FRI Natural language processing course 2021
Project report

University of Ljubljana
Faculty of Computer and
Information Science

Offensive language analysis

Mateo Kalem, Peter Horvat, Jernej Zupančič

Abstract

In this report we will present an overview of our task where we conducted an analysis into the recognition of offensive language, otherwise known as hate speech, racism, sexism, etc. We will take a look at a few datasets which contain messages and posts which contain harmful language and some different approaches which use natural language processing and machine learning to detect harmful speech.

Keywords
Offensive language, Hate speech, Natural language processing, Machine learning

Advisors: doc. dr. Slavko Žitnik

Introduction

Hate speech is a public form of speech that expresses hate or encourages violence towards a person or group based on a factor that is not common to both parties such as race, religion, sex, or sexual orientation. In the last few years we have seen its rise, especially in online forums or social media pages such as Twitter, Facebook, Reddit, etc. The detection of such speech is a critical challenge in *NLP (Natural Language Processing)*. Despite the existence of numerous *NLP* algorithms and methods dedicated to detect hate speech and offensive language, the results are sometimes very low. In our report we will take a look at some of these approaches and evaluate their effectiveness at detecting offensive language.

Related works

In this section we will take a look at some of the related works in this field of hate speech detection and NLP. The first related work is a paper [1] in which the authors combined multiple deep learning methods (such as RNN, CNN, ...) to detect hateful speech in tweets. Other works which we found are not contained in papers or articles, rather they are public Github repositories in which the authors implemented algorithms related to the problem which we are studying. In one of these projects [2] the author uses the NLTK toolkit for hate speech detection, the dataset which he uses contains tweets with harmful text within them.

In another related paper [3] the authors used Context Aware Models to detect hate speech in the comment sections of Fox News articles. With this method they reported an accuracy of about 75 % and a precision rate of 55 %.

Algorithms

A brief description of some of the projects we looked at:

- Classification of Offensive tweets, in which authors divided the algorithm in to three sub-tasks: Offensive language identification, Automatic categorization of offense types, Offense target identification
- Offensive Language Detection, authors here have divided their application in to different stages, they firstly pre-processed the text to remove any unnecessary stop words, emojis, mentions, urls and all kinds of noise, then the text is tokenized using TF-IDF vectorizer, and lastly this project used 6 different classifiers.
- Machine Learning and NLP methods for Automated Hate-Speech and Offensive Language Detection, here the authors focused mainly on tokenization. In particular, they used NLP methods to create feature spaces including weighted TF-IDF scores, TF-IDF matrix, N-grams, sentiment scores, and dependency-based features.
- HateSonar: Hate Speech Detection, is a python library which already includes pre-trained models. Just like in the other projects authors here have firstly cleaned the text, then computed the weights then classified it with scikit-learn.

Datasets

For the purpose of completing our task we must also specify some datasets which are related to our topic. One of the most widely used datasets is a set which contains Twitter hate speech annotations [4]. This dataset contains nearly 17.000 entries, which is perfect for our project. Another dataset which we intend to use is the Fox News Comments Corpus [5].We briefly mentioned this dataset when we described the article in the Related works section. Here are some of the datasets we found that we are lookin forward to use them in the testing phase.

- Fox News User Comments [5]
- Hate Speech Twitter annotations [4]
- Predicting the Type and Target of Offensive Posts in Social Media [6]
- Automated Hate Speech Detection and the Problem of Offensive Language[7]
- Hate Speech Dataset from a White Supremacy Forum [8]
- NLP and CSS 2017: Second Workshop on Natural Language Processing and Computational Social Science at ACL 2017 [9]
- Hateful Users on Twitter [10, 11]
- A Benchmark Dataset for Learning to Intervene in Online Hate Speech [12]
- MLMA Hate Speech [13]
- Hate Speech Twitter Datasets [14, 15]
- A Quality Type-aware Annotated Corpus and Lexicon for Harassment Research [16]

Since the datasets size and quality is of grave importance we collected as many datasets as possible. This way we can select the ones that will fit our requirements the best. If we will get to a point where we will need a even larger dataset, we will attempt to merge the best datasets. By comparing the results of the individual datasets and the merged dataset we will be able to conlcude if there was any deterioration in the results.

Initial idea

The initial idea of our project is to take a few different NLP methods and compare their accuracy when it comes to detecting hate speech. We also wish to compare them on different datasets. For this purpose we must also find many more datasets which originate from different forums and social media pages. These ideas are subject to change along the course of the 2nd semester.

References

[1] Nicolò Frisiani, Alexis Laignelet, and Batuhan Güler. Combination of multiple deep learning architectures for offensive language detection in tweets. *arXiv preprint arXiv:1903.08734*, 2019.

[2] Aman Saha. Hate speech detection. https://github.com/aman-saha/hate-speech-detection, 2018.

[3] Lei Gao and Ruihong Huang. Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395*, 2017.

[4] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics.

[5] Lei Gao and Ruihong Huang. Fox news comments. https://github.com/sjtuprog/fox-news-comments, 2017.

[6] 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.

[7] 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.

[8] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[9] 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.

[10] Manoel Ribeiro, Pedro Calais, Yuri dos Santos, Virgilio Almeida, and Wagner Meira Jr. "like sheep among wolves": Characterizing hateful users on twitter. 12 2017.

[11] Manoel Ribeiro, Pedro Calais, Yuri dos Santos, Virgilio Almeida, and Wagner Meira Jr. Characterizing and detecting hateful users on twitter. 03 2018.

[12] Jing Qian. A benchmark dataset for learning to intervene in online hate speech. https://github.com/jing-qian/A-Benchmark-Dataset-for-Learning-to-Intervene-in-Online-Hate-Speech, 2017.

[13] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. Multilingual and multi-aspect hate speech analysis. In *Proceedings of EMNLP*. Association for Computational Linguistics, 2019.

[14] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. Peer to peer hate: Hate instigators and their targets. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*, ICWSM '18, 2018.

[15] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Y. Wang, and Elizabeth Belding. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*, ICWSM '18, 2018.

[16] M. Rezvan, S. Shekarpour, L. Balasuriya, K. Thirunarayan, V. Shalin, and A. Sheth. A quality type-aware annotated corpus and lexicon for harassment research. arxiv. https://github.com/Mrezvan94/Harassment-Corpus, 2018.