Improving Hate Speech Detection With Twitter

Project Proposal Report

By Ben Schmidt and Armand Heydarian

**Problem**:

With the increased use of social networking platforms such as Twitter, people are using such platforms to disseminate information on a large scale, they have been increasingly utilized for the diffusion of hate-based ideologies and world-views. Online hate-based content has fueled extremism throughout the world. This problem is then only prolonged due to the anonymous environment and mobility of the platform. There remains a problem of a lack of an efficient automatic hate speech detection model based on natural language processing and machine learning. One of the key challenges of hate speech detection in the realm of social media is the need to separate offensive language from what is considered to truly be