Examining the Prevalence and Impact of Toxicity and Hate Speech in Online Comments

Abstract. Online platforms have become a prevalent part of modern life, providing a space for communication and connection. However, these platforms have also been used to spread hate speech and engage in toxic behaviour, leading to negative consequences for individuals and society. In this abstract, we propose investigating the prevalence and impact of hate speech and toxicity on online platforms. This will involve a review of existing literature on the topic and collecting and analysing data from online platforms via the database provided by Civil Comments to understand the predictability of this behaviour. This research aims to shed light on the issue of hate speech and toxicity and identify strategies for predicting and combating these behaviours to create a safer and more inclusive online environment for all users. Find the analysis here.
Keywords: NLP; Sentiment Analysis; Violence; Civil Comments; Hate Speech

Introduction

Civil Comments was founded in 2015 to facilitate civility in online discussions with a commenting plugin for independent news sites. The idea was to use a peer-review system to imitate face-to-face interactions, asking users to score the civility of three random comments before their own could be published for review. Knowing that others will also rate their comments motivates them to moderate their posts before submitting them (Bogdanoff, 2018). This dataset has seven leading labels that crowd workers created, their values ranging from 0 to 1, indicating the ratio of users choosing the given label. The platform shut down in 2017 due to insufficient funds; however, they made the comments available for future research. Therefore, the comments were written between 2015 and 2017 and appeared on about fifty English-language news sites. Jigsaw extended the original dataset by annotating additional labels for identity mentions, toxicity and covert offensiveness (*Civil_Comments | TensorFlow Datasets*, n.d.).

Problems and Background

Social media platforms have allowed billions of people to connect online and share their opinions. However, while it is a great help, it also has negative consequences, like online harassment and cyberbullying. *Hate speech* is a specific offensive language that utilizes stereotypes and minorities to express hate (Warner & Hirschberg, 2012).

Twitter defined *hate speech* as the following: 'any tweet that promotes violence against other people based on race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease' (*Twitter's Policy on Hateful Conduct | Twitter Help*, 2022). Numerous governmental and social media sites are trying to constrain these kinds of posts, but it is still an immense problem in our society (Mathew et al., 2019). People are more likely to engage in aggressive online behaviour due to the anonymity given by this digital environment (Burnap & Williams, 2015). Furthermore, hate speech tends to have specific targets from specific religions, sexuality or gender (Fortuna & Nunes, 2018).

Due to this pressing issue, research on security in social media has also rapidly increased in the past decade. Numerous datasets are available for this research; however, most research on abusive language has used binary classification with one positive and negative label. A study by Dinakar et al. (2011) suggests that those training systems have relied way too much on the frequency of offensive words.

Furthermore, the automatic detection methods are vague, inefficient, and lack training data (Fortuna & Nunes, 2018).
Based on the above reasons, this study will try complementing the current detection systems, targeted explicitly at specific genders. We hypothesize that hate speech often occurs with well-known minorities as target groups and has a specific vocabulary. Furthermore, we also suggest that their sentiment value is vastly negative.

Methods

Software Framework

This analysis was conducted on MacBook Pro (13-inch, 2017, Two Thunderbolt 3 ports) with 8GB RAM, which runs the macOS Big Sur operating system. The data was processed in the Jupyter Notebook programming environment (Kluyver et al., 2016) using Python 3 (Van Rossum & Drake, 2009).

