The Telegram chronicles of online harm

Tatjana Scheffler[a]* and Veronika Solopova[b] and Mihaela Popa-Wyatt[c].

[a]*German Studies, Ruhr-Universität Bochum, Bochum, Germany;* [b]*Dahlem Center for Machine Learning and Robotics, Freie Universität Berlin, Berlin, Germany;* [c]*Zentrum Allgemeine Sprachwissenschaft, Berlin, Germany*

Abstract

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

Harmful and dangerous language is frequent in social media, in particular in spaces which are considered anonymous and/or allow free participation. In this paper, we analyse the language in a Telegram channel populated by followers of Donald Trump, in order 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, but also more indirect ways of poisoning online discourse, such as divisive speech and the glorification of violence. We apply this taxonomy to a large portion of the corpus. Our data gives empirical evidence for harmful speech such as in/out-group divisive language and the use of codes within certain communities which have not often been investigated before. Second, we compare our manual annotations to several automatic methods of classifying hate speech and offensive language, namely list based and machine learning based approaches. We find that the Telegram data set still poses 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.

(1) Introduction

Digital media can cause harm in different ways. In addition to language that directly causes harm, such as bullying, hate speech attacks, and trolling, there are many more implicit avenues for online harm. These can include the spread of disinformation and the use of dehumanising, offensive, and incendiary language. Such language causes harm even when not directed at or read by individuals of target groups. It does so by poisoning public discourse, facilitating a networking platform for individuals with extreme views and thus creating a climate for normalising harmful practices with consequences beyond the online forums where they are primarily trafficked (Popa-Wyatt, forth.).

In this paper, we analyse a public channel from the direct messaging platform *Telegram*, which is rife with such indirect forms of offensive language. The channel is a platform where right-wing views are exchanged among like-minded people, expressing harmful views both explicitly and implicitly. Our data shows how extreme online discussions may lead to extreme actions, as the users of this channel 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 to finally discussing the aftermath of the event. We argue that the use of harmful language on this channel can serve to enable the feeling of group membership and thus facilitate the incitement of violent actions.

Our paper makes the following contributions: First, we apply several automatic annotations of hate speech and offensive language to our data, chronicling the prevalence of hateful language in this Telegram corpus. Second, we define a taxonomy of harmful expressions in online discussions which includes both direct and indirect forms. Third, we manually annotate a subset of the corpus with our taxonomy. Finally, we evaluate the currently available automatic methods of hate speech detection by comparing them with our manual annotations, giving pointers for future work.[1]

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

(2) Online harm and social media

Online hate 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 approach to analysing hate speech is by studying text corpora containing such 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[2] has already collected several dozen text corpora with hate speech in different languages (Vidgen & Derczynski, 2020).

We see two challenges with the existing datasets: (i) diversity of topics, platforms, and collection methods; and (ii) a lack of agreement of what constitutes "hate speech". We aim to address both in this contribution. As to (i), most existing corpora of hate speech were collected opportunistically from easily accessible media, mainly Twitter. For practical reasons, many datasets further are restricted to certain specific domains (sexism, anti-immigrant rhetoric). Finally, many corpora and annotations concentrate on quite overt forms of hate speech and offensive language. In contrast, we are interested in studying how even very indirect forms of harmful language in an online community can influence members' thinking and over time lead to concrete harms in society. In this paper, we are trying to establish an empirical basis for such implicit expressions of online harm in the context of supporters of former US president Donald Trump.

Challenge (ii) is the lack of a common definition of "hate speech", which is reflected in the various uses of terms such as "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). Here we shall adopt a broader umbrella term of '*online harm*'. Secondly, not all datasets provide clear definitions for what kind of language is considered offensive (Vidgen & Derczynski, 2020), and where definitions are provided, they often contradict each other. Here, we aim to give a taxonomy of the language of online harm as applies to our type of online (in-group) discussion.

[2] http://ckan.hatespeechdata.com/

(2.1) Telegram

Telegram is an encrypted mobile messaging platform which is popular as an alternative to less secure messengers such as What'sApp 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 (for example, 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 analyse the language included in this large sample. We are not aware of any previous Telegram corpora addressing offensive language or online harm. Following the recommendations by Vigden & Derczynski (2020), we create a new dataset from a Telegram channel for which we expect that it contains a significant amount of harmful and dangerous language. In the following subsection, we introduce our working definition and taxonomy for analysis.

