Journal of Open Humanities Data

A Telegram Corpus for Hate Speech, Offensive Language, and Online Harm

DATA PAPER

ubiquity press

VERONIKA SOLOPOVA
TATJANA SCHEFFLER
MIHAELA POPA-WYATT

*Author affiliations can be found in the back matter of this article

ABSTRACT

We provide a new text corpus from the social medium Telegram, which is rich in indirect forms of divisive speech. We scraped all messages from one channel of Donald Trump supporters, covering a large part of his presidency, from late 2016 until January 2021, including the January 6 Capitol riot. The discussion among the group members, over this long time period, includes the spread of disinformation, disparaging of out-group members, and other forms of harmful speech. To enable research into the role of harmful speech in political discourse, we added two types of annotations to the corpus: (i) automatic annotations of offensive language for all messages, and (ii) our own manual annotations of harmful language for a portion of the posts leading up to the January 2021 Capitol riot and its aftermath.

CORRESPONDING AUTHOR:
Veronika Solopova
Freie Universität Berlin, Germany
verosolopova@gmail.com

KEYWORDS:
social media; offensive language; harmful speech; linguistics; philosophy

TO CITE THIS ARTICLE:
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, pp. 1–5. DOI: *https://doi.org/10.5334/johd.32*

1 OVERVIEW

REPOSITORY LOCATION

https://osf.io/ck3gd/

CONTEXT

Scheffler, T., Solopova, V., & Popa-Wyatt, M. 2021. The Telegram chronicles of online harm. *Journal of Open Humanities Data*, *7*: 8, 1–15. DOI: https://doi.org/10.5334/johd.31

2 METHOD

STEPS

The data collection represents one public channel from the platform Telegram, encompassing four years of Donald Trump Jr.'s presidency, through the prism of his supporters' conversations, leading up to and including discussion of the January 6 Capitol riot. The data comprises 26,431 messages in a continuously evolving isolated "echo-chamber" discussion, produced by 521 distinct users. While many similar channels introduced the policy of daily chat history purge, this channel essentially preserved its integrity from the day it was created on December 11, 2016. It thus represents a unique testimony of a controversial period of American history, by providing a rich source of harmful speech and practices.

The content and metadata were mined using the Telethon[1] Python package. This is an interface to the Telegram API which facilitates interaction with Telegram and application development. Our data also contains the metadata, including date and time of post creation, message ID, user ID, the ID of the message replied to, any attached media (e.g., image, video, sticker), as well as the message text itself. This may be useful for further research modelling the interactions among participants in the community.

We provide multiple independent automatic and manual annotations of the corpus. We automatically annotated the corpus, firstly, using two lists of offensive language in English (Shutterstock 2020; Anger 2020) and, secondly, applying HateSonar (Nakayama 2020), an open-source automated hate speech detection library for Python based on (Davidson et al. 2017). In addition, after having statistically analyzed the channel activity, and after considering its crucial social context, we chose to manually annotate 4505 messages sent from November 1, 2020 to January 9, 2021, according to our own fine-grained taxonomy of harmful speech. We manually classified messages into five broad categories: incitement; pejorative words and expressions; insulting, offensive and abusive uses; in-out-group (divisive speech); and codes.[2]

QUALITY CONTROL AND LIMITATIONS

Out of the 4505 manually annotated posts, we doubly-annotated 711 messages and measured the inter-annotator agreement using Cohen's κ, in order to ensure annotation quality and identify the complexity of the task itself. Measuring the agreement on message-level assignment of 5 categories of harmful language (+the "none" category) revealed substantial agreement ($\kappa = 0.65$). Problematic instances have been discussed among the authors, and the discussion, in turn, helped refine the proposed taxonomy.

As a result of the controversial nature of the data, 3,619 additional messages originally posted in the channel appear to have been deleted prior to collection, leaving blank message content, which we filtered out. This also reduced the initial 1,068 unique users to 521. ***Figure 1*** illustrates this trend in 2018, as reflected in the small number of messages posted and the absence of new users.

