Explaining the qualitative differences between different classifiers on cyberbullying datasets

Yiyang Cheng
12075760@student.uva.nl

Amaan Syed
amaansyed099@gmail.com

Tori Baral
toriburi@gmail.com

Abstract

Twitter is a platform on which 15,000 cyber-bullying tweets are sent out daily. Tweets can be reported as abuse by users, but Twitter also aims for auto moderatation of tweets, and removes abusive tweets even before they are published using artificial intelligence. In this paper we aim to test various classifiers on their ability to classify a tweet as abuse or not, as well as using explainable Artificial Intelligence (XAI) to pinpoint which specific words contribute to the classification of a tweet, using LIME. We also visualize our data and results for further clarity of what contributes to the bias for abusive language. Our project aims to provide insight on the decisions made by auto moderation classifiers for social media platforms.

1. Introduction

Our project aims to answer the following research questions: Which classifiers perform best in classifying a tweet as abuse? What words contribute to classifying tweets as abuse?

2. Related Works

We were inspired by several academic sources while doing our project. These include works about dealing with tweets and short text data in general, as well as insights into explainable AI which is a new topic for us all, not discussed much in class.

2.1 Target-dependent Twitter Sentiment Classification (Jiang et al., 2011)

This paper gave us information about classification of Twitter data. They use a target-based approach for sentiment analysis of tweets. However, we gathered from this paper more insights on how to work with short and ambiguous tweets and classifying them with such limited content and context.

2.2 Short Text Topic Modeling Techniques, Applications, and Performance: A Survey (Qiang et al., 2019)

This paper too was more insightful on how to work with short text data, such as tweets. Short text data analysis depends less on word-cooccurrences unlike long text data. However, this paper focuses more on the above problem in the context of topic modeling, which we ultimately did not add to our project.

2.3 Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI (Arietta et al., 2020)

This paper was very helpful to us as we began looking into explainable AI. It was significant in that it introduced concepts related to XAI, as well as comparing different models' performance on the same task.

Especially the table on page 14, showing the explainability of various Machine Learning models. Their methods of visualisation also inspired us in our project.

2.4 Explanation in Artificial Intelligence: Insights from the Social Sciences (Miller, 2018)

This paper generally gave us a broad understanding of XAI, and how explainable and interpretable AI can have significance when applied on non-scientific data- in our case, tweets. This was insightful for us when reviewing our results, classifying twitter data as bullying or not.

2.5 "Why Should I Trust You?" Explaining the Predictions of Any Classifier (Ribeiro et al., 2016)

This reading was significant to us as it introduced LIME, which we also implemented in our project for XAI. LIME simplified our project a lot, and facilitated us to compare the XAI results of our different classifiers. This paper also touched upon the comparison of different models, and deepened our understanding of 'trusting' biases present in artificial intelligence and the importance of explaining and understanding the biases present in the results of a classifier

3. Data Collection

The data for training was taken from here: Cyberbullying datasets
We specifically chose these files from the above link:

1. twitter_racism_parsed_dataset.csv
2. twitter_sexism_parsed_dataset.csv
3. twitter_parsed_dataset.csv
4. kaggle_parsed_dataset.csv

This is because they were all of manageable size, labeled, not very noisy, and tweets of a similar format, i.e., to facilitate pre-processing.

4. Data set Description

The above files contain 25596 tweets in total, after filtering out duplicates. The .csv files contain columns with the tweets' text, annotation such as 'racism', 'sexism'; and a Label, which is 1 for abusive tweets and 0 for non-abusive tweets. There were also columns such as Date and ID, which we chose to neglect for our training.
We chose this website because of its labeled, cleanly formatted data. The compilation of tweets contains a 2:1 non-abusive to abusive ratio (after removing duplicates).
There are a few instances of non-english tweets, but we also chose to neglect these as language detectors we tried (such as lang-detect, Spacey, and Google's) proved to be very unreliable for us.
For testing, we chose this labeled dataset from the same website: attack_parsed_dataset.csv
This contained 115668 tweets (after removing duplicates) and were a bit noisy, but we filtered out the noise and stopwords.

