").

4 Mentions of "hate speech" below are unavoidable given that the tools in computational linguistics that we use for the detection of online harmful speech have adopted the term "hate speech" as a key category.

I. incendiary speech (assaultive speech, extreme speech, dangerous speech, the glorification of violence)
II. pejorative words and expressions - dehumanizing - canonical slurs - descriptive slurs - gendered slurs and expressions - pejorative nicknames - stereotyping expressions - pejorative words used pejoratively - expletives - swear words - generic pejoratives
III. insulting/abusive/offensive uses - jokes - insulting rhetorical questions - insulting metaphors - inventive abusive uses - non-pejorative words used pejoratively
IV. in/out-group (divisive speech)
V. codes

Table 1 Taxonomy of online harm used in manual annotation.

Category II includes pejorative expressions which are derogatory in virtue of their conventional meaning. These are evaluative terms which serve to express a speaker's feelings or attitudes towards the target. Subtypes include canonical slurs that harm individuals of minority and oppressed groups simply in virtue of their group membership, group-based slurs based on a political affiliation (e.g., "commie", "libtard"), and other types of descriptive and gendered slurs, pejorative nicknames, etc. It also includes straightforward pejoratives which target individuals based on personal properties (e.g., "scum", "idiots", "retarded"), as well as expletives (e.g., "damn"), and swear words (e.g., "fuck", "shit"), among others.

In contrast, category III includes expressions that are being *used* derogatively, but are not inherently pejorative in their conventional meaning. This includes jokes and inventive forms of speech, which serve to put down specific individuals or groups (e.g., "DemoRATS", "Commiefornia"), insulting metaphors (e.g., "they are a sickness"), as well as non-pejorative words when used pejoratively in context (e.g., "Jews", "gay").

Category IV includes expressions which cause harm more implicitly by "othering" (Culpeper, 1996; Palmer et al., 2020) another (target) group. This covers general divisive rhetoric which serves to create and reinforce in-group vs. out-group distinctions (e.g., "The Chinese are godless society they operate on like the mafia... unless they are Christian can be trusted", "Are women banned from this chat? If not, why the fuck not?", "The left will kill 100 million Americans if they ever get the chance", "I mean if BLM is a thing. Why not WLM").

Finally, category V includes coded expressions that mark in-group affiliation within a like-minded community of people able to decode the expression (e.g., "Trump train", "GIVE THAT MAN A BRICK", "glow", "fedposter", "when shit hits the fan", "Patriot").

Further examples can be found in the data repository, as well as in the annotation guidelines provided there.

With this taxonomy as our starting point, we now proceed to empirically ground it by illustrating various forms of harmful speech with a corpus from a Telegram channel. Our goal is to provide a qualitative and quantitative analysis of the corpus we gathered and thus provide a basis towards improving tools for automatic detection of online harmful speech.

(3) METHODS AND DATA

For our study, we created a new corpus from one Telegram channel used by supporters of former US President Donald Trump, which covers the period from December 11, 2016, to January 18, 2021 (Solopova, Scheffler, & Popa-Wyatt, 2021). After removing empty messages, the dataset consists of 26,431 messages, produced by 521 distinct users.

(3.1) MESSAGES AND USER ACTIVITY

We start our analysis by providing surface statistics on the vocabulary composing the dataset. We measured the average length of the message in tokens and characters (12.65 and 76.11, respectively). This means that Telegram messages are on average comparable to tweets, which average 11–14 tokens and 70–84 characters each (Boot et al., 2019).

Reflecting Telegram's status as a news sharing platform, we note a number of messages containing links (454), many of which consist of only the link without any other textual content (333). Similarly to Twitter, Telegram allows user referencing with the "@" sign (699), but these mentions appear mostly in reposted tweets and are not used much for communication among the channel participants, probably due to the higher level of anonymity this social network provides.[5] However, @ is used more frequently than hashtags (only 209), which are an integral part of Twitter.

As we can observe in ***Table 2***, the number of messages is not homogeneous throughout the 4.5 years. Although this dataset covers only 18 days of 2021, this is the most active period, recording 15,603 messages. This sudden spike in new Telegram users suggests the possibility of a "migrating" effect,[6] due to the temporary closure of Parler, the ban on Reddit's r/The_Donald, and the introduction of new Discord policies.[7] ***Figure 1*** shows this trend. 2021 records the highest number of new users and the highest number of old users reviving their activity. It's difficult to assess the activity between 2016–2018, because we start with December 2016 when the channel was created, and also because in 2018 the channel had massive message deletion, which is reflected both in the small number of messages posted and of new users.

	2016	2017	2018	2019	2020	2021
Number of messages	410	2601	1456	5865	4059	15,603
Max. daily messages	75	49	22	449	663	3,696
Most active day	Dec. 14	Feb. 8	Sep. 16	Jun. 27	Dec. 23	Jan. 9

Table 2 Annual message statistics.