(2.2) Online harm and hate speech

"Hate speech" and "dangerous speech" are two key terms in the legal context of regulating discrimination. This is particularly important when it comes to regulating online content. There are two challenges to this project. One is a definitional problem: we lack a univocal definition of what forms of speech count as hate/dangerous speech (Brown, 2017; Benesch et al., 2918; Gelber, 2019), which is to serve as a guiding policy for detection of online harmful content. The other challenge is a 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). Here we focus on the first problem, though our goal is not to settle a definition of hate/dangerous speech. Our goal is more modest: we shall provide a classification of various forms of hate/dangerous speech and illustrate them empirically with a corpus from a Telegram channel. The qualitative and quantitative analysis we provide below will help towards developing and improving tools for automatic detection of online harm.

It is difficult to define and circumscribe hate/dangerous speech because it includes a host of heterogeneous phenomena that share a certain number of common features. We follow Brown (2017) in taking hate speech to function more like a "family resemblance" rather than unified by a single essential feature common to all discursive phenomena typically labelled as hate speech. Among these features, we discern five:

(1) Hate speech disproportionately harms vulnerable target groups, e.g. (historically) oppressed, disadvantaged, marginalised, victimised members of society and groups of persons identified by certain characteristics that make them vulnerable and thus in need of protection (e.g. based on race, religion, ethnicity, sexual orientation, gender identity or disability, etc.). It does so by facilitating and perpetuating acts of subordination and oppression (Matsuda et al., 1993, Langton, 2012; Maitra, 2012; McGowan, 2019; Popa-Wyatt & Wyatt, 2018).

(2) Hate speech incites to violence and hatred, serves to provoke, stir up hatred, harass, threaten and advocate discrimination, vilify, intimidate, defame, and acts that serve to justify and glorify violence against target groups. This often correlates with hate crimes (see the category of "fighting words" which is legally protected in the US) (Matsuda et al., 1993; Tirrell, 2012; Oster, 2015).

(3) Hate speech recruits, encourages and enables by-standers, e.g. in the form of racist propaganda espousing the inferiority of certain races, which also may lead to promoting racial discrimination, hatred, violence, and persecution (Langton, 2012; Tirrell, 2012; Stanley, 2015).

(4) Hate speech is socially divisive and destructive of social cohesion in diverse societies, thus reinforcing in-group vs. out-group views. This may be seen to (likely to) cause a breach of peace, leading to conflict or even genocide (Brown, 2017; Tirrell, 2012)

(5) Hate speech undermines people's reputation and assurance that they are members of society in good standing and who deserve to be treated as equal citizens (Waldron, 2014).

The difficulty with operationalising the category of hate/dangerous speech is that various forms of speech may fall into more than one of the types above. So it will be a matter of context which functions are performed at any one time. Recent efforts of finding a consensual definition of hate speech come from the Council of Europe. In the context of hate speech disseminated through the media, the European Commission against Racism and Intolerance has updated an internationally adopted definition:

"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)

Clearly, the list of harms and wrongs caused by hate/dangerous speech is much wider than the features included in the above definition. We could add to this acts of disempowering, marginalising, silencing, insulting, disparaging, degrading, humiliating, disheartening, harassing, persecuting, threatening, provoking, inciting hatred, discrimination, violence, misrecognising, etc. Also, offensive acts of expressing (awakening) emotions and attitudes such as hate, contempt, dislike, disgust, despisement, aversion, loathing, antipathy, enmity, hostility, etc. may play a role in spilling over into acts of discrimination or violence. Note also acts of putting down, ranking targets as inferior, outcast and unwelcome. These might range as offensive and insulting acts rather than the more restricted category of hate/dangerous speech. Though clearly not everything that is offensive counts as hate/dangerous speech, it's important to keep these in mind because they all together contribute to poisoning the discourse in various communities, and further legitimate discriminatory speech and conduct. When anchored in a discourse of power, dominance and control, and incubated in like-minded communities particularly as in online echo-chambers, this may be the imperceptible beginning that leads from speech to action.