5. Methods & Main Algorithms

To pre-process our data, we focused on removing stopwords, punctuation, numbers, common words such as 'rt' which signifies 'retweet', as well as hashtags and @ handlers. We used nltk's TweetTokenizer to tokenize our tweets.

We explored the preprocessed dataset in three ways.

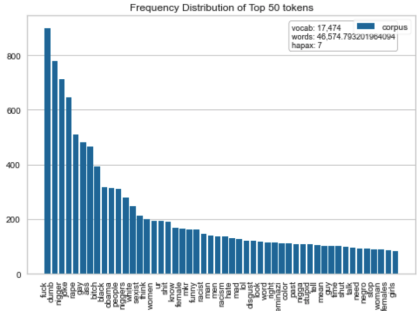
Firstly, we visualized the dataset and marked all occurrences of words, along with their annotations and labels. We were able to map the token frequency of our dataset.

After an analysis of the dataset we moved on to train three machine learning classifiers based on the dataset- Support Vector Machine (SVM), Random Forest classifier (RF), and Neural Network (NNw). For building the models, we implemented the scikit-learn (sklearn) package. By training these models we aimed to find the best performing model to predict abusive language. We also planned to look inside those models, inspired by explainable AI, but due to limited time we will put off this idea into future plans.

We were fascinated by LIME, a package that explains machine learning classifiers. We explored this toolkit in order to give individual predictions for text classifiers or classifiers that act on our dataset. We also took advantage of its built-in visualization. At the final stage, we aimed to make a prediction for each token whether it is abusive language or not.

6. Results & Findings

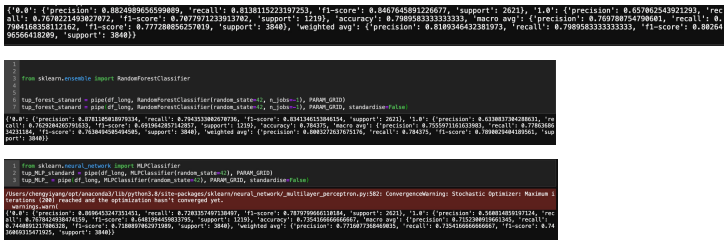
The word frequency of our text dataset had a good coherence to Zipf Law, with the frequency decreasing geometrically as the percentile of the word falls.



[pic1: 50 most common words in ‘abusive’ tweets]

We obtained the performance scores of all of our models. The SVM performs the best with highest overall accuracy, interestingly. Possibly because of the relatively small input data size.

It is also important to mention that all classifiers when they are not standardly scaled (for mean, standard deviation, etc.)



[pic3, 4: Results of SVM, RF, NNw]

[pic4: a selection of LIME results]

Limitations include selecting only the best performing results of models in this report.

7. Conclusions

A large proportion of costs of social media like Twitter is spent to counter abusive language that attacks individuals using the platform. Manually cleaning cyberbullying speech, although precise, is too slow and costly. Automatic detection of abusive messages is therefore a valuable issue. We obtained a labelled dataset online about cyberbullying and, after some brief preprocessing, visualized the clean data, trained classifier models and attempted to predict whether a single word would be abusive.

Further improvement of this topic include introducing state-of-the-art models such as BERT, fine-tuning datasets so as to minimize

class imbalance, and visualization techniques are more compatible with PyTorch and TensorFlow rather than sklearn which hindered us mightily.

8. References

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-Dependent Twitter Sentiment Classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (pp. 151–160). Association for Computational Linguistics.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38.

Qiang, J., Zhen-yu, Q., Yun, L., Yun-hao, Y., & Xin-dong, W. (2019). Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. arXiv: Information Retrieval.

Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (pp. 97–101). Association for Computational Linguistics.