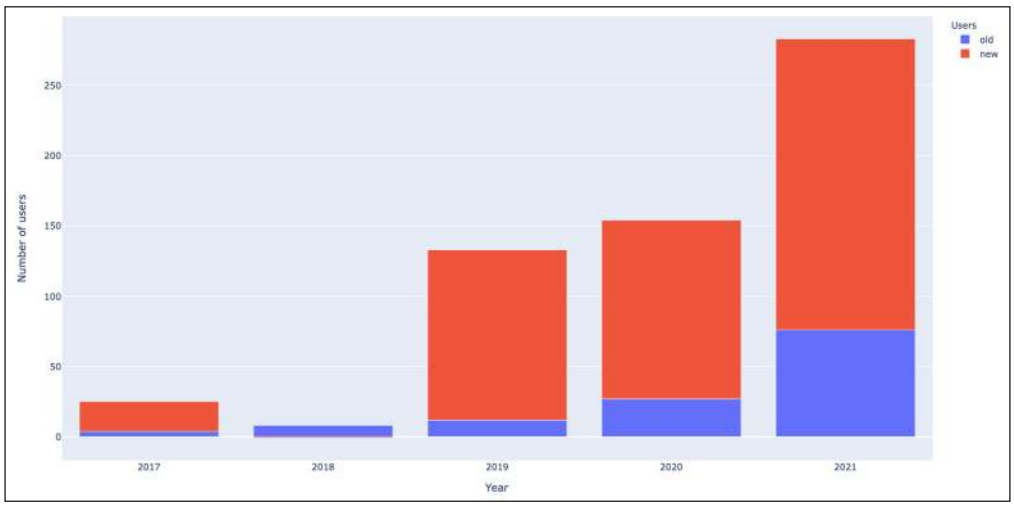

Figure 1 User statistics per year: newly added users and users active the previous year.

The period between 2019–2020 shows interesting trends. 2019 is more active in the number of messages, but less in the number of new users. This is perhaps because of an increase of public discussion on topical issues, e.g., the US government shutdown and the state of national emergency in order to secure funds for the Southern border construction. The activity on June 27, 2019 is associated with the r/The_Donald subreddit being quarantined by Reddit admins due to excessive reports and threatening of public figures in the context of the 2019 Oregon Senate Republican walkouts. The subreddit lost revenue opportunities and was removed from feeds and search, which outraged a lot of its users. Thus, more intense activity on the channel seems to correlate with responses to provocative tweets from the @realDonaldTrump account.

5 In contrast, over 20% of tweets are replies in some Twitter corpora (Scheffler, 2014).

6 See Chandrasekharan et al. (2017) discussing migrating effects following bans on social media.

7 *https://www.nytimes.com/2017/08/15/technology/discord-chat-app-alt-right.html*.

2020 marked a gradual increase in activity up to 2021 (see *Figure 2*), which tends to be concerned with the COVID-19 pandemic, though this is not a topic of high priority for this group. December 23, 2020, records the highest activity, with discussions related to Donald Trump's issuing of pardons and commutations, and tweeting about the "stolen election" ("This was the most corrupt election in the history of our Country, and it must be closely examined!" — Donald J. Trump (@realDonaldTrump)). During this period, there is also the first evidence of planning an assembly on January 6, 2021. January 9, 2021, marks the highest number of messages overall (3,696), which reflects discussions in the aftermath of January 6, the Capitol Hill insurrection. *Figure 3* shows that this activity started on January 7, gradually growing to the 9th and then decreasing to its usual average on the 16th.

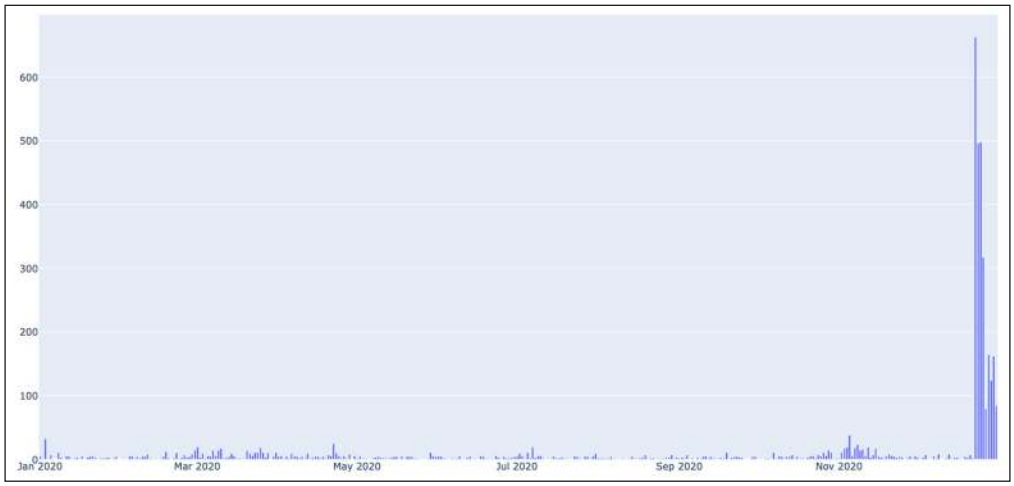

Figure 2 Messages per day in 2020.

