FRI Natural language processing course 2021
Project report

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

Offensive language exploratory analysis

Juš Hladnik, Miha Torkar

Abstract

In this project we focused on exploratory analysis regarding hate speech and offensive language.

Keywords

Offensive language, hate speech, exploratory data analysis

Advisors: Slavko Žitnik

Introduction

In this project we do exploratory analysis on hate speech where we explore different datasets, how they are structured and labeled and try to find importance of specific words and phrases and relationship between types of hate speech. We make an overview of similar aproaches to detecting hate speech and we propose baseline methods that we will use to learn importance of specific words and relationships between them.

Datasets and literature overview

- Gomez et al. [1] create manually annotated multimodal hate speech dataset formed by 150,000 tweets, each one of them containing text and an image which they call MMHS150K. They classify each tweet in one of 6 categories: No attacks to any community, racist, sexist, homophobic, religion based attacks or attacks to other communities. They focus on detecting categories of hatespeech with both ext and images from tweets. First they train a LSTM on tweet text where they use pretrained GLoVe embeddings. They improve this model by incorporating also text from images. They then use various multimodal architectures to use text and images simultaneously.
- Qian et al. [2] introduce 2 datasets with full conversations instead of single posts, which is meant to provide context. Datasets consist of 5K conversations retrieved from Reddit and 12k conversations retrieved from Gab, both of whiche are social media platforms. For each conversation they provide labels for posts (hate speech/not hate speech) and human written intervention responses. For binary hate speech detection they use logistic regression and support vector machines with features as TF-IDF values of up to 2-grams. They also use CNN and RNN models for sentence classification with word2vec embeddings.
- Chung et al. [3] propose a large scale multilingual dataset with hate speech type and subtype annotations for English, French and Italian. They also provide counter speech responses in an effort to counter hate speech
- de Gibert et al. [4] create a dataset with sentences from Stormfront a white supremacist forum and they label each sentence whether it contains hate speech or not. Dataset consists of 10568 sentences. They have then trained a classifier on this dataset to detect hate speech, where they used SVM with bag-of-words features, they also used CNN and RNN with LSTM
- Davidson et al. [5] create a dataset with 24802 posts from twitter and they label them as 3 possible categories: hate speech, offensive language and neiher of those. They use this dataset to build features such as bigra, unigram, triram TF-IDF, and they also include features such as length of tweet, number of heshtags, mentions, retweets,... They then test a variety of models t: logistic regression, naıve Bayes, decision trees, random forests, and linear, SVMs.
- Mandl et al. [6] proposed a competition in which datasets in 3 languages were provided (1 of them in English). Dataset is developed from Twitter and Facebook and every entry is binary (hateful and offensive or not) and more fine-grained (hate, offensive and profane) classified, further for each it is declared if the entry is targeted or not. For English dataset consists of circa 8000 entries. For all tasks 321 experiments with different approaches has been submitted.
- Fersini et al. [7] similarly propose a competition in

which goal was to identify misogyny and a dataset was provided in English and Spanish, the English one consising of around 4000 tweets, which were labeled as mysognistic or not. Misogynistic tweets were further classified into 5 more categories and also the target was depicted. For classification in English 11 different approaches were submitted. All used more basic models, such as SVM, Bag of Word etc.

- Wulczyn et al. [8] provide a corpus of more then 100.000 comments from Wikipedia, which are labeled by human as attacks or not and if the comment has aggressive tone. Since more people were tasked with labeling we also get a fraction of people that labeled each comment as offensive. Researchers represent data as bag-of-words and build few models (logistic regression, multi-layer perceptrons). The models are then used on more then 63 million comments so the dataset we could potentially use is huge.
- Zampierieta et al. [9] also conducted a competition in identifying and categorizing offensive language in social media and provided a dataset of 14.100 3-level annotated tweets. First level is only a label if tweet is offensive or not, second and third level are about targets of offensive speech. They experiment classification with use of SVMs, BiLSTM (bidirectional Long Short-Term-Memory model) and CNNs. For competition part[10] more then 100 different models were submitted for all 3 parts of classification problem, for first level classification most successfull approaches used BERT, meanwhile on average on all 3 levels the most succesfull approach used ensembles.

Initial Ideas

- Combine and preprocess datasets so they have the same (or very similar) structure so we will be able to compare them.
- Compute and visualise word and phrase statistics, and how are they correlated with hate speech and different types of hate speech.
- Make different features for posts (tweets): bag of words, TF-IDF on unigrams, bigrams, trigrams. Try to improve with dense embeddings.
- Make clustering on words or posts to see if we can get clear clusters which we would then compare with actual labels of hate speech types. Also visualise which words are similar by measuring the distance between them.
- Train classifiers on different features to detect hate speech.

References

[1] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications, 2019.

[2] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. A benchmark dataset for learning to intervene in online hate speech, 2019.

[3] 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. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[4] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum, 2018.

[5] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515, 2017.

[6] 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*, pages 14–17, 2019.

[7] Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. Overview of the task on automatic misogyny identification at ibereval 2018. *IberEval@ SEPLN*, 2150:214–228, 2018.

[8] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399, 2017.

[9] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*, 2019.

[10] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), 2019.