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. 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. 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. (Logistic regression, SVM, Deep neural nets) also (ngram, sentiment analysis, pos tags) ========== TALK ABOUT SOME DIFFERENT APPROACHES HERE ==========

social media (using twitter data for this proj), specific NLP models and their background / papers they were developed on. Mention collecting multiple data sources together for a large range of different hate speech types. try to get 4 or 5 references in here, multiple references from the same paper is fine too.
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

A. subsection 1

Subtopic 1

III. Methodology

Methodology used in project

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).