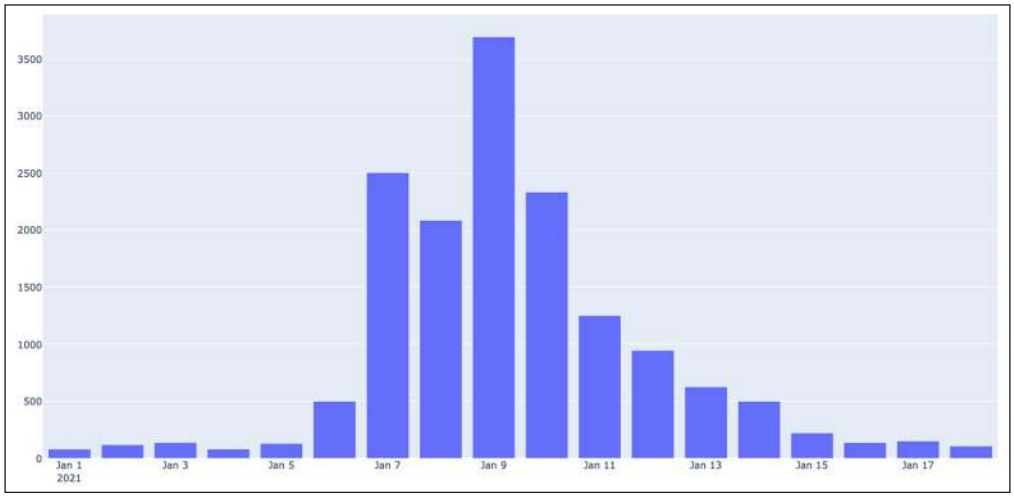

Figure 3 Messages per day in 2021.

We now turn to describing first our automatic annotation of the entire corpus, and then our manual annotation of a subset of the messages.

(3.2) OFFENSIVE LANGUAGE OVER TIME

As described by Solopova, Scheffler, and Popa-Wyatt (2021), we automatically annotated the entire corpus by using methods of surface-matching to two kinds of open source word lists (Anger, 2017; Shutterstock, 2020). One list focuses on the detection of offensive language; the other on the detection of messages with higher risk of offensive language. Both lists are commonly used to detect subtypes of harmful language in online contexts. In both cases, the focus is on explicit forms of harmful speech, especially pejorative words and expressions in virtue of their conventional meaning (i.e., category II in our taxonomy). This enabled us to detect some of the categories of our taxonomy, e.g., offensive uses (44 messages), codes (24 messages), incendiary speech (29 messages); see *Table 4*. However, while useful in this respect, we shall argue that the word lists have serious limitations in detecting implicit forms of harmful speech.

Further, it is worth noting that the second type of word list, designed to detect messages with higher risk of offensive language, contains not only offensive terms, but also neutral words occurring in contexts more prone to harmful uses. This may explain why surface-matching methods tend to create false positives when target words occur in otherwise neutral discussions (see ***Table 4***). Furthermore, these word lists were unable to detect implicit as well as creative uses of harmful speech, thus resulting in false negatives. These are important findings in themselves, since given that such word list methods are commonly used by administrators of social media groups and channels, it is important to draw attention to their limitations.

To render our detection methods more robust, in a next step of analysis we also applied the open-source automated hate-speech/offensive-language detection library HateSonar (Nakayama, 2017). This is a detection method trained with neural networks and contextual embeddings, and is one the newest automated tools available. It assigns the labels "hate speech", "offensive", or "neutral" to input text. We also measured the attributed tags quantitatively for each year. Our results show that the messages labelled as harmful speech by the word lists, on the one hand, and by the automatic hate speech detection, on the other, do not overlap. In short, there was no convergence in tagging the same messages using word list methods and machine-learning-based methods. This suggests that greater care in evaluating the findings using these methods is needed.

Mirroring the increase in user activity from 2019 to 2021, we also found both a quantitative and qualitative increase in the use and varieties of harmful speech, as flagged by the automatic methods (***Figure 4***). We should note, however, that the overall ratio of the different types of (automatically annotated) harmful speech remains constant across the three years (around 17% of messages).

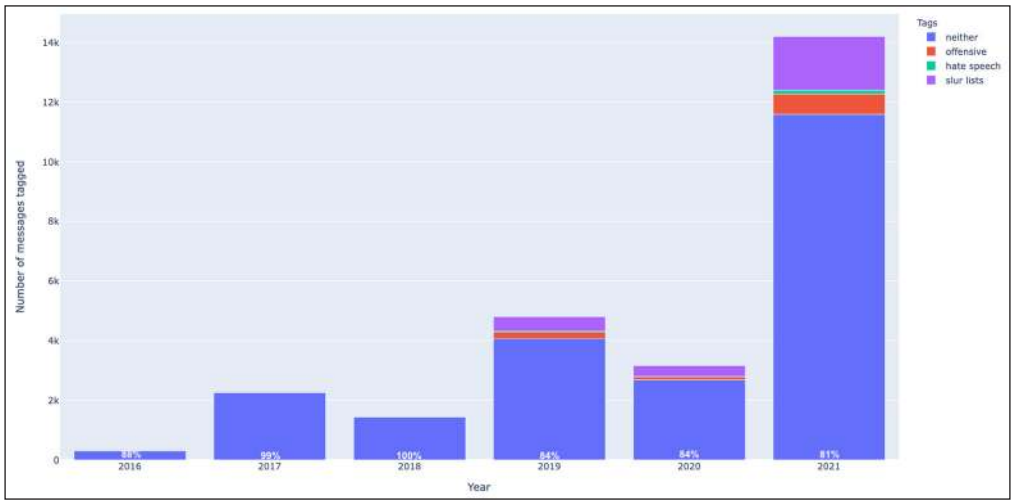

Figure 4 Automatically predicted offensive language labels over time. White numbers show the fraction of the 'neither' tag.

