University *of Ljubljana*
Faculty *of Computer and Information Science*

Offensive language exploratory analysis

Matic Fučka, Anže Alič and Aljaž M. Eržen

Abstract

In this assignment we analysed the typical structure of hate speech. We used methods ranging from traditional to neural and even some statistical ones.

Keywords
Hate speech, NLP

Advisors: Slavko Žitnik

Introduction

In this assignment we will try to analyse the structure of typical offensive language. We will tackle this problem with a standard statistical analysis and extend it with a use of traditional natural language processing methods. We will also improve on those with the use of neural approaches. In the end we will also check if we are able to correctly identify offensive language cross domain (different datasets).

Existing solutions

Several papers have already addressed hate post analysis and identification. Albadi *et al.* [1] have used different statistical methods, such as χ^2 test. to determine the terms with statistically most significant association to one of the hate speech classes. Davidson *et al.* [2] have proposed using n-grams weighted by its TF-IDF and classifying it using logistic regression. Qian *et al.* [3] have classified hate speech using several different non-contextual neural approaches. Xiang *et al.* [4] have used BERT model to predict the toxicity of a post and detect the part of the post from which the toxicity stems.

For our exploration we will use several different datasets. The first one was proposed by Mandl *et al.* [5]. It consists of 7005 English posts from Twitter and Facebook. For each post we have in our dataset its id, its text and three labels. First label categorizes whether the post is hate speech or not. The second label categorizes hate speech posts into subcategories(hate, offensive and profanity). And the last label tells us against whom was the hate speech meant(targeted and untargeted). The dataset is already split into a training and test set. The training set contains 5852 posts, from which 2261 posts are hate speech. The test set contains 1153 posts, from which 288 posts are hate speech.

Another similar dataset was proposed by Founta *et al.* [6], who collected the data from the dataset. The dataset consists of tweet ids and a label for each tweet. There are 7 different labels: offensive, abusivee, hateful speech, aggressive, cyberbullying, spam and normal. The impressive thing about this dataset is that it consists of 80 000 tweets.

A broader approach was taken by Ousidhoum *et al.* [7] where they created a dataset of 5647 tweets with labels for 5 different aspects: directness (direct/indirect), hostility (abusive, hateful, offensive, disrespectful, fearful and normal), the target (origin, gender, sexual orientation, religion, disability, other), target group(individual, other, women, special needs, African descent) and response from an annotator (disgust, shock, anger, sadness, fear, confusion, indifference).

Rezvan *et al.* [8] created annotated corpus usable for research. Dataset has 75000 tweets labelled with different labels such as sexual, racial, appearance-related, intellectual, political and labels for any other types of harassment. Each sample in the dataset was labelled by three different researchers.

Waseem [9] created a dataset where he classifies tweets into four classes sexist, racist, neither, both. Dataset was annotated by feminist and antiracism activists. It contains 4033 tweets.

Initial ideas

We will tackle the problem similarly as we had described them in existing solutions. First we will check different traditional statistical and natural language processing methods, such as χ^2 test, n-grams, to determine the typical structure of hate speech in our datasets, particullary we will check if there are some non-slur words that affect the post heavily to be classified under a specific label. For example we predict that we will find that significant part of hate speech posts will

contain political terms, so such tests might have a bias towards it.

Later we will try to gain deeper understanding of it with the use of neural approaches, such as Word2Vec to find some potential synonims in posts with the same label. In the end we will use Bert to analyse the contextial structure even further and try to find out which specific part of the post indicates most heavily towards a specific label.

If time will permit us we will also construct different classification models, underlying on our analysis and check their cross dataset performance. For example if both datasets will contain racist posts we will check if a classifier trained on one dataset will correctly predict the post on the other dataset. With these we might even find some bias that was put subconsciously when creating the dataset.

References

[1] Nuha Albadi, Maram Kurdi, and Shivakant Mishra. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In Ulrik Brandes, Chandan Reddy, and Andrea Tagarelli, editors, *IEEE/ACM 2018 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018, Barcelona, Spain, August 28-31, 2018*, pages 69–76. IEEE Computer Society, 2018.

[2] Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009, 2017.

[3] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth M. Belding, and William Yang Wang. A benchmark dataset for learning to intervene in online hate speech. *CoRR*, abs/1909.04251, 2019.

[4] Tong Xiang, Sean MacAvaney, Eugene Yang, and Nazli Goharian. Toxccin: Toxic content classification with interpretability, 2021.

[5] Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA, 2019. Association for Computing Machinery.

[6] Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. *CoRR*, abs/1802.00393, 2018.

[7] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. Multilingual and multi-aspect hate speech analysis. *CoRR*, abs/1908.11049, 2019.

[8] Mohammadreza Rezvan, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie L. Shalin, and Amit Sheth. A quality type-aware annotated corpus and lexicon for harassment research. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, page 33–36, New York, NY, USA, 2018. Association for Computing Machinery.

[9] Zeerak Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *NLP+CSS@EMNLP*, 2016.