Data Acquisition and Processing

The dataset was downloaded from an online community platform specifically made for data scientists, called Kaggle (*Jigsaw Unintended Bias in Toxicity Classification | Kaggle*, n.d.). Due to the great size of the dataset, it is not available in my repository but can be accessed via this link, the sheet called *all_data.csv*. The corpus consisted of 46 columns and about two million entries. First, 1000 entries were randomly sampled to facilitate the data processing. Then the dataset was tokenised by removing all special characters and stopwords and stemming the remaining corpus. Stemming is a process of reducing the target word to its root format. This is necessary for clustering and data classification later on. All the steps were carried out using the stem and tokenise modules from the nltk package. Having the processed comments, word clouds were created to visualise the most frequent words used in the comments labelled with the 'threat' and 'toxicity' categories using the wordcloud package. This might illustrate some overlap between the two topics. Afterwards, a violin correlation plot was applied to investigate the relationship between the gender of the target person and the toxicity of the given comment. Since not many entries were labelled as toxic, another random sampling was made from the whole dataset, choosing 50.000 data points. To visualise the correlation, a new data frame was made with a column containing the gender, the value of the gender and the toxicity. The value and the toxicity ranged between 0 and 1, indicating the fraction of users who categorised the comment with the given label. Having the new data frame, the identity and the toxicity were plotted. To investigate this relationship further, Pearson correlation coefficients were calculated between identity value and toxicity strength, categorised by gender. Lastly, sentiment analysis was conducted to investigate the relationship between the sentiment of the comment and whether it got approved or rejected. First, the sentiment of each comment was defined with the sentiment modul of nltk. This results in four different scores: a score of 'negative', 'neutral', 'positive' and 'compound', all of them ranging between 0 and 1. While the first three scores add up to 1.0 altogether, the compound is normalised. However, for this research, only the compound factor will be used. In our random sample, there were a total of 919 approved comments and 81 rejected comments. Boxplots were made based on their compound values after sorting them into two different data frames based on their acceptance. To assess the strength of this relationship, Students' t-tests were conducted as well.

Empirical Results

Figure 1 and 2. The most frequent keywords in comments labelled with threat (left) and toxicity (right).

As we can see, the majority of the words are everyday use words, therefore it barely indicates specific vocabulary.

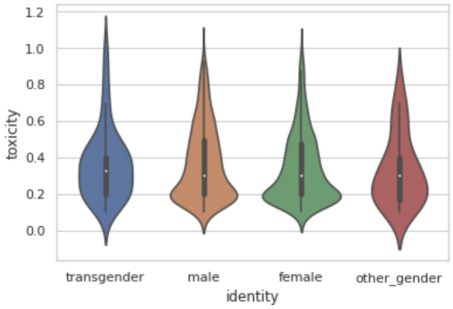

Figure 3. Violin plots of the correlation between toxicity and gender.

All of the categories seem to be the strongest around the toxicity value of 0.2-0.3, none of the Pearson correlation coefficients were significant and had no strong relationship (see **Table 2**).

	Pearsons Correlation	**P Value**
Transgender and Toxicity	0.05	.743
Male and Toxicity	-0.06	.152
Female and Toxicity	-0.07	.119
Other genders and Toxicity	0.26	.211

Table 1. Pearsons correlation between the identity value and toxicity value.

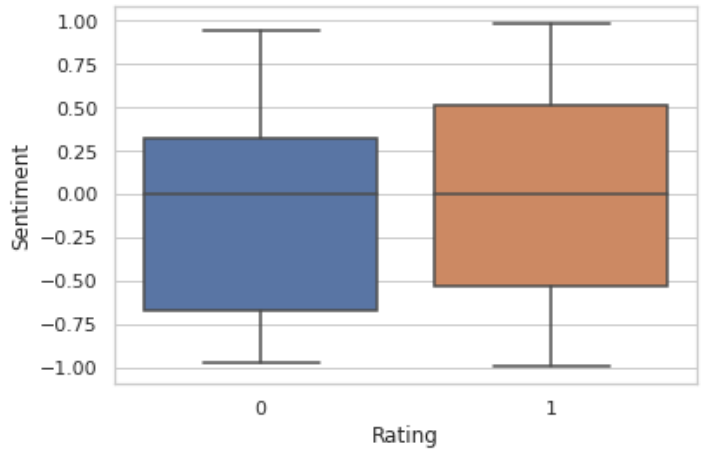

Figure 4. The sentiment value based on the rating of the comment; 0 being 'Rejected' and 1 being 'Accepted'.

The results from the t-test indicated that there is a significant relationship between the approval of the post and its sentiment, the accepted posts being more positive, $t(998) = 45.9$, $p < .001$. This suggests that the sorting method of the comments might be predictable based on the negativity of the wording.

Critical Evaluation

Looking at he word clouds, they do not suggest a specific vocabulary for either label, except for *'kill'*, *'trump'* and *'tax'*. However, it is interesting that both categories have *'people'* and *'time'*, since both of them are neutral words. In future research, bigrams could be created to assess in what kind of context these two words are used. Furthermore, a more thorough elimination could be done by removing all the neutral and frequent words in everyday language, only leaving less used words. After removing frequently used words, a corpus could be made from strictly words with a negative compound sentiment value. This would ensure that only potentially threat-related words would appear. That is, specific vocabulary might be better represented like this.