Critically, we also note that the list-matching methods flag more messages than the machine learning classifier. One possible reason for this is that the machine learning tool assigns the offensive tag almost 3 times more frequently than the hate speech tag. This might be because, according to Davidson et al. (2017), hate speech is marked by a high intensity level of "offensiveness". Another reason might be the type of data these methods are trained on. For example, the machine learning tool was trained on Twitter data, which differs in many ways from Telegram messaging. These differences may be the reason why its performance on our data is less optimal.

(3.3) MANUAL ANNOTATION

One of the authors manually annotated about one fifth of the data according to the taxonomy established in Section (2.2). We chose the time period from November 1, 2020 to January 7, 2021, to cover discussions and reactions related to the 2020 US election, including the January 6 Capitol riot. We also added January 9, the most active day of our corpus. This tallied 4,505 messages for annotation. We used the *BRAT* tool (Stenetorp et al., 2012), which enables manual annotation of both token-level instances and message-level labels. This is useful

because it is expected to provide more fine-grained data for analysis of linguistic markers of harmful language.

In order to validate the annotation schema and determine the difficulty of the task, we chose a continuous thread of 711 messages from the most active day (January 9, 2021), to be re-annotated by another linguist. This second annotator was provided with the taxonomy and examples for the individual categories. After annotation of 200 messages, we discussed the taxonomy and several difficult cases with both annotators, before the re-annotation was completed. We computed inter-annotator agreement between the two independent annotations on a message level, first on the binary decision task (namely, "Should the message be flagged as containing harmful language?"), and then taking our 5 top-level categories into account.

The annotators show substantial agreement (Cohen's $\kappa = 0.70$) on the binary harmful/neutral distinction. When evaluating the 6-way classification (5 categories plus "neutral"), we counted any overlap in categories of harmful language between the two annotators as agreement (i.e., if one annotator found only category II, pejoratives, while the other annotator found both II and III, we counted the message as an agreement between annotators). The fine-grained distinction leads to an overall agreement of $\kappa = 0.65$, which is promising given the difficulty and subjective nature of the task. Clearly, more comprehensive annotation guidelines and a dedicated adjudication process between annotators is likely to raise the inter-annotator stability of our taxonomy categories. However, we leave this further evaluation for future work.

(4) ANALYSIS AND DISCUSSION

Overall, out of the chosen 4,505 messages, we manually identified as harmful language a total of 787 messages and 831 instances (i.e., phrases). This means that 44 messages contained more than one instance of harmful language. Among the major categories, category II "pejorative words and expressions" is the largest (273 messages). Category V "codes" is the second-largest (261 messages). Category III "offensive uses" and I "incendiary speech" are roughly similar in size (115 and 98 messages, respectively), while category IV "in/out-group", which is the most complicated by definition and can often coincide with other classes, is the smallest (40 messages). ***Table 3*** summarizes the distribution across the categories. Note that we found 310 instances (37 of which appear more than once in one message) which we annotated as "pejorative words and expressions", and 121 instances which we annotated as "insulting/offensive/abusive uses" (only six times in the same message). The instance level statistics on the other categories coincide with those on the message level.

TAG	I. INCENDIARY	II. PEJORATIVE	III. OFFENSIVE USES	IV. IN/OUT-GROUP	V. CODES	ALL
Number of messages	98	273	115	40	261	787
Fraction	12%	35%	15%	5%	33%	100%

Table 3 Statistics on 5 main categories: incendiary speech, pejorative words and expressions, insulting/offensive/abusive uses, in/out-group, code words.

Figure 5 shows the distribution of sub-categories of II and III in the manually annotated subcorpus. Starting with the largest category of "pejoratives" (II), this includes among the most frequent sub-categories: pejorative words used pejoratively (85 messages, 88 instances), pejorative nicknames (90 messages, 97 instances), and swear words in offensive contexts (56 messages, 60 instances). This category also includes canonical slurs (27 messages), generic pejoratives (12), descriptive slurs (9), gendered slurs and expressions (6), stereotyping expressions and expletives (2). We found no examples of dehumanizing speech. In the category of "offensive uses" (III), the largest sub-categories include non-pejorative words used pejoratively (39 messages) and insulting metaphors (38 messages). We also found offensive jokes (21 messages), inventive offensive instances (21), and insulting rhetorical questions (6).