We are interested in describing and classifying the varied forms of expression used to cause harm as outlined above (see Jeshion, 2021, for a recent taxonomy of pejoratives). In working with the data in our corpus, we have identified the following 5 major categories of such expressions. The full taxonomy, including subcategories, is listed in Table 1.

Category I includes expressions of extreme or dangerous speech, assaultive speech, and language which glorifies or incites violence ("just burn in the sun", "DEATH TO CHINA").

Category II is used for pejorative expressions. These are words or phrases that are inherently insulting, derogative, etc. All of these forms are primarily evaluative in that they serve to express a speaker's feelings or attitudes towards the target. Subtypes include different types of slurs, which are meant to harm individuals simply because of their

group membership, pejoratives (e.g., “scum”, “idiots”), expletives (e.g., “damn”), swear words (e.g., “fuck”, “shit”), and others.

In contrast, category III is reserved for expressions that are being *used* derogatively, but are not inherently pejorative in their conventional meaning. This category includes jokes and inventive uses of language meant to put down other individuals (“DemoRATS”, “Commiefornia”), insulting metaphors (“they are a sickness”), as well as non-pejorative words when used pejoratively in context (e.g., “commie”, “Jews”).

Category IV is used for expressions which cause harm more implicitly by “othering” (Culpeper, 1996; Palmer et al., 2020) another group and covers general divisive language creating distinctions between an in- and an out-group (“The Chinese”, “Are women banned from this chat? If not, why the fuck not?”).

Finally, category V is used for coded expressions that can harm the discourse by communicating different messages to a like-minded in-group who is able to decode the expression than to others (“Trump train”, “GIVE THAT MAN A BRICK”). Notably, this category includes dog whistles which may seem innocuous to outsiders (“Patriot”).

Table 1: Taxonomy of online harm used in manual annotation.

I.	**incendiary speech** (assaultive speech, extreme speech, dangerous speech, the glorification of violence)
II.	**pejorative words and expressions** - dehumanising - 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 - rhetorical questions - insulting metaphors - inventive - non-pejorative words used pejoratively
IV.	**in/out-group** (divisive speech)
V.	**codes** - dog whistles

(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 Jr., 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, given that tweets 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 whom consist of only the link without any other textual content (333). Similarly to Twitter, Telegram allows user referencing with the help of the '@' sign (699), but these mentions for the most part appear 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[3]. However, @ is used more frequently than hashtags (only 209), which are an integral part of Twitter.

As we can observe in Table 3, the number of messages is not homogeneous throughout the 4.5 years. Although this data set only covers 18 days of 2021, this is the most active period, recording 15,603 messages. This sudden spike in new users on Telegram supports the idea of a migrating effect, following the temporary closure of Parler, the ban on Reddit's r/The_Donald, and the introduction of new Discord policies. We can observe this trend in Figure 1, where 2021 records the highest number of new users, and also the highest number of old users being 'revived' and maintaining their activity. It is difficult to assess the activity during 2016 and 2018, first because we only cover December of 2016 following the creation of the channel. The second reason is that

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

the channel saw massive message deletion in 2018, which is reflected in a small number of overall messages posted that year and also no new users being added to the channel in 2018 (Figure 1). However, we observe interesting trends in 2019 and 2020. In particular, 2019 is more active in the number of messages, but less so in the number of users added (see Figure 1). This can be explained by the fact that 2019 was full of events worth discussing, such as the US government shutdown and state of national emergency being introduced in order to secure sufficient funds for the Southern border construction. During 2020, we can see a gradual increase in activity up to 2021 (see Figure 2), which is mainly concerned with the pandemic, though this is not a general topic of interest for this group.

In 2020, we see the highest activity on December 23 in relation to Donald Trump issuing a flurry of pardons and commutations, and also 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). Interestingly, during this period there is first evidence of planning an assembly on January 6, 2021. The activity on the 27th of June 2019 is associated with the r/The_Donald subreddit being quarantined by Reddit admins due to excessive reports and threatening of the public figures in the context of the 2019 Oregon Senate Republican walkouts. The subreddit also lost revenue opportunities and was removed from feeds and search, leading to outrage from its users. The most active days in 2018, 2017 and 2016 seem to be reactions to other provocative tweets posted by the @realDonaldTrump account.