The similarity between the genders and toxicity might be explained by the overall mean of toxicity being 0.35. This either suggests that the raters were not as strict with the labelling or that the given comments were truly not that toxic. However, the label rating does not seem to be diverse enough to produce significant differences.

However, there was an interestingly strong connection between the sentiment value of the wording and its rating of it, indicating that the more positive a comment is, the more likely the post is to be approved by the raters. This is consistent with previous research.

Conclusion

Based on the results, this dataset could have been more optimal for correlation analysis. This might be because not every comment had a label with 'toxicity' (1149 entries out of 2.000.000); therefore, a broader investigation involving all labels might result in stronger relationships. Furthermore, a general label could be generated from all negative labels. That is, no matter whether the comment was 'toxic' or a 'threat', it would be added to the corpus. This would provide a greater pool for the analysation. Furthermore, the majority of the labels and rating aspects have not been used yet, therefore an extended research could also investigate that. For instance, analysing the dynamics behind comments aiming religions or sexualities. These three categories then could be compared by visualising which one is most exposed to hate speech. Lastly, further research could also involve machine learning to predict offensive

comments and whether they got accepted or rejected. The Jigsaw dataset on Kaggle provides a training and a testing dataset for this exact purpose.

Required Metadata

Nr	Software metadata description	
S1	Current software version	Python 3
S2	Permanent link to Github repository where you put your script or R project	https://github.com/Digital-Methods-HASS/au668705_Juli_Furjes
S3	Legal Software License	3-Clause BSD License
S4	Computing platform / Operating System	macOS BigSur (version 11.6.2 (20G314))
S5	Installation requirements & dependencies for software not used in class	You need to have Python 3 and the dataset downloaded
S6	If available Link to software documentation for special software	
S6	Support email for questions	juli.furjes@au.dk

Table 1. Software metadata.

Nr	Metadata description	
D1	all_data.csv	A collection of comments rated on the scales of oxicity, obscenity, threat, insult, identity attack and sexual explicity, besides identifying the target person's sexuality, religion and race. The raters were civil people over the years and the database was made available by Civil Comments in 2017.

Table 2. Data metadata (use the template below or create your own metadata table).

References

Bogdanoff, A. (2018, May 17). *Saying goodbye to Civil Comments - Aja Bogdanoff*. Medium. https://medium.com/@aja_15265/saying-goodbye-to-civil-comments-41859d3a2b1d

Burnap, P., & Williams, M. L. (2015). Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy &Amp; Internet*, *7*(2), 223–242. https://doi.org/10.1002/poi3.85

civil_comments | TensorFlow Datasets. (n.d.). TensorFlow. https://www.tensorflow.org/datasets/catalog/civil_comments

Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the Detection of Textual Cyberbullying. *International Conference on Weblogs and Social Media*, *5*(3), 11–17. https://ie.technion.ac.il/~roiri/papers/3841-16937-1-PB.pdf

Fortuna, P., & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, *51*(4), 1–30. https://doi.org/10.1145/3232676

Jigsaw Unintended Bias in Toxicity Classification | Kaggle. (n.d.). https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/data

Kluyver, T., Ragan-Kelley, B., Perez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. *International Conference on Electronic Publishing*, 87–90. https://doi.org/10.3233/978-1-61499-649-1-87

Malmasi, S., & Zampieri, M. (2017). Detecting Hate Speech in Social Media. *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*. https://doi.org/10.26615/978-954-452-049-6_062

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of Hate Speech in Online Social Media. *Proceedings of the 10th ACM Conference on Web Science*. https://doi.org/10.1145/3292522.3326034

Twitter's policy on hateful conduct | Twitter Help. (2022, February 10). https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. *CreateSpace EBooks*. https://dl.acm.org/citation.cfm?id=1593511

Warner, W., & Hirschberg, J. (2012). Detecting Hate Speech on the World Wide Web. *Proceedings of the Second Workshop on Language in Social Media*, 19–26.