[1] https://docs.telethon.dev.

[2] See (Scheer et al. 2021) and the annotation guidelines for examples and specifications.

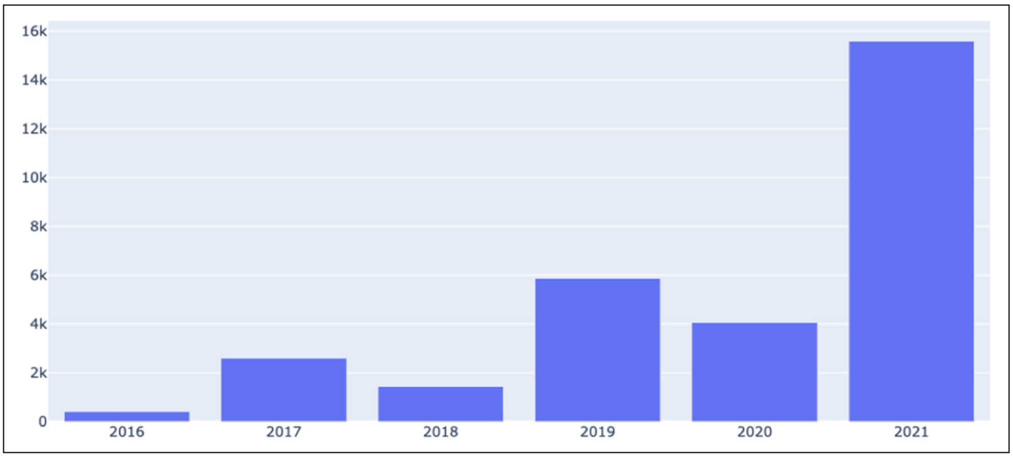

Figure 1 Number of messages in the channel per year.

3 DATASET DESCRIPTION

OBJECT NAME

1) telegram_corpus.tsv: annotated data in TSV format 2) telegram_corpus.json: annotated data in JSON format

FORMAT NAMES AND VERSIONS

TSV and JSON

CREATION DATES

2016-12-11 – 2021-01-18

DATASET CREATORS

Tatjana Scheffler, Veronika Solopova, Mihaela Popa-Wyatt

LANGUAGE TEXT

Mainly English; Chinese and Russian less than 0.1%.

METADATA

Date and time; message ID; posting user ID; reply-to-message ID; media attached; automated offensive speech tags; manual annotation.

LICENSE

CC-BY

REPOSITORY NAME

OSF

PUBLICATION DATE

2021-06-16

4 REUSE POTENTIAL

Telegram is a widely used social media platform that allows asynchronous, anonymous communication between individuals, within a range of broad thematic channels. Our corpus documents a complete channel in this platform, from its creation in December 2016 until January 2021, leading up to and including the January 6 Capitol riot. Telegram differs from other platforms with regards to its user base and content. Notably, it has not been the direct focus of computational linguistics studies to date. Thus, our data provides useful insights into the types of content contributions and users' interactions, which are a valuable source to compare with other media (e.g., Twitter, Reddit, Facebook) along a range of dimensions and within several scientific fields. Linguists will be interested in studying the way asynchronous dialogue is structured in our corpus.

The channel was chosen specifically to document online use of harmful language, among a like-minded group of users. This will allow follow-up studies to refine definitions and taxonomies of harmful speech and online harm in linguistics, philosophy, communication, and media studies.

In addition, the data can be used to validate computational methods for harmful language detection. Various computational methods have already been developed based on data from other media and domains, e.g., discourse around immigration on Facebook, gender on Twitter (see Poletto et al. 2020 for a comprehensive overview of available datasets). Our data suggests there is utility in evaluating these methods based on novel data, such as our corpus. This is because the in-group community that tends to populate Telegram channels has a different dynamic than that of more open and heterogenous communities present on Facebook, Twitter, or YouTube. The enclosed nature of the Telegram community is arguably a key factor to an environment amenable to incubating hatred, and where harmful language is the norm.