January 9, 2021, is the day with 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 starts on January 7, gradually grows to the 9th and then decreases to its usual average on the 16th.

Table 2. Annual message statistics

	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

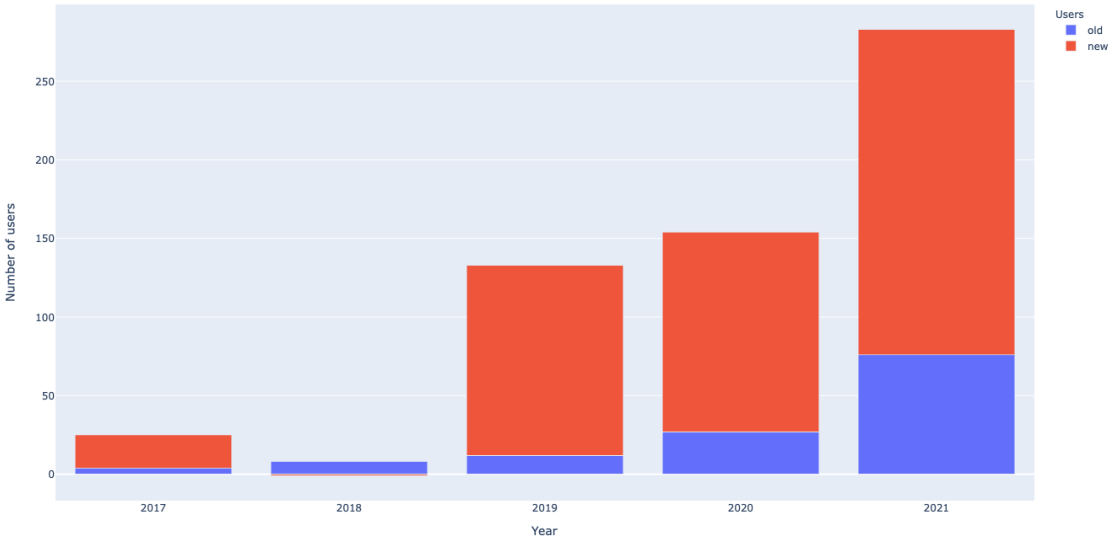

Figure 1. User statistics per year: newly added users and users active the previous year.

Figure 2. Messages per day in 2020.

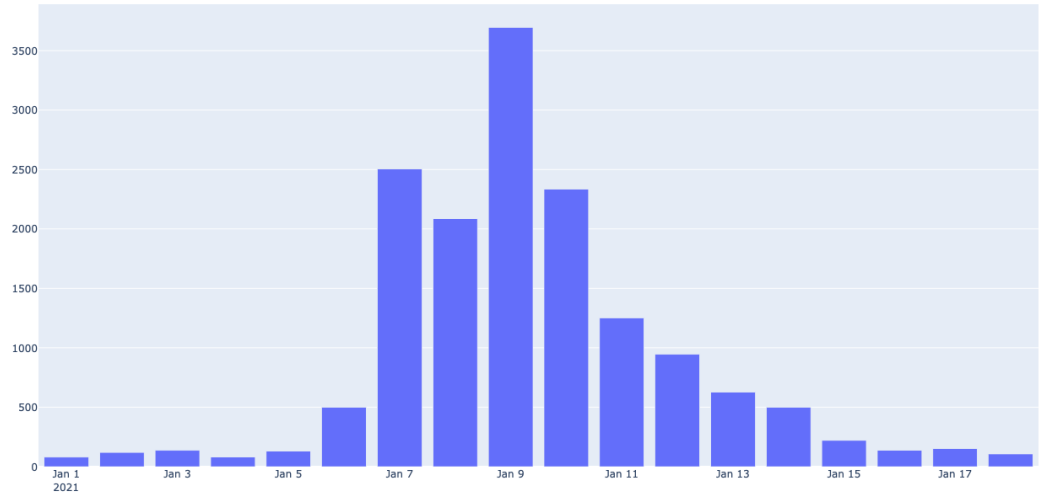

Figure 3. Messages per day in 2021.