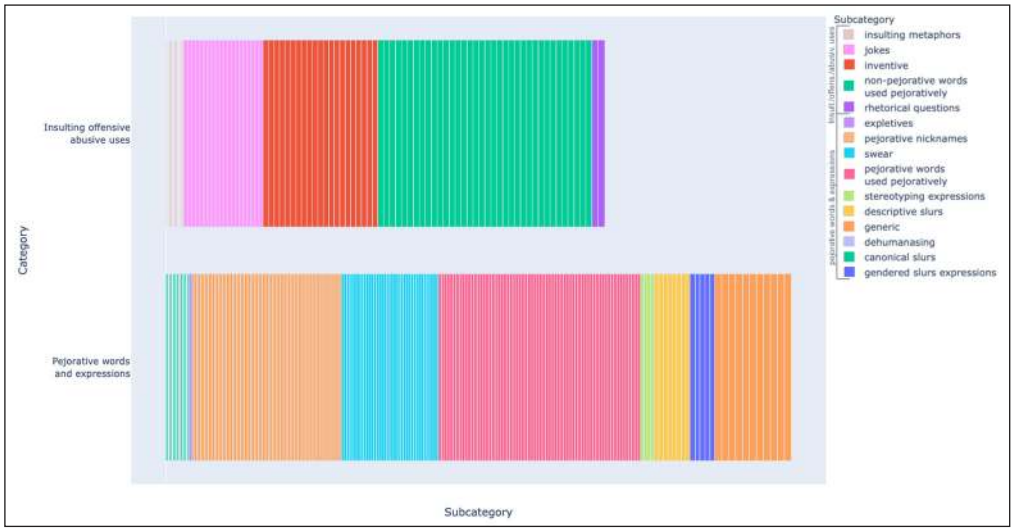

Figure 5 Sub-category statistics for pejorative words and expressions and insulting/offensive/abusive uses categories.

(4.1) COMPARISON OF MANUAL AND AUTOMATED ANNOTATIONS

We performed a binary and multi-class comparison of the manual and automated annotations.[8] The binary comparison (see ***Figure 6***) shows how many manually marked messages containing harmful language are also identified by either word-list-based or machine-learning-based automated methods. Out of the 4,505 messages in the manually annotated subcorpus, 3,395 remained untagged both by the manual annotation and the automated one (true negatives). Only 275 harmful messages have been correctly tagged by the automated systems as either offensive or hate speech (true positives). Consequently, we have 835 messages tagged differently by different methods.

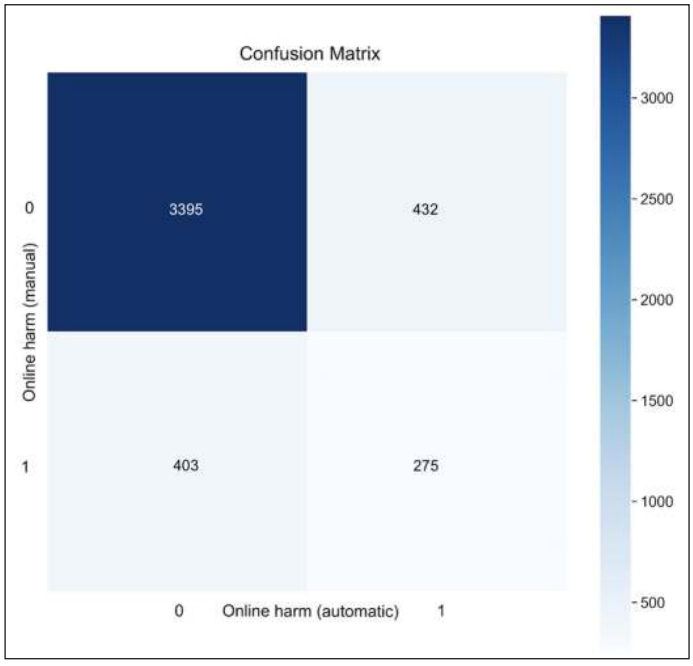

Figure 6 Binary confusion matrix for manual and automated annotation.

Specifically, the HateSonar classifier or offensive word lists falsely tagged some messages as offensive/hate speech (432 false positives), or failed to find a manually identified harmful message offensive (403 false negatives). Using these numbers, we calculated the balanced F-measure for the binary classification for the joint performance of word lists and HateSonar as 40% (with 39% precision and 41% recall). These low scores confirm the low generalizability of solutions for detection of offensive language when applied to a new dataset (Yin & Zubiaga, 2021). One reason is that our data originates from a different platform than the training data. In addition, our data contains many more different types of complex and implicit harmful language, which may be too difficult for automated detection.

8 In the following, we reserve "harmful language" for our own manual annotations, and call the automatic annotations "offensive language", since this is their main target label.

These results demonstrate a need for additional corpora of harmful language which would be able to provide more diverse and fine-grained annotations, and which provide necessary context for identifying harmful speech. This is illustrated in the multi-class confusion matrix (***Table 4***). Across our five categories of harmful language, the word lists identified more harmful messages, most of which overlap with our category of pejorative words and expressions. However, the list-based detection also flagged many neutral messages as harmful (317), which indicates low precision. At the same time, the HateSonar machine learning system had low recall in our data, whereas its tag of hate speech had better precision. Overall, we can conclude that existing automated tools are generally better at finding neutral messages. We contend this is a by-product of the prevalence of neutral messages in the data.

