Project #15: Twitter hate speech detection 2

Janne Eskola Teemu Ikävalko Toni Kuosmanen Tapio Kursula

***Abstract*—Abstract**

***Index Terms*—natural language processing, sentiment analysis, hate speech**

I. Introduction

The Cambridge dictionary defines hate speech as:

"public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation" [1]

In practice, defining clear boundaries between hate speech and normal expression can be hard. Legal definitions of hate speech vary by country with some countries setting much stricter definitions of hate speech. [2]

It could be argued that today, the largest platforms for public speech are provided by the social media companies, such as Facebook or Twitter. Both these companies forbid hate speech on their platforms [3], [4] but the sheer volume of posted content makes it hard to enforce these rules. The services rely on users reporting hateful content when it is found on the platform. Facebook has also utilized machine learning algorithms in detecting hate speech but the results have not been optimal. [5]

Hate speech can be spread by individuals or extremist groups seeking to advance their agendas. Hate speech is usually directed at minorities with the aim of demonizing and dehumanizing the targeted group. The effects of hate speech can be severe. For example, the UN fact finding mission sent to Myanmar, after the government crackdown on the Rohingya minority, found that hate speech spread on Facebook contributed significantly to sparking tensions in the region [6].

The goal of our project is to find efficient methods for identifying hate speech on twitter. Our aim is to find a set of features that could be used to identify hate speech content.

II. Problem description

III. Data sets

A. Data set 1: Labeled hate speech

Tada

B. Data set 2: Active hate speakers

Todo

IV. Characterization of the labeled data set

A. Sentiment analysis

TODO

B. LIWC features

To identify common themes and topics in the labeled tweet data set, we used Empath, which is an open source alternative to proprietary LIWC software. [7]

C. Emoticon usage

TODO

D. Named entities

TODO

E. Named phrases

TODO

V. Radicalization of of active hate speakers

TODO

VI. Results

TODO

VII. Conclusions

References

[1] (2020) Cambridge dictionary. [Online]. Available: https://dictionary.cambridge.org/us/dictionary/english/hate-speech

[2] (2020) Hate speech. Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Hate_speech

[3] (2020) Hateful conduct policy. Twitter. [Online]. Available: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

[4] (2020) Facebook community standards. Facebook. [Online]. Available: https://www.facebook.com/communitystandards/hate_speech

[5] B. Perrigo. (2019, nov) Facebook says it's removing more hate speech than ever before. but there's a catch. [Online]. Available: https://time.com/5739688/facebook-hate-speech-languages/

[6] (2018) U.n. investigators cite facebook role in myanmar crisis. Reuters. [Online]. Available: https://www.reuters.com/article/us-myanmar-rohingya-facebook/u-n-investigators-cite-facebook-role-in-myanmar-crisis-idUSKCN1GO2PN

[7] E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 4647–4657.