(3.2) Offensive language over time

As described in (Solopova, Scheffler, & Popa-Wyatt, 2021), we automatically annotated the entire corpus by surface matching to the offensive language lists by Anger (2017) and Shutterstock (2020), and also by applying the open-source automated hate-speech/offensive language detection library HateSonar (Nakayama, 2017). We then measured the attributed tags quantitatively for each year. Interestingly, the messages labelled by the slur lists and by the automatic hate speech detection method are completely distinct; i.e., no message was tagged both using the offensive word dictionaries and by the machine learning based method. In parallel with the aforementioned trend on user activity in 2019, 2020, and especially 2021, we can see an increase in offensive language (Figure 4), although the overall fraction of these kinds of speech stays the same for all three years, at around 17% of messages. We can also see that we captured more messages using list matching than using the automated classifier. For the automated tool, the offensive tag is attributed almost 3 times more frequently than the hate speech category, because hate speech indicates a higher level of intensity of "offensiveness" (Davidson et al., 2017). We also have to note that the tool was trained on Twitter data, which, as we have seen above, differs in many ways from Telegram messages. These differences can influence its performance on our data negatively.

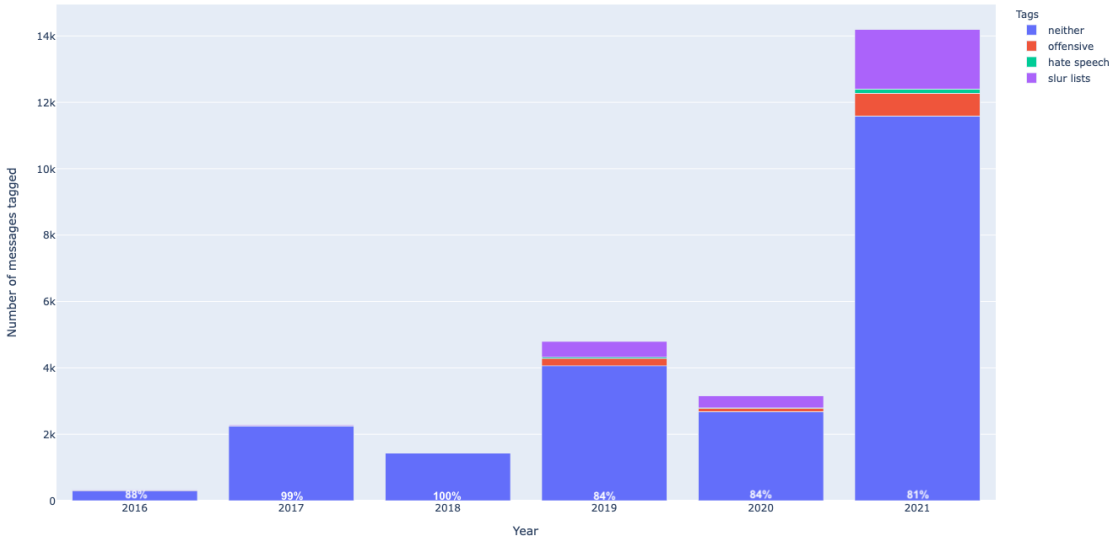

Figure 4. Automatically predicted offensive language labels over time. White numbers show the fraction of the 'neither' tag.

(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 the time from the US election up to and including the January 6 Capitol riot. We added part of January 9, to represent the most active day of our corpus. This resulted in 4,505 messages to annotate. We chose Brat (Stenetorp et al., 2012), which also permitted us to annotate token-level instances and not only on the sentence level.

In order to validate the annotation schema and determine the difficulty of the task, we chose a continuous thread of 711 messages from January 9, 2021, the most active day, to be reannotated by another linguist. This second annotator was provided with the taxonomy and examples for the individual categories. After annotation of 200 instances, we discussed the taxonomy and several difficult cases with both annotators, before the reannotation was completed. We computed inter-annotator agreement between the two independent annotations on a message level, first on the binary decision task (= "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 κ=0.70) on the binary harmful/neutral distinction. When evaluating the 6-way classification, 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 found both II and III, we counted the message as an agreement between annotators). The fine grained distinction leads to overall lower agreement (κ=0.65), reflecting the difficulty of the task. We find these scores promising for such a difficult and potentially subjective task and believe that more comprehensive annotation guidelines and in particular a dedicated adjudication process between annotators will be able to further raise the inter-annotator stability of our taxonomy categories. We leave this, as well as a fine-grained token-level agreement evaluation, for future work.