Table 4 Confusion matrix of multi-class annotation.

MANUAL\AUTO	WORD LISTS	HATE SPEECH	OFFENSIVE	NEITHER	RECALL
I incendiary	29	0	8	61	0.378
II pejorative	85	13	30	89	0.590
III offensive uses	44	3	10	54	0.514
IV in-/out-group	12	5	4	18	0.538
V codes	24	0	8	181	0.150
neither	317	11	104	3395	0.887
precision	0.380	0.656	0.366	0.894	

In a further step, we compared the automatically assigned tags to the categories of our taxonomy on a message basis. We found that recall is much higher across the automatic methods (the automatic tags do not overlap in any messages) for our categories of pejorative expressions (II), offensive uses (III), and in/out-group divisive speech (IV). Incendiary speech (I) and codes (V) are, however, harder to detect with current automatic methods, given the context-sensitivity of such uses.

(4.2) EXPLORATORY ANALYSES

In a next step, we explore the extent to which use of harmful speech can create a toxic environment in which more harmful content becomes not only permissible but tolerated.[9] We base this analysis on the noisy automatic offensiveness labels available for the entire corpus. We found that 24,629 message-response pairs, which is 93% of all messages, are both neutral. There are also 836 (3%) neutral responses to offensive messages, 585 (2.2%) offensive responses to a neutral message, and only 327 (1.2%) are both offensive. This means that while neutral messages receive an offensive response in about 2.3% of cases, this chance increases in offensive messages by more than a factor of 10, to 28.1%. Due to the large number of messages, this is a highly significant result with a moderate effect size (χ^2 = 2,208.64, Cramér's V = 0.28). However, as noted above, the automated detection methods for offensive language have low accuracy, so this research needs to be repeated after a revision of the automated tags in order to confirm the results. Nevertheless, these initial results suggest that these data may be used for studying whether harmful speech creates a hospitable environment that is inducive to further harmful actions and practices (Tirrell, 2017).

We also explored the hypothesis that Trump's narrative may be reflected in his supporters' language. We analyzed quantitative and semantic similarities between our Telegram corpus and the Trump Twitter Archive (Brown, 2016), containing all tweets posted by Donald Trump since 2016. However, we found a negative correlation (Spearman's ρ = –0.098) for overall activity. Thus, our hypothesis that more tweets would correlate with more Telegram messages is disconfirmed. However, such a measurement is inherently difficult, given that often discussions on Telegram can follow tweets with at least a day delay (e.g., discussions on June 27, 2019, following the r/The_Donald ban on June 26, or discussions on January 7–9, following the Capitol riot on January 6). Other days, however, don't register any channel activity, so our data set cannot measure specific delayed responses.

9 Tirrell (2017) argues that the phenomenon of hate speech or toxic speech more generally can be thought of by analogy to the spread of a toxin.

Finally, we started analyzing the semantic similarity of our corpus and Trump's tweets using contextual embeddings (word2vec, Mikolov, et al., 2013; and fastText, Bojanowski, et al., 2016). We compared semantically similar words for keywords of Republican narrative (e.g., "guns", "China", "immigrants", etc.) for a vector model trained on Trump tweets and another on our Telegram corpus. We only found a semantic similarity for "Antifa", "socialist", "communist", and "masks". The latter is qualified by both as "illegal" and "leftists". The first three are semantically associated with "funded" and "criminal", being used interchangeably, and also strongly correlated with the Democratic party. We believe that building on this preliminary method will allow us to measure how message embedding vectors annotated according to our taxonomy differ from the neutral vocabulary.

(5) SUMMARY

We presented a new corpus of a channel of the instant messaging platform Telegram, which we chose for its large potential for harmful language. The corpus consists of over 25,000 messages spanning a period of over 4 years, and includes discussions leading up to and following the January 6, 2021, US Capitol riot. We argued for a broad notion of online harmful speech, which includes not only direct attacks and pejorative expressions, but also divisive rhetoric and code words which seek to establish and reinforce an in-group identity and sense of community in order to justify attacks against target groups. To this end, we introduced a comprehensive taxonomy of harmful speech and provided manual annotations on a subset of our corpus. Comparing these annotations with automatically obtained labels of "hate speech" and offensive language (two commonly studied subsets of harmful speech), we showed that both lexicon-based methods and machine learning algorithms trained on other datasets and platforms are unable to fully detect the varieties of harmful speech that we encounter on the Telegram channel. This suggests that much more research is needed both to ground the philosophical foundations of harmful speech so that we operate with a consensual definition, and to improve our tools for detecting the linguistic manifestations of harmful speech.

Our contribution to this research consists in diversifying the available empirical data in terms of the platform, content, and the kinds of annotations we cover. In contrast to previous work, which focuses mostly on personal derogation, we have distinguished between various forms of harmful speech, some of which include pejorative expressions which conventionally express negative attitudes, while others include pejorative uses of non-pejorative expressions, which are used to attack or put down a person or group, as well as more direct forms of harmful speech inciting to hatred and violence. In addition, we have also identified group-internal codes (i.e., not intended to be addressed to or understood by outsiders) and divisive rhetoric which seeks to reinforce in-/out-group distinctions. We argued that these latter categories are not easily identified with word lists, and thus pose additional challenges for detection. We invite researchers from related fields to use our data to further address the question of what constitutes online harm, how to detect it, and how to mitigate it.

