Cyberbullying Auto Detection

Wenqu Wang
School of Information
University of California, Berkeley
Berkeley, CA, USA
jackw@ischool.berkeley.edu

Casey Yoon
School of Information
University of California, Berkeley
Berkeley, CA, USA
caseyhyoon@ischool.berkeley.edu

Abstract

Cyberbullying has had an immense impact on mental health for its victims on social media platforms. For these victims, creating a system to address/prevent these negative encounters would help create a more positive community. This project explored the use of neural networks and feature extraction techniques to classify sentiment on Twitter data.

***Index Terms*—Machine Learning, Natural Language Processing**

I. Introduction

The development of mobile internet has made it possible for social media platforms, like Twitter and YouTube, to amass such popularity and have largely altered people's lifestyles by providing users with a new way of communication that is free, convenient, and entertaining. However, these platforms have also given way to cyberbullying, a new form of electronic violence characterized by its lack of retribution for its perpetrators and whose victims range from normal personnel to celebrities or popular influencers/bloggers. Cyberbullying has had an immense impact on mental health for its victims, many of whom suffer depression and anxiety; as a result, suicide has become an unfortunate consequence due to the mental and even physical reactions from cyberbullying.

The challenges to this problem, like most sentiment analysis problems, are the fluidity of language and the context in which cyberbullying comments are posted. This context includes the target of said cyberbullying and topic of conversation. In this project, we will utilize and compare several algorithms to find the highest classification accuracies on datasets from Twitter. By applying natural language processing (NLP) techniques, cyberbullying comments can be detected, screened/blocked in the hopes of creating less hostile online communities.

II. Overview

A. Background

Cyberbullying is becoming a serious problem from the rapid growth in the number of internet users. There are about 4.66 billion internet users all over the world today, and this number grew by 319 million in the past 12 months. With the vast amount of users online, it is very difficult to control for inappropriate behavior. Fortunately, the growth of techniques related to natural language processing provides a potential solution, which is efficient and accurate, to the problem of cyberbullying. By analyzing user sentiment using algorithms, cyberbullying comments can be detected and further blocked by the server.

In the rest of this section, we will discuss our approach to solving this problem using NLP.

B. Dataset

We initially explored Cyberbullying Datasets from Mendeley Data [1], which contains labeled user comments data collected from various platforms like Kaggle, Twitter, and Youtube. We chose the following three datasets for sentiment analysis on Twitter:

Twitter_parsed_dataset.csv
Twitter_racism_parsed_dataset.csv
Twitter_sexism_parsed_dataset.csv

We are interested in tweets that identify to be racist or sexist – these negative traits go against community guidelines and are the focus of our classification. Merging all three datasets together, the dataset has 45197 entries. 34503 (76.33%) of entries are neutral, while the other 10694 (23.67%) tweets have been labeled either racist or sexist.

C. Sample Tweets

RT @Mooseoftorment Call me sexist, but when I go to an auto place, I'd rather talk to a guy.

A good Muslim is good despite his bad religion, not because of it.

@DianH4 Islam doesn't answer anything. It pretends to answer with illogical and delusional superstition.

The above messages contains sexism or racism content and is marked positive in the dataset.

Woo can't wait to see what happens!!! #mkr

@halalflaws @biebervalue @greenlinerzjm I read them in context.No change in meaning. The history of Islamic slavery.

@OneLegSandpiper @DblBlackDs Show me some pictures of them beheading people and reinstating slavery. Leftist moral equivalents are stupid.

These messages are examples of non-bullying tweets. Noticed that some of these tweets also contain bad sentiment or racist keywords, but the main focus of the tweet is not to bully someone.

[1] https://data.mendeley.com/datasets/jf4pzyvnpj/1

D. Data Cleaning

We were given raw tweets and the fluidity of language present in twitter data, prompted much consideration before tokenization:

- Addressing Twitter handles (@user)
- Removal of punctuation, numbers and special characters
- Lowercasing

Additionally, we did not remove hashtags and retweet labels (RT). We include hashtags into our vocabulary in order to not take away unique sentiment values related to certain hashtags and we don't remove retweet labels as they are superficial. Once we clean our tweets, we split them into tokens to prepare for featurization.