(4) Results and discussion

Overall, out of the chosen 4,505 messages we manually identified harmful language in 787 messages and 831 instances, meaning that 44 messages had more than one tag associated with it. 'II. Pejorative words and expressions' is the most represented category with 273 messages. The second-largest category is 'V. code words' (261). 'III. Insulting/offensive/abusive uses' and 'I. incendiary speech' are similar in size with 115 and 98 messages, respectively. The least represented class, 'IV. in/out-group' (40), is also the most complicated by definition and can often coincide with other classes. The distribution is shown in Table 3. As for the instance level, there are 310 (37 instances are found more than once in one message) instances annotated as 'pejorative words and expressions' and 121 (only six times in the same message) annotated as 'insulting/offensive/abusive uses', while statistics on the other categories coincide with those on the message level.

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

Tag	I. incendiary	II. pejorative	III. offen-sive uses	IV. in/out-group	V. codes	All
Number of messages	98	273	115	40	261	787
Fraction	12%	35%	15%	5%	33%	100%

Looking closely at the sub-categories in Figure 5, we can also observe that they are not evenly distributed. Let us start with pejoratives, as the biggest category, with its most frequent sub-categories being pejorative words used pejoratively (85 messages, 88 instances) and pejorative nicknames (90 messages, 97 instances), while swear words in

an offensive context account for 56 messages and 60 instances. There are also 27 canonical slurs, 12 generic pejoratives, 9 descriptive slurs, and only 6 gendered slurs and expressions, as well as stereotyping expressions and 2 expletives, while we found no examples of dehumanising speech in the manually annotated subcorpus. Within 'insulting/offensive/abusive uses' we found that the largest categories are non-pejorative words used pejoratively (39) and insulting metaphors (38). There are also 21 offensive jokes, 21 inventive-creative offensive instances, and 6 rhetorical questions.

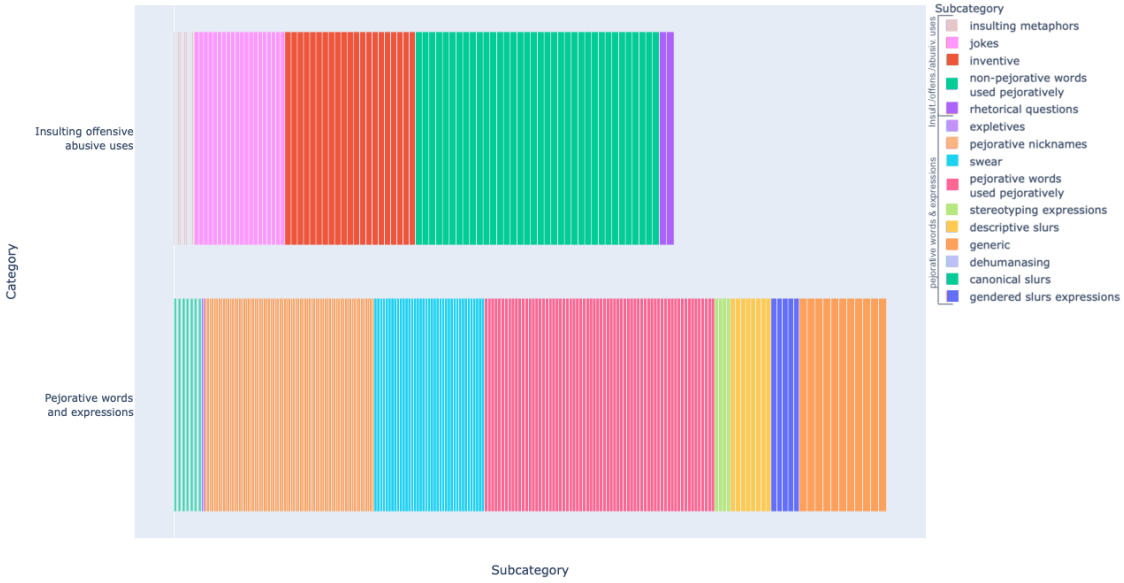

Figure 5. Sub-category statistics for pejorative words and expression and insulting/offensive/abusive uses categories.

