Methodology Report

By Ben Schmidt and Armand Heydarian

**Problem:**

Twitter has been increasingly utilized for the diffusion of hate-based ideologies and world-views. This becomes prolonged due to the anonymous environment and mobility of the platform. There is a lack of an efficient automatic hate speech detection model based on natural language processing and machine learning, particularly with regards to differentiating between racist and anti-racist speech. One of the key challenges of hate speech detection in the realm of social media is the need to separate offensive language from true hate speech. The lexical detection methods that are used tend to have low precision as they classify all messages containing particular terms as hate speech; with previous work using supervised learning continuing to fail to distinguish between the two categories. This project aims to analyze and compare posts on the Twitter platform to improve the accuracy rate of the model, particularly related to false positives (offensive language being flagged as hate speech). Prior attempts to do so utilizing other NLP methods have shown to have high false-positive rates. In order to accomplish this comparative analysis, this project will be utilizing a Twitter dataset containing hate speech labeled from a 2017 paper (CITATION) which used older NLP methods. This project will seek to improve accuracy and speed with iterations of BERT. BERT’s advantage in this effort is it’s capability of working across a framework and with different pre-trained models, which can either be: contextual or context free; and unidirectional or bidirectional.

**Methodology:**

For effective comparison this project uses the benchmark dataset created by \_\_\_\_\_. After gathering the data, the researchers need to preprocess the data by tokenizing the Tweets. The Tweets will then be converted into a list of words, with the researchers removing unwanted characters and words as well as words that are not part of the English dictionary. This list will be the corpus of words. The researchers will utilize BERT embeddings to convert a particular word in the Tweet to its numerical vector representation. Next, the Researchers will employ the usage of batching and padding to convert each word to a vector representation of fixed/uniform size. This enables a fixed size of the input data, which is a precondition for any machine learning model. Following tokenization of the Tweets, the training and testing sets need to be formed. To form the training and testing data, the researchers have used a 70-30 split respectively. In this BERT based neural network, the structure consists of several layers. The first input layer is the embedding layer, in simpler terms this means the embeddings which the researchers generated from preprocessing act as an input to the neural network. The next three layers are convolution layers using the ReLu activation function with max pooling layers in between for dimensionality reduction, after this the model produces a dense layer to generate a scalar representation of the data. Finally, the model produces the output layer with three output units using soft max activation for generating the output class. The researchers then move on to testing the model and once the model is trained for a predefined number of epochs, it is tested on the testing data in terms of performance or evaluation metrics such as accuracy, precision and recall.  Additionally, this project seeks to use additional data gathered from #BlackLivesMatter to compare against the benchmark data to study impact on the model against anti-racist speech.

**Literature Review:** Based on our initial research and literature review, we found that if we conﬂate hate speech and offensive language then we erroneously consider many people to be hate speakers and fail to differentiate between commonplace offensive language and serious hate speech. Since there could be legal and moral implications of hate speech, it is important that we are able to accurately distinguish between the two. We found that other attempts with this dataset using lexical methods are an effective way to identify potentially offensive terms but are inaccurate at identifying hate speech; only a small percentage of tweets ﬂagged by the hate base lexicon were considered hate speech by human coders. It was also found that automated classiﬁcation methods can produce relatively high accuracy at differentiating between these different classes, close analysis shows that the presence or absence of particular offensive or hateful terms can both help and hinder accurate classiﬁcation. Since a sizeable portion of the tweets that are considered most hateful contain multiple racial and homophobic slurs, this helps us to easily identify some of the appaling instances of hate speech but it also means that we are more likely to misclassify hate speech if it doesn’t contain any curse words or offensive terms. To more accurately classify such cases we should ﬁnd sources of training data that are hateful without necessarily using particular keywords or offensive language. We plan on adding additional data relating to current events as previously mentioned.

**Key Differentiation:** Unlike the previous project, which used TF-IDF and bigrams/trigrams for word representation this project is employing word embeddings which is basically one level deep. In addition to bigrams/trigrams features, this project employed bi-directional relationships between the words of a particular tweet to learn the semantics. Another difference in terms of models, they are using Naive Bayes, random forest and linear SVMs, but this project is employing a multi-layer convolution neural network to gather insights at a deeper level than previous attempts.

BLOCK Diagram for Model



Preliminary Results (Changing learning rate is the next major attempt)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Epochs | mcc | f1 | acc |
| Bert | 5 | 0.7338331950032495 | 0.9043892896442108 | 0.9043892896442108 |
|  |  |  |  |  |
|  |  |  |  |  |

**Next Major Steps:**

**TEST MODEL WITH NEW DATA and EVALUATE the results**

**Literature Sources:**

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. "Automated Hate Speech Detection and the Problem of Offensive Language." Proceedings of the 11th International Conference on Web and Social Media (ICWSM).

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media.

Ribeiro, M., Calais, P., Santos, Y., Almeida, V., & Meira Jr, W. (2018). Characterizing and Detecting Hateful Users on Twitter.