Reliable algorithms for detecting online harmful speech are so much needed in the context of our societies, in which hate speech, misinformation and right-wing extremism are on the rise. There is, thus, an immediate practical application of our data and analysis to research on digital language.

In addition to contributing to computational methods for detecting harmful speech, our corpus is also useful to validate general methods in natural language processing, such as coreference resolution and dialogue act tagging, which have been developed based on data from other media. The corpus can also be used as a resource for teaching in corpus and computational linguistics.

Notably, given the time period we chose, leading up to and following the January 2021 US Capitol riot, the corpus provides valuable data for political scientists, sociologists, and communication scientists, interested in the conditions and consequences of critical political events in the United States, how they shape public opinion by mobilizing and radicalizing layers of population into a dangerous right-wing base.

Finally, it is worth flagging that as a corpus assembled by mining a social media channel, the data has certain limitations. While this particular corpus may be of interest in the current political climate of US politics, it may lose its topical impact or become superseded by future political changes and trends in public discourse. Thus, though most of the available posts were created within the last few months, they may become obsolete for future studies relying on more recent data.

Another inherent difficulty with online content in general is that supplementary material, such as URL links included in the posts, may be deleted, and thus render the context for the content contributions insufficient or simply unavailable. This is unfortunately the fragile nature of online content and data, which is a common problem for this type of research.

A further difficulty concerns users' anonymity. The corpus contains anonymous posts, which make it difficult to ask permission for explicit consent from the users to be included in a scientific research study. Thus, the data should only be used in aggregate and ideally for automatic analyses.

Given these difficulties, our goal is to make a contribution both in terms of providing empirical data of an increasing corrosive force in our democracies in the form of online harm, and in terms of refining our computational tools to improve performance in detection of harmful speech.

Our data collection is created in accordance with the FAIR principles (Wilkinson et al. 2016) meaning that it is Findable, Accessible, Interoperable and Reusable, as it is publicly available through the OSF platform; it is open-source and presented in two widely used formats, TSV and JSON; in (Scheer et al. 2021) we analyze its content showing that it contains a big variety of information, inviting further interdisciplinary research.

ACKNOWLEDGEMENTS

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

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 partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1412, 416591334, and SFB 1287, 317633480.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Veronika Solopova: Data curation, formal analysis, investigation, methodology, resources, software, visualization, writing – original draft. *Tatjana Scheffler*: Conceptualization, data curation, formal analysis, methodology, supervision, validation, writing – original draft, writing – review & editing. *Mihaela Popa-Wyatt*: Conceptualization, methodology, writing – original draft, writing – review & editing, funding acquisition.

AUTHOR AFFILIATIONS

Veronika Solopova *orcid.org/0000-0001-7498-6202*
Freie Universität Berlin, Germany

Tatjana Scheffler
Ruhr-Universität Bochum, Germany

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

REFERENCES

Anger, Z. (2020). List of profanity in English. *https://github.com/zacanger/profane-words*

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*

Nakayama, H. (2020). HateSonar: Hate speech detection. *https://github.com/Hironsan/HateSonar*

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

Scheffler, T., Solopova, V., & **Popa-Wyatt, M.** (2021). The Telegram chronicles of online harm. *Journal of Open Humanities Data*, *7*: 8, 1–13. DOI: *https://doi.org/10.5334/johd.31*

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

Wilkinson, M., Dumontier, M., Aalbersberg, I., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(160018). See also: *https://www.go-fair.org/fair-principles/*. DOI: *https://doi.org/10.1038/sdata.2016.18*

TO CITE THIS ARTICLE:
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, pp. 1–5. DOI: *https://doi.org/10.5334/johd.32*

Published: 05 July 2021

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