(4.1) Comparison of manual and automated annotations

We performed a binary and multi-class comparison of the manual and automated annotations. The binary comparison (see Figure 6) shows how many potentially harmful messages (annotated manually) are also identified by either word list based or machine learning based automated methods. Out of the 4,505 messages in the doubly annotated subcorpus, 3,395 are tagged neither by the manual annotation nor by the automated one (true negatives), while only 275 were correctly found by the automated systems as either offensive or hate speech (true positives). Consequently, we have 835 messages annotated differently, where the HateSonar classifier or offensive word lists falsely attributed some messages as offensive/hate speech (432 false positives) or did not 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 joined

list+HateSonar performance as 40% (with 39% precision and 41% recall). These low scores confirm the low generalizability of solutions for offensive language detection when switching to a new dataset (Yin & Zubiaga, 2021), an effect exacerbated by the fact that our data originates from a different platform than the training data and contains many different types of more complex and implicit harmful language.

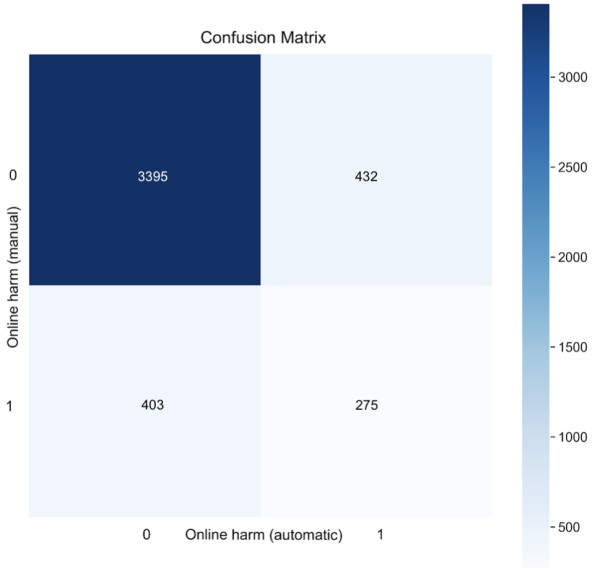

Figure 6. Binary confusion matrix for manual and automated annotation.

The results demonstrate a need for additional corpora of harmful language and in particular for more diverse and fine-grained annotations, as the multi-class confusion matrix shows (Table 4). Considering the main categories of harmful language, we note that the slur lists overall captured more harmful messages, and mostly overlap with category II, pejorative words and expressions. However, list-based detection also flagged many neutral messages as harmful (317), indicating low precision. At the same time, the machine learning-based system HateSonar has a low recall in our data while in particular its hate speech tag exhibits better precision. Overall we can see that the automated tools are generally better at finding neutral messages, an effect caused by the prevalence of this category in the data.

On the other hand, we can study our taxonomy using the automatically assigned tags as comparison. We see that recall is much higher across the automatic methods (again, the automatic tags do not overlap in any messages) for our categories II, III, and IV - pejorative expressions, offensive uses of not exclusively derogatory words, and in/out-

group divisive speech. Incendiary language calling to violence as well as, particularly, veiled codes are in contrast harder to detect by current automatic methods.

Table 4. Confusion matrix of multi-class annotation.

manual \ auto	slur 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	

(4.2) Exploratory analyses

Drawing on the hypothesis that harmful language poisons online discourse by further normalising hate speech, we expect that messages will be prone to contain more offensive language when they respond to already offending messages. 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=2208.64, Cramér's V=0.28). However, as noted above, the automated detection methods for offensive language have low accuracy, so this research shall be repeated after a revision of the automated tags in order to confirm the results.

We also explored the idea of Trump's narrative being reflected in his supporters' language. We analysed 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 only found a negative correlation (Spearman's ρ=-0.098) for overall activity, thus our expectation that more tweets would lead to more Telegram messages was not confirmed. At the same time, we think that this measurement

is inherently difficult as, on the one hand, the discussion can follow tweets with at least a day delay, as shown in the case of 27 June 2019, discussing the r/The_Donald ban of June 26, and on the activity of January 7-9, discussing the Capitol riot on the 6th. On the other hand, some days are lacking channel activity completely, so our data set does not always have the next day's data to measure this kind of delayed response.

