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

Offensive language exploratory analysis

Maša Kljun, Matija Teršek

Abstract

In this paper we focus on the exploratory analysis of 10 different subgroups of hate speech. We use natural language processing techniques in order to find the underlying structure and connections/relations between the subgroups. We focus on data extracted from Twitter and online forums. First we use classic approaches, such as TF-IDF, BoW, and LDA, then we move on to more sophisticated methods such as embeddings. Note to professor: More to come in the next submission.

Keywords

Hate speech, TF-IDF, embeddings, exploratory analysis, NLP ...

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]. 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. In this paper we focus on ten subgroups of hate speech - *abusive, hateful, spam, general hate speech, profane, offensive, cyberbullying, racism, sexism,* and *benevolent sexism*. With understanding the structure of these groups, the goal is to also find similarities and connections between them.

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. Some examples include [8], who try to separate bullying from other social media posts and try to discover topic of bullying using topic modeling with Latent Dirichlet Allocation (LDA). [9] model hate speech against immigrants on Twitter in Spain. They try to find underlying topic of hate speech using LDA, discovering features of different dimensions of hatespeech, including foul language, humiliation, irony, etc. [10] conduct a survey about hate speech detection and describe key areas that have been explored, regarding the topic modeling, as well as sentiment analysis.

Note to professor: Our goal for the future is first a deeper analysis of the obtained data. We will try to extract some underlying topics for each label and present the main findings (LDA). Use TF-IDF to find the most common n-grams for each label. And in the future focus on the newer approaches, such as embedding with the further goal of exploring the

relationships between types of hate speech. We hope that we can determine a more precise plan on the first submission defense.

This paper is organized as follows: TODO at the end of the report

Data

We use six publicly available datasets for our exploratory analysis. We combine datasets [3], [6], and [11] 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 ([11]) 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)[12] that we use has 3 categories - *abusive, hateful, spam*. As this is the original dataset no additional merging is needed. We obtain a dataset with 13776 tweets with the 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.

Additionally we use the dataset of comments extracted from the League of Legends community [13]. We preprocess the dataset given in the SQL format to a more readable CSV form and keep only the posts that are annotated as harassment. We obtain 259 examples of cyberbullying examples.

The last dataset that we use was designed for the problem of the hate speech identification and classification, but we use the labels from the train and test set and merge them into one big dataset that we use for our analysis. It provides tags of *hatespeech, profane, and offensive*, so we refer to the dataset as HPO. It consists of 2549 tweets, distribution of which can be seen in Figure 3. For each subgroup of hate speech, we provide an example from the datasets.

Racism - ”He can’t be a server at our restaurant, that beard makes him look like a terrorist.” Everyone laughs. #fuck-thanksgiving

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”

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

Cyberbullying - plot twist she’s a fggt

Hatespeech - Johnson you liar. You don’t give a flying one for the Irish

Offensive - #FuckTrump And retired porn star Melania too.

Profane - Fuck Trump and anybody who voted for that Lyin POS! #FuckTrump

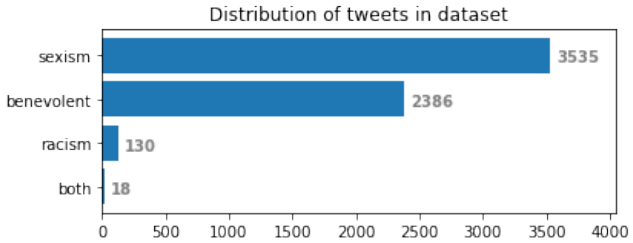

Figure 1. Distribution of tweets in SRB dataset. This figure shows the distribution of hate speech categories in the SRB dataset. We can see that *sexism* and *benevolent* are well represented, whereas *racism* and *both* are far less frequent. Original set contains more tweets labeled *racism*, but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

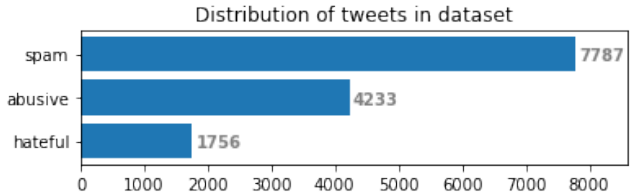

Figure 2. Distribution of tweets in AHS dataset. We see that the *spam* is the most represented label in the dataset, which represents the majority of the dataset. This is followed by the *abusive* tweets and there is the least *hateful* tweets. We can see that categories in this dataset are well represented.

Data preprocessing

Before applying any methods we first preprocess all of our data. We separate datasets into subgroups only, where each contains multiple documents - tweets belonging to this category. 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.

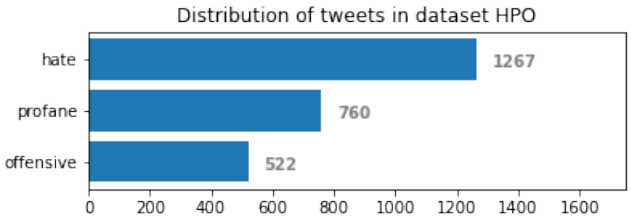

Figure 3. Distribution of tweets in HPO dataset. The most used label is *hatespeech*. It is followed by *profane* and then *offensive*, which have a similar number of tweets.

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

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
cyberbullying	TODO
hate	TODO
profane	TODO
offensive	TODO

Table 1. bb

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] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666, 2012.

[9] Carlos Arcila Calderón, Gonzalo de la Vega, and David Blanco Herrero. Topic modeling and characterization of hate speech against immigrants on twitter around the emergence of a far-right party in spain. *Social Sciences*, 9(11):188, 2020.

[10] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10, 2017.

[11] 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.

[12] 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.

[13] Uwe Bretschneider and Ralf Peters. Detecting cyberbullying in online communities. 2016.