SUPPLEMENTARY FILES

Solopova, V., Scheffler, T., & **Popa-Wyatt, M.** (2021). A Telegram corpus for hate speech, offensive language, and online harm. *Journal of Open Humanities Data*, 7: 9, 1–5. DOI: *https://doi.org/10.5334/johd.32*

Data repository: *https://osf.io/ck3gd/*

ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewers for their detailed and helpful comments. In addition, we are grateful to the audience of the workshop on "Oppressive Speech, Societies, and Norms" at ZAS Berlin for discussion.

FUNDING STATEMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 841443. (HaLO project—"How Language is Used to Oppress"). The project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1412, 416591334.

We acknowledge support by the Open Access Publication Funds of the Ruhr-Universität Bochum.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Tatjana Scheffler: Conceptualization, data curation, formal analysis, methodology, supervision, validation, writing – original draft, writing – review & editing, funding acquisition.

Veronika Solopova: Data curation, formal analysis, investigation, methodology, resources, software, visualization, writing – original draft, writing – review & editing.

Mihaela Popa-Wyatt: Conceptualization, methodology, writing – original draft, writing – review & editing, funding acquisition.

AUTHOR AFFILIATIONS

Tatjana Scheffler *orcid.org/0000-0001-7498-6202*
German Studies, Ruhr-Universität Bochum, Bochum, Germany

Veronika Solopova
Dahlem Center for Machine Learning and Robotics, Freie Universität Berlin, Berlin, Germany

Mihaela Popa-Wyatt *orcid.org/0000-0001-9239-9247*
Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin, Germany

REFERENCES

Anger, Z. (2017). List of profanity in English. Retrieved from *https://github.com/zacanger/profane-words*

Baumgartner, J., Zannettou, S., Squire, M., & **Blackburn, J.** (2020). The Pushshift Telegram dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*(1), 840–847. Retrieved from *https://ojs.aaai.org/index.php/ICWSM/article/view/7348*

Benesch, S., Buerger, C., & **Glavinic, T.** (2018). Dangerous speech: a practical guide. *http://dangerousspeech.org*. DOI: *https://doi.org/10.15868/socialsector.34064*

Bleich, E. (2011). The rise of hate speech and hate crime laws in liberal democracies. *Journal of Ethnic and Migration Studies*, *37*(6), 917–934. DOI: *https://doi.org/10.1080/1369183X.2011.576195*

Bojanowski, P., Grave, E., Joulin, A., & **Mikolov, T.** (2016). Enriching word vectors with subword information. *arXiv:1607.04606*. DOI: *https://doi.org/10.1162/tacl_a_00051*

Boot, A. B., Tjong Kim Sang, E., Dijkstra, K., et al. (2019). How character limit affects language usage in tweets. *Palgrave Commun*, *5*(76). DOI: *https://doi.org/10.1057/s41599-019-0280-3*

Brison, S. J., & **Gelber, K.** (2019). *Free Speech in the Digital Age*. Oxford: Oxford University Press. DOI: *https://doi.org/10.1093/oso/9780190883591.001.0001*

Brown, A. (2017). What is hate speech? Part II: Family resemblances. *Law and Philosophy*, *36*, 561–613. DOI: *https://doi.org/10.1007/s10982-017-9300-x*

Brown, A. (2019). The Meaning of Silence in Cyberspace: The Authority Problem and Online Hate Speech. In S. J. Brison & K. Gelber (Eds.), *Free Speech in the Digital Age* (pp. 207–223). Oxford University Press. DOI: *https://doi.org/10.1093/oso/9780190883591.003.0013*

Brown, B. (2016). Official Trump Twitter Archive V2 source. Retrieved from *https://www.thetrumparchive.com*

Chandrasekharan, E., Pavalanathan, U., Srinivisan, A., Glynn, A., Eisenstein, J., & **Gilbert, E.** (2017). You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proceedings of the ACM on Human-Computer Interaction*, *1*(2), 1–22. DOI: *https://doi.org/10.1145/3134666*

Culpeper, J. (1996). Towards an anatomy of impoliteness. *Journal of Pragmatics*, *25*(3), 349–367. DOI: *https://doi.org/10.1016/0378-2166(95)00014-3*

Davidson, T., Warmsley, D., Macy, M., & **Weber, I.** (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1). Retrieved from *https://ojs.aaai.org/index.php/ICWSM/article/view/14955*

ECRI. (2016). General Policy Recommendation No. 15 On Combating Hate Speech, December 8, 2015, Strasbourg. Retrieved from https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01

Gelber, K. (2019). Differentiating hate speech: a systemic discrimination approach. *Critical Review of International Social and Political Philosophy*, *24*(4), 394–414. DOI: https://doi.org/10.1080/13698230.2019.1576006

Heinze, E. (2016). *Hate Speech and Democratic Citizenship*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780198759027.001.0001

