FRI Natural language processing course 2021
Project report

University of Ljubljana
Faculty of Computer and
Information Science

Offensive language exploratory analysis

Rok Petrovčič Vižintin, Marko Krajinović and Tim Štromajer

Abstract

...

Keywords
Keyword1, Keyword2, Keyword3 ...

Advisors: Slavko Žitnik

Introduction

The internet has become a very common tool for people to express their opinions and communicate with each other. This role has become even more important in the latest virus era. But unlike in the real world, where different minded people usually live apart, here on the internet everyone is much closer. Different groups of people or individuals are being targeted and harassed with the use of social media, blogs, online gaming and more.

In this project we are working with the already annotated posts from different social networks, some of which contain offensive language like hate-speech, racism, sexism, stereotype and other. We analyze each offensive language class to find its specific features and compare it with other classes to find similarities. For this task, we use different methods. Term frequency-inverse document frequency (TF-IDF) reflects how important a word is to a document in a collection of documents. With this method, we obtain the most important words for each offensive language class and even which classes use similar vocabulary. With K-means clustering and hierarchical clustering we get a picture which offensive language classes are most similar to each other. Besides traditional methods, we also use contextual word embedding Word2Vec and non-contextual word embedding BERT. The Word2Vec algorithm is used to learn word associations from a large corpus of text to detect synonymous words and BERT is used for classification. For each conducted method we provide visualization to better clarify our findings.

For visualizations we utilize different techniques based on the text representations. Sparse vector representations are visualized with rooted and un-rooted dendrograms. Dense representations are linearly reduced to an intermediate lower dimension using PCA. The lower dimensional data is further reduced using a non-linear t-SNE embedding to produce 2 and 3 dimensional scatterplot visualizations. Another form of dense representation visualization is created with radviz. The best attributes are selected from the reduced space through a machine learning approach using vizrank.

Related work

0.1 Predicting the Type and Target of Offensive Posts in Social Media [1]

The authors of [1] set out to create an annotated dataset of offensive tweets called *OLID (Offensive Language Identification Dataset)* and test different natural language processing algorithm's performance on that dataset.

The dataset uses a three level annotation schema:

1. Level A: offensive or not offensive
2. Level B: targeted or untargeted insult
3. Level C: who the target of the insult is - a group, an individual or other

The data set was compiled using the Twitter API and searching for specific keywords, 50% of those political and 50% non-political. It contains 14,100 English tweets, around 30% of those offensive. The set was annotated by hand by crowdsourced annotators.

The authors tested the dataset with a Support Vector Machine (SVM), bidirectional Long Short-Term-Memory model (BiLSTM) and a Convolutional Neural Network (CNN). The models were evaluated on the basis of detecting all three annotation levels, using a macro-averaged F1 score. They found that the CNN yields the best results.

0.2 Unsupervised Cyberbullying Detection via Time-Informed Gaussian Mixture Model [2]

The authors of [2] present an unsupervised approach for cyberbullying detection in social media comments. Their motivation for an unsupervised approach was the time and labor intensity of manual data annotation and to future-proof the model against the evolution of language. Their method differentiates itself from other methods by using comment inter-arrival times of a social media session, which enables the classification of cyber-bullying instances using the full commenting history. Their model consists of a representation learning network which learns the multi-modal representations of a session and a multi-task learning network, which predicts if a social media session contains bullying. The representation learning network uses a Hierarchical Attention Network for textual features and a Graph Auto-Encoder for user and network features. The two learning models are optimized in a way that boosts each-other's learning effectiveness.

The experiments they conducted were done on two datasets of Instagram and Vine sessions containing videos, captions and comments, annotated as cyberbullying or non-cyberbullying by crowdsourced annotators. From each session they gathered a user's social connections, text as a bag-of-words representation of the captions and comments and the timestamps of all the session's elements. They compared their model with different unsupervised and supervised learning models. They used Precision, Recall, F1 and AUROC metrics to compare the models' performance. They found that their model outperforms other unsupervised methods, and achieves competitive performance against supervised models.

0.3 Detecting Hate Speech in Social Media [3]

The authors of [3] explore the topic of detecting hate speech in social media content. They use a data set of English tweets annotated as *hate speech*, *offensive* or *not offensive*. Their aim is to establish a lexical baseline to discriminate between hate speech and profanity on the used dataset.

They use a linear Support Vector Machine classifier and three groups of features: surface n-grams, word skip-grams, and Brown clusters. For result evaluation they use 10-fold cross-validation with folds created via stratified cross-validation. The results of their method are compared to a majority class baseline and an oracle classifier, to establish the theoretical upper limit performance for the dataset. They found that character 4-grams, word unigrams and 1-skip word bigrams produced the best accuracy of around 77% compared to the theoretical limit of 91.6%. They also note that accuracy increases with the amount of training instances, plateauing around 15,000 instances.

References

[1] 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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Lu Cheng, Kai Shu, Siqi Wu, Yasin Silva, Deborah Hall, and Huan Liu. Unsupervised cyberbullying detection via time-informed gaussian mixture model. 08 2020.

[3] Shervin Malmasi and Marcos Zampieri. Detecting hate speech in social media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 467–472, Varna, Bulgaria, September 2017. INCOMA Ltd.