Identifying Hate Speech Categories On Social Media

Jack Ryan Cracknell
School of Electronic Engineering and Computer Science
Queen Mary University of London
London
jcracknell123@gmail.com / EC20046@qmul.ac.uk

***Abstract*—With the ever increasing popularity of social media within our society, there exists a growing amount of people using such platforms to cause harm. Through natural language processing models, it is possible to detect when a social media post contains hate speech. These models can be used to autonomously moderate sites and remove hateful or offensive messages. One issue with this approach is that it may be difficult for data to be collected by such models that show their effectiveness at removing various subgroups of hate speech. In this paper, a model will be created which aims to identify the specific type of hate speech that a message contains. Through this model, social media sites could collect data on which hate speech is most rampant, and make adjustments to their post moderation algorithms accordingly.**

***Index Terms*—Social media, Hate speech**

I. Introduction

Hate speech has been defined as 'any communication that disparages a person or a group on the basis of some characteristics such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics' [1].

With the ever growing population of social media sites, there has been an explosion of posts that have been uploaded to them. Due to this, there has also been an increase in the amount of hateful activity online. Some of this hateful activity has been broadcast in real time on social media sites, such as the Christchurch shooting in 2019 [6]. Successfully identifying these hate speech messages through machine learning models aids in removing not only the messages themselves but also similar posts, stemming the amount of hate that can be spread. Having a state-of-the-art hate speech classification model may also increase the speed and accuracy at which social media sites can identify possible perpetrators of hateful actions in the real world, since many individuals who commit violent hate crimes have previously posted hate speech online [7] Furthermore, having the ability to classify hate speech posts into the most relevant subclass can give insight into what type of hate is most prevalent.

Before automatic natural language processing (NLP) models were introduced, websites would aim to block hate speech by having filters around specific words and phrases, but this is extremely time consuming. Word filters also scale poorly due to the quick evolution of language [2], [3]. Therefore an effective NLP model must be adaptable to new language and efficient in parsing through huge datasets.

There are various NLP methods for tackling hate speech online. Some classical approaches include Logistic regression, Support Vector Machines and Deep neural networks. In recent years, neural network models have become the industry standard state-of-the-art approach [4], although in some cases, the results given by neural networks can be difficult to interpret [5].

social media (using twitter data for this proj), Mention collecting multiple data sources together for a large range of different hate speech types.

Talk about some issues past papers (bias, unbalanced classes, different perceptions of hate speech) have come across and how they will be answered by my research

II. Background and related work

Related papers on hate speech detection, word filters, nlp models, find a paper on 'cyberbullying' Background topics specific NLP models and their background / papers they were developed on.

A. Classification models

(Logistic regression, SVM, Deep neural nets)

B. Feature space reduction

(ngram, sentiment analysis, pos tags)

C. subsection 1

Subtopic 1

III. Methodology

Methodology used in project / maybe data cleaning here?

A. subsection 2

Methods subsection

IV. Experiments and Results

Experiments and results here

V. Conclusion and Future work

Final summation

Acknowledgement

Any relevant acknowledgements

REFERENCES

[1] John T. Nockleby. 2000. Hate Speech. In Leonard W. Levy, Kenneth L. Karst, and Dennis J. Mahoney, editors, *Encyclopedia of the American Constitution*, pages 1277–1279. Macmillan, 2nd edition.

[2] Schmidt, A. and Wiegand, M., 2017, April. A survey on hate speech detection using natural language processing. *In Proceedings of the fifth international workshop on natural language processing for social media* (pp. 1-10).

[3] Badjatiya, P., Gupta, S., Gupta, M. and Varma, V., 2017, April. Deep learning for hate speech detection in tweets. *In Proceedings of the 26th international conference on World Wide Web companion* (pp. 759-760).

[4] Zimmerman, S., Kruschwitz, U. and Fox, C., 2018, May. Improving hate speech detection with deep learning ensembles. *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018).

[5] MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N. and Frieder, O., 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8), p.e0221152.

[6] Regan, A., Gunter, J. (2019). 'Reaction to NZ mosque attacks', *BBC*, 15 March. Available at: https://www.bbc.co.uk/news/live/world-asia-47578860 (Accessed: 26 March 2021).

[7] Hatzipanagos, R. (2018). 'How online hate turns into reallife violence', *Washington Post*, 30 November.
Available at: https://www.washingtonpost.com/nation/2018/11/30/how-online-hate-speech-is-fueling-real-life-violence/ (Accessed: 26 March 2021).