E. Word Embeddings

In order to train classification models, our preprocessed data need to be converted to features. Tf-idf vectorization, which accounts for word frequencies, will be used as our baseline featurization method. Counting the number of times a certain token appears in a corpus is a simple plan for us to identify relationships between frequencies and text sentiment; however, nothing more.

Several existing research papers have proposed use of Word2Vec word embeddings in future research to initialize classification models, specifically convolutional neural networks. [1](Huang, Qianjia, et al.) Word2Vec, as a shallow neural network used to obtain word vectors, better represents our tokenized tweets as it utilizes word context.

Finally, Bidirectional Encoder Representations from Transformers (BERT) was another pre-trained processing model that piqued our interest because it was a higher level tokenizer that maintained sequence information.

III. Models

A lot of existing machine learning algorithms and models can be applied to the problem of cyberbullying auto detection. The authors focus specifically on the performance of the combinations of different word embeddings and models. In this section, the authors will explain four models: Baseline Model, Tf-idf with Feed-Forward Neural Network, Word2Vec with Convolutional Neural Network, and BERT.

A. Baseline Model

The authors built a naive one-class classification model with an accuracy of 0.76, that is, if we simply classified a tweet to be racist or sexist, it would be correct about 76% of the time. We expect any potential models to be at least better than this model.

Instead, Tf-idf word embeddings with a logistic regression model was used as a primary baseline comparison as we introduce alternative models and also word embeddings that hopefully capture significantly better results. We achieve an accuracy of about 85% with this baseline.

B. Tf-idf with Feed-Forward Neural Network

By training a 2-hidden-layer feed-forward neural network with Tf-idf embedded data, the model achieved an accuracy of 94.8%. In order to handle the input dimension for neural networks, authors decided to use 500 hundred as the max number of features of Tf-idf so that each input has 500 features. After the input layer, the data is then passed to two hidden layers with ReLU activation. Finally, a sigmoid layer will map the hidden layer output into a number between zero and one so that the final output could be determined.

The training process takes about one second to run for each epoch, which is pretty efficient.

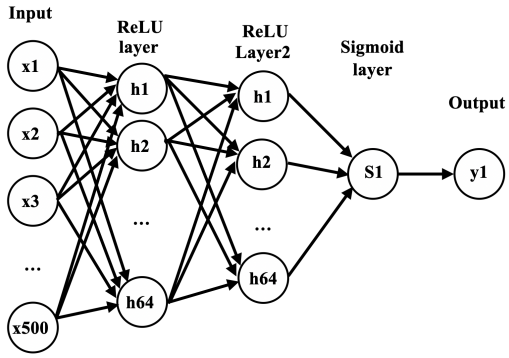

Fig. 1. Feed-Forward Neural Network Architecture

C. Word2Vec with Convolutional Neural Network (CNN)

Entering the tokenized tweets into a size-200 word2vec, the model generated a vocabulary of size 19645, which is approximately 20000. As a result, a 10000 * 200 embedding matrix is built and will be put into CNN as the embedding layer. After the embedding layer, the word vectors are then passed into a layer of 1D convolution and another max-pooling layer, and finally the fully connected layers.

Under such architecture, the model achieved accuracy of 98.04%, which improved significantly compared to the previous model. As for efficiency, each batch takes about 83 seconds to train, which is fairly efficient.

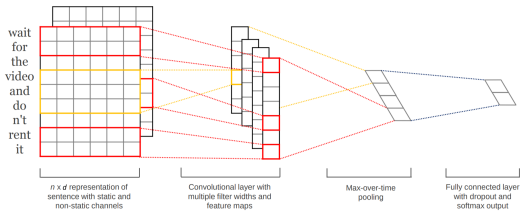

Fig. 2. Convolutional Neural Network Architecture [2]

D. BERT

Initially the authors employed a BERT tokenizer, pretrained on an uncased corpus, and utilized default settings. However, we quickly learned that classification with BERT default

parameters yields predicting the majority class. With specific optimizer parameters and freezing a few layers, it allowed for the model to learn better than its default counterpart.

