A Telegram corpus for hate speech, offensive language, and online harm

Repository location:
https://osf.io/ck3gd/?view_only=e54bb3172c974c009bad8aee0332f923

Paper Authors

Solopova, Veronika; Freie Universität Berlin, Germany
Scheffler, Tatjana; Ruhr-Universität Bochum, Germany
Popa-Wyatt, Mihaela; Zentrum Allgemeine Sprachwissenschaft, Berlin, Germany

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 supporters of Donald Trump, covering a large part of his presidency from late 2016 until January 2021. 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 offensive speech. To encourage research into such practices of poisoning public political discourse, we added automatic annotations of offensive language to all messages. We further added manual annotations of harmful language to a portion of the posts in order to enable the analysis of more implicit forms of online harm.

Keywords

social media; offensive language; linguistics; philosophy

Context

Tatjana Scheffler, Veronika Solopova, and Mihaela Popa-Wyatt. 2021. The Telegram chronicles of online harm. JOHD.

(2) Methods

Steps

The data collection represents one public channel from the platform Telegram, encompassing 4 years of Donald Trump Jr.'s presidency through the prism of his supporters' conversations. 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 11th, 2016, and thus represents not only a unique witness to this highly controversial time of American history but also a very particular source of harmful speech and offensive language.

The content and metadata were mined with the help of the Telethon Python package, which is an interface to the Telegram API, facilitating interaction with Telegram and application development. We included the metadata we found useful for research purposes, namely date and time of post creation, message-id, user id, the id of the message replied to, if applicable, presence of media attached (e.g., image, video, sticker), and the message text itself.

We then automatically annotated the corpus, firstly, using two lists of offensive language in English [1,2] and, secondly, applying HateSonar [3], an open-source automated hate speech detection library for Python based on [4].

Finally, after having statistically analysed the channel activity, as well as considering its crucial social context, we chose the period from November 1, 2020 to January 9, 2021 to manually annotate 4505 messages according to our fine-grained taxonomy of offensive and harmful speech. We attributed messages to 5 categories: incitement; pejorative words and expressions; insulting, offensive and abusive uses; in-out-group (divisive speech); and code words.

Quality Control and Limitations

As a result of the controversial essence of the data, 3619 additional messages in the channel appear to have been deleted, leaving blank message content, which we filtered out. This also reduced the initial 1068 unique users to 521. According to our data, 2018 was a year especially influenced by this trend, which can be observed in Figure 1, reflected by a small number of messages posted during 2018, and also by the fact that no new users were added to the channel that year.

As for the manually annotated posts, we used Cohen's κ to measure the inter-annotator agreement of doubly-annotated 711 messages 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 (κ=0.65). Problematic instances were discussed by the authors, in order to refine the taxonomy.

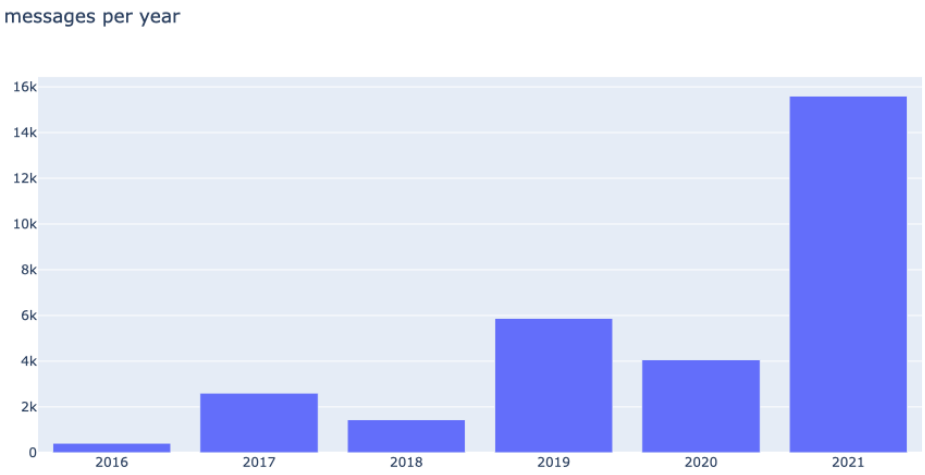

Figure 1. Number of messages in the channel per year.

(3) Dataset description

Object name
telegram_corpus.tsv : annotated data in TSV format
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
n/a

Language
Text: English; Chinese and Russian less than 0.1%
Metadata: date and time; message-id; posting user-id; reply to message-id; media attached; automated hate speech tag; manual annotation.

License
CC-BY

Repository name
OSF

Publication date

Data will be made public upon acceptance.

(4) Reuse potential

Telegram is a widely used social media platform that allows asynchronous, anonymous communication between individuals within a range of broadly thematic channels. Our corpus documents a complete channel in this platform from its creation in December 2016 until January 2021. Telegram differs from other platforms in its user base and content, but has so far not been the direct focus of a large number of studies yet. This data provides a snapshot of the type of contributions and interactions available on this platform that will enable future comparison to 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 dialog is structured in our corpus. The channel was chosen specifically to document offensive or harmful language, among a like-minded group of users. This will allow follow-up studies on the definition and analysis of hate speech and online harm in linguistics, philosophy, communication, and media studies. In addition, the data can be used to validate computational approaches to the detection of offensive language. Such methods have been developed based on data from other media and domains, but must be evaluated based on novel and more indirect data as included in this corpus. Reliable algorithms for detecting hate speech online are also a highly sought after practical application of research on digital language. In addition to hate speech detection in particular, the corpus can also be used to validate general methods in natural language processing such as coreference resolution and dialog act tagging, which have been developed based on data from other media.
The corpus also includes the time period leading up to and following the January 2021 U.S. Capitol riot. It provides valuable data for political scientists, sociologists, and communication scientists interested in the organization of and fall-out from these events in a public forum aligned with the political right in the United States.
Finally, the corpus can be used as a resource for teaching in corpus and computational linguistics.
As a corpus assembled by crawling a social media channel, the data also has some limitations. In particular, while the corpus is current now and most of the available posts were created within the last few months, in a few years it may be out of date for studies relying on recent data. Links included in the posts as well as missing contributions (because posts have been deleted before crawling) may make some of the context unavailable. In addition, the corpus includes only anonymous posts which make it impossible to get explicit consent from the authors to be included in a scientific research study. Thus, the data should only be used in aggregate and ideally for automatic analyses.

Our data collection is created in accordance with the FAIR principles [5], meaning that it is Findable, Accessible, Interoperable and Reusable, as it is

publicly available through OSF platform; it is open-source and presented in two widely used formats TSV and JSON; in [6] we analyze its content showing that it contains a big variety of information, inviting further research relevant for many different disciplines.

Acknowledgements
The authors would like to thank Lesley-Ann Kern for additional 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".

References

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

[2] Zac Anger. (2020). List of profanity in English. https://github.com/zacanger/profane-words

[3] Hiroki Nakayama. (2020). HateSonar: Hate speech detection. https://github.com/Hironsan/HateSonar

[4] 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

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