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

Offensive language exploratory analysis

Maša Kljun, Matija Teršek

Abstract

asdasdasd

Keywords
Keyword1, Keyword2, Keyword3 ...

Advisors: Slavko Žitnik

Introduction

In the last few years social media grew exponentially and with it also the ability of people to express themselves online. By enabling people to write on different online platforms without even identifying themselves it lead to a new era of freedom of speech. As this new medium for communication and writing brought many positive things, it also has its downside. Social media has become a place where heated discussions happen and often result in insults and hatred. It is an important task to recognize hate speech and to prevent it.

Hate speech is defined as *abusive or threatening speech or writing that expresses prejudice against a particular group, especially on the basis of race, religion, or sexual orientation.*[1]. However, we can see that the definiton is very vague. Having said that, the goal of our paper is to help distinguish different types of hate speech and find the specific keywords of its subgroups in order to explain its structure. This could help with its identification and classification.

There has been done a lot of research regarding the hate speech, however these works are usually focused on the classification of hate speech. One of the first works include [2] who built the decision tree based classifier Smokey for abusive message recogniton and classification. Some other works that focus mainly on classification include [3] who compare the classification accuracy of models trained on expert and amateur annotations, [4] who use convolutional neural networks for classification into four predefined categories, and [5] who use different natural language processing techniques for expanding datasets with emotional information for better classification. In the last years, especially deep learning models are often used for detection and classification of hate speech, such as [6] who propose a sophisticated method that is a combination of a deep neural network architecture with transfer learning. There is a also a lot of related work that focuses on creating large datasets such as [7] who create a large-scale, multilingual, expert based dataset of hate speech.

What is less common in the research area of hate speech is analysis of relationships between different types of hate speech and the importance of specific keywords.

That is why we we use smth to do smth etc. torej neki o teh modelih ki jih boma uporabljala. Pa neki o tem kere classe npr mama oz keri obstajajo pa kere boma midva raziskovala

This paper is organized as follows:...

Data

We use four publicly available datasets for our exploratory analysis. We combine datasets [3], [6], and [8] into one large dataset (reffered to as Dataset SRB) as they include same categories of hate speech. We make labels *sexism*, *racism*, and *both* from [3] and [6]. The third dataset ([8]) that we use contains label *hostile sexism*, where marked tweets are already included in the first two datasets under *sexism*, and label *benevolent sexism*, which we rename to *benevolent*. We obtain a dataset with 6069 samples that are labeled either *sexism*, *racism*, *both*, or *benevolent*.

The fourth dataset (reffered to as Dataset AHS) that we use [9] has 4 categories - *abusive, hateful, spam*. As this is the original dataset no additional merging is needed. We obtain a dataset with 13776, with mentioned labels. Note that we exclude *None* label from both datasets, as we do not need it for the analysis.

We show the distribution of individual categories from datasets SRB and AHS in Figures 1 and 2, respectively. Note that the numbers of samples might not match the numbers in the original papers, due to the Twitter removing the tweets, making them unavailable for us to analyze.

We also provide an example for each label from both datasets. Some examples from the SRB dataset:

Racism - "He can't be a server at our restaurant, that beard makes him look like a terrorist." Everyone laughs. #fuckthanksgiving

Sexism - #katieandnikki stop calling yourselves pretty and hot..you're not and saying it a million times doesn't make you either...STFU

Benevolent - It's "NEXT to every successful man, there's a woman"

Examples from the AHS dataset:

Spam - RT @OnlyLookAtMino: [!!] #WINNER trending #1 on melon search

Abusive - You Worried About Somebody Bein Ugly... Bitch You Ugly...

Hateful - i hope leaders just kick retards that fake leave teams today

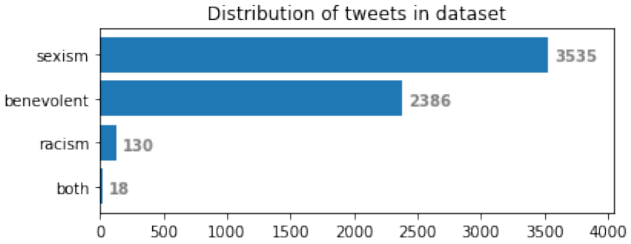

Figure 1. Distribution of tweets in SRB dataset. This figure shows the distribution of hate speech categories in the SRB dataset. We can see that *sexism* and *benevolent* are well represented, whereas *racism* and *both* are far less frequent. Original set contains more tweets labeled *racism*, but due to their removal we cannot obtain them.

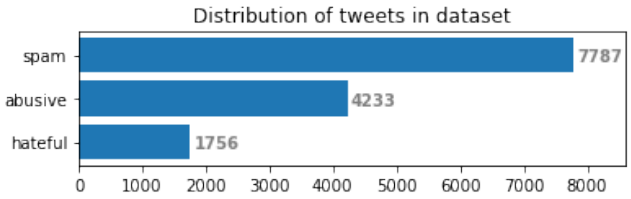

Figure 2. Distribution of tweets in AHS dataset. We see that the *spam* is the most represented label in the dataset, which represents the majority of the dataset. This is followed by the *abusive* tweets and there is the least *hateful* tweets. We can see that categories in this dataset are well represented.

category	unigrams with highest tf-idf score
racism	white, terror, coon
sexism	hot, hard, feminazi
benevolent	nasty, sassy, classy
abusive	hoe, fake, love
hateful	lgbt, religion, discrimination
spam	game, laptop, giveaway

Table 1. bb

Data preprocessing

Before applying any methods we first preprocess all of our data. We separate a whole dataset AHS on three parts, one for each category (abusive, hateful, spam), each containing multiple documents - tweets belonging to this category. Similarly we separate a whole dataset SRB on three parts, one for each category (sexist, racist, benevolent). We remove retweet text RT, hyperlinks, hashtags, taggings, new lines, and zero length tweets. We further filter out tokens that not contain letters, e.g., raw punctuation.

Methodology

TF-IDF

We start off with a traditional method TF-IDF as we want to see the most relevant words for each category of offensive language. We show the results in Table 1

Discussion

Use the Discussion section to objectively evaluate your work, do not just put praise on everything you did, be critical and exposes flaws and weaknesses of your solution. You can also explain what you would do differently if you would be able to start again and what upgrades could be done on the project in the future.

References

[1] *hate speech*. Lexico.com.

[2] Ellen Spertus. Smokey: Automatic recognition of hostile messages. In *Aaai/iaai*, pages 1058–1065, 1997.

[3] Zeerak Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142, 2016.

[4] Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90, 2017.

[5] Ricardo Martins, Marco Gomes, Jose Joao Almeida, Paulo Novais, and Pedro Henriques. Hate speech classification in social media using emotional analysis. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 61–66. IEEE, 2018.

[6] Marian-Andrei Rizoiu, Tianyu Wang, Gabriela Ferraro, and Hanna Suominen. Transfer learning for hate speech detection in social media. *arXiv preprint arXiv:1906.03829*, 2019.

[7] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. Conan–counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. *arXiv preprint arXiv:1910.03270*, 2019.

[8] Akshita Jha and Radhika Mamidi. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16, 2017.

[9] Antigoni 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. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.