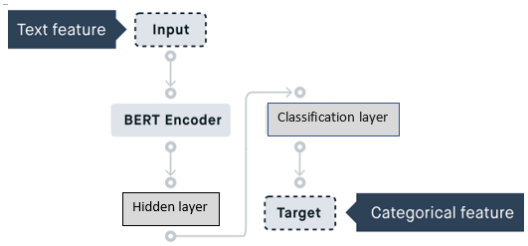

Fig. 3. BERT Structure

IV. Results

TABLE I
Model Results

Models	Accuracy
Tf-idf with Logistic Regression	84.85%
Tf-idf Feed-Forward Neural network	95.08%
Word2Vec with CNN	98.04%
BERT	89.05%

Tf-idf with Logistic Regression is too low to use as a model baseline; Tf-idf Feed-Forward Artificial Neural network is a good baseline as it accurately classifies a good portion of the data with room for improvement. With our own research Word2Vec with CNN performed the best with a test accuracy of 98.04%. However, BERT on its own, did not perform as well as we had hoped with an accuracy lower than our baseline at 89.05%.

V. Error Analysis

A. False Positive Tweets

RT @Juliet777777: Video Pat Condell "Nothing to do with Islam" The death rattle of a dhimmi society..

@Lithobolos @PoliticalAnt @ZaibatsuNews Christianity is dying. Less attendance all the time. Total liberalization of the religion.

RT @adriarichards: With so many factual errors @NYTimes piece about online harassment by @jonronson, it's a disservice to everyone.

Most false positive tweets contain words with negative sentiment and display emotionally charged words without specifically attacking any individual or group.

B. False Negative Tweets

Kat's not a morning person. Or a midday person. Or an afternoon person. Or an evening person. Wait. Is she even a person?? #mkr #MKR2015

RT @itsbariecool @Mccheesy904 it happens vice versa but men are smarter naturally #NotSexist

Maybe the girls should have less tickets on themselves and worry about the cooking. #MKR

Most false negative tweets employ an underlying tone of sarcasm and surprisingly a good amount of them contain hashtags like #mkr; which may represent some form of unique sarcastic sentiment. False negative tweets may be more difficult to classify due to hidden meanings and less direct forms of attack. Also, some bullying tweets contain hashtags like #notSexist or #notRacist, which makes it harder for the model to discern them from actual neutral tweets. More examples of such attacks may be needed for an algorithm to detect such outliers.

VI. Conclusion

Word2Vec with CNN performed the best as the training and test accuracies are above 98%. With little room for improvement, no amount of optimizations will be able to increase training and test accuracies and error analysis indicates that improvements will require targeting of specific tweets; our NLP models are not able to detect such outliers. Cyberbullying will not be completely eliminated as it evolves and rears its head in different forms. Though our model correctly classified 98% of tweets, we are still very concerned about our type II error rate as the impact of an unprevented cyberbullying tweet is immense. With 4.66 billion internet users, if 1% of this population were to receive a cyberbullying message, that would be tremendous. For global communications platforms, a sentiment analysis model with an error rate of <1% is essential to preserving a safe community. CNN along with feature extraction methods and a word embedding feature superior to Word2Vec would potentially be a more successful candidate for model improvement. One possible combination would be to utilize BERT with CNN.

References

[1] Muneer, Amgad, and Suliman Mohamed Fati. "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter." MDPI, Multidisciplinary Digital Publishing Institute, 29 Oct. 2020, www.mdpi.com/1999-5903/12/11/187

[2] Hee, etcl., 2018. "Automatic Detection of Cyberbullying in Social Media Text", Arxiv. https://arxiv.org/pdf/1801.05617.pdf

[3] Huang, Qianjia, et al. "Cyberbullying Intervention Based on Convolutional Neural Networks." ACL Anthology, Association for Computational Linguistics, Aug. 2018, www.aclweb.org/anthology/W18-4405/

[4] Zhao, etcl., 2016. "Automatic Detection of Cyberbullying on Social Networks based on Bullying Features", ACM Digital Library. https://doi.org/10.1145/2833312.2849567

[5] Sahay, etcl., 2018. "Detecting Cyberbullying and Aggression in Social Commentary using NLP and Machine Learning", International Journal of Engineering Technology Science and Research. http"//www.ijetsr.com/images/short_pdf/1517199597_1428-1435-oucip915_ijetsr.pdf