Howard, J. (2019). Dangerous Speech. *Philosophy & Public Affairs*, 47, 208–254. DOI: https://doi.org/10.1111/papa.12145

Jeshion, R. (2021). Varieties of pejoratives. In J. Khoo & R. Sterkin (Eds.), *Routledge Handbook of Social and Political Philosophy of Language* (pp. 211–231), New York: Routledge. DOI: https://doi.org/10.4324/9781003164869-17

Langton, R. (2012). Beyond belief: Pragmatics in hate speech and pornography. In I. Maitra & M. K. McGowan (Eds.), *Speech and harm: Controversies over free speech* (pp. 72-93). Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199236282.001.0001

Maitra, I. (2012). Subordinating speech. In I. Maitra & M. K. McGowan (Eds.), *Speech and harm: Controversies over free speech*. DOI: https://doi.org/10.1093/acprof:oso/9780199236282.003.0005

Matsuda, M., Lawrence, C., Delgado, R., & **Crenshaw, K.** (Eds.). (1993). *Words that wound: Critical race theory, assaultive speech, and the first amendment*. Colorado: Westview Press.

McGowan, M. K. (2019). *Just words: on speech and hidden harm*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/oso/9780198829706.001.0001

Mikolov, T., Chen, K., Corrado, G., & **Dean, J.** (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv*:1301.3781.

Nakayama, H. (2017). Hate Speech Detection Library for Python. Retrieved from https://github.com/Hironsan/HateSonar

Oster, J. (2015). Incitement to hatred. In *Media Freedom as a Fundamental Right* (Cambridge Intellectual Property and Information Law (pp. 223–240). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781316162736.013

Palmer, A., Carr, C., Robinson, M., & **Sanders, J.** (2020). COLD: Annotation scheme and evaluation data set for complex offensive language in English. *Journal for Language Technology and Computational Linguistics*, *34*(1), 1–28. Retrieved from https://jlcl.org/content/2-allissues/1-heft1-2020/jlcl_2020-1.pdf#page=11

Popa-Wyatt, M., & **Wyatt, J. L.** (2018). Slurs, roles and power. *Philosophical Studies*, *175*(11), 2879–2906. DOI: https://doi.org/10.1007/s11098-017-0986-2

Prucha, N. (2016). IS and the Jihadist information highway – Projecting influence and religious identity via Telegram. *Perspectives On Terrorism*, *10*(6). Retrieved from http://www.terrorismanalysts.com/pt/index.php/pot/article/view/556/1102

Poletto, F., Basile, V., Sanguinetti, M., et al. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, *55*, 477–523. DOI: https://doi.org/10.1007/s10579-020-09502-8

Scheffler, T. (2014). A German Twitter Snapshot. *Proceedings of LREC*, Reykjavik, Iceland. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1146_Paper.pdf

Shehabat, A., Mitew, T., & **Alzoubi, Y.** (2017). Encrypted Jihad: Investigating the role of Telegram app in lone wolf attacks in the West. *Journal of Strategic Security*, *10*(3), 27–53. DOI: https://doi.org/10.5038/1944-0472.10.3.1604

Shutterstock. (2020). List of Dirty, Naughty, Obscene, and Otherwise Bad Words. Retrieved from https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words

Solopova, V., Scheffler, T., & **Popa-Wyatt, M.** (2021). A Telegram corpus for hate speech, offensive language, and online harm. *Journal of Open Humanities Data*, *7*: X, 1–5. DOI: https://doi.org/10.5334/johd.32

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & **Tsujii, J.** (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. *Proceedings of the Demonstrations Session at EACL 2012*. *Association for Computational Linguistics*. https://www.aclweb.org/anthology/E12-2021

Tirrell, L. (2012). Genocidal language games. In I. Maitra & M. K. McGowan (Eds.), *Speech and harm: Controversies over Free Speech* (pp. 174–221). Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199236282.003.0008

Tirrell, L. (2017). Toxic Speech: Toward an Epidemiology of Discursive Harm. *Philosophical Perspectives* *45*(2), 139–161. DOI: https://doi.org/10.5840/philtopics201745217

Vidgen, B., & **Derczynski, L.** (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE*, *15*(12), e0243300. DOI: https://doi.org/10.1371/journal.pone.0243300

Waldron, J. (2014). *The harm in hate speech*. Cambridge, MA: Harvard University Press.

Yayla, A. S., & **Speckhard, A.** (2017). Telegram: The mighty application that ISIS loves. *International Center for the Study of Violent Extremism. Technical Report*. Retrieved from *https://www.icsve.org/telegram-the-mighty-application-that-isis-loves/*

Yin, W., & **Zubiaga, A.** (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. *arXiv preprint arXiv*:2102.08886. DOI: *https://doi.org/10.7717/peerj-cs.598*

TO CITE THIS ARTICLE:
Scheffler, T., Solopova, V., & Popa-Wyatt, M. (2021). The Telegram Chronicles of Online Harm. *Journal of Open Humanities Data*, 7: 8, pp. 1–15. DOI: *https://doi.org/10.5334/johd.31*

Published: 05 July 2021

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.