We also started analysing 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 ("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 related to "funded" and "criminal", with interchangeable usage of all three and also in strong relation to 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, selected for its large potential for harmful language. The corpus consists of over 25,000 messages spanning a period of over 4 years and including discussion leading up to and in particular following the January 6, 2021, US Capitol riot. We argued for a broad notion of harmful language online, which includes not only direct attacks on persons and pejorative expressions, but also divisive language and statements meant to poison public discourse. To this end, we introduced a broad taxonomy of this type of online harm and provide manual annotations on a subset of our corpus. Comparing these annotations with automatically obtained labels of hate speech and offensive language, we showed that both lexicon-based methods and machine learning algorithms trained on other datasets and platforms are unable to detect the various subtle and implicit types of harmful language we encounter in our Telegram channel. We therefore believe, along with other authors, that much more research is needed in the philosophical foundations of online harm and their possible linguistic expressions. We attempt to contribute to this research by diversifying the available empirical foundation for these types of investigations in terms

of the platform, content, and the kinds of annotations we cover. In contrast to previous work, which often focussed on personal derogation, we specifically distinguish between pejoratives, which conventionally denote offensive attitudes, and offensive or derogatory *uses*, where non-pejorative expressions are used to attack or put down a person or group. These uses, as well as group-internal codes (which are not intended to be addressed to or understood by outsiders) and divisive language focussing in in-/out-group distinctions, can not easily be identified with lexical items and thus pose additional challenges for detection. Finally, the inclusion of language glorifying or inciting violence is important in our opinion, as we argue that such speech can lead to corresponding action, but cannot itself be easily identified by list based or state of the art machine learning based means. 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.

Acknowledgements

The authors would like to thank Lesley-Ann Kern for help with the annotations.

Funding statement

The research in this article was supported with funds from the research project grant "*HaLO* - How Language is Used to Oppress" (No 841443) from the Marie-Skłodowska Curie, ZAS Berlin. Further funding was supplied by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 317633480 – SFB 1287, Project A03, "Discourse Strategies across Social Media: Variability in Individuals, Groups, and Channels".

Competing interests

The authors have no competing interests to declare.

References

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

Benesch, S., Buerger, C., & Glavinic, T. (2018). Dangerous speech: a practical guide. http://dangerousspeech.org

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

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.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv:1607.04606*.

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 University Press.

Brown, A. (2017). What is hate speech? Part II: Family resemblances. *Law and Philosophy, 36*, 561–613.

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

Culpeper, J. (1996). Towards an anatomy of impoliteness. *Journal of Pragmatics, 25*(3), 349–367.

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*: 1-22.

Haraszti, M. (2012). Foreword: Hate speech and the coming death of the international

standard before it was born. In M. Herz & P. Molnar (Eds.), *The content and context of hate speech*. New York, NY: Cambridge University Press, xii–xviii.

Heinze, E. (2016). *Hate speech and democratic citizenship*. Oxford: Oxford University Press.

Howard, J. (2019). Dangerous Speech. *Philosophy & Public Affairs* 47: 208-254.

Jeshion, R. (2021). Varieties of pejoratives. In Justin Khoo and Rachel Sterkin (Eds.), *Routledge Handbook of Social and Political Philosophy of Language*.

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. 94–120). Oxford: Oxford University Press.

Maitra, I. (2012). Subordinating speech. In I. Maitra & M. K. McGowan (Eds.), *Speech and harm: Controversies over free speech* (pp. 94–120). Oxford: Oxford University Press.

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 University Press.

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.

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.

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

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: http://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; submitted). A Telegram corpus for hate speech, offensive language, and online harm. *JOHD*. (= accompanying data paper)

Stanley, J. (2015). *How propaganda works*. Princeton: Princeton University Press.

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.

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE, 15*(12), e0243300. 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.

Supplementary Files (optional)

Any supplementary/additional files that should link to the main publication must be listed, with a corresponding number, title and option description. Ideally the supplementary files are also cited in the main text.

Note: supplementary files will not be typeset so they must be provided in their final form. They will be assigned a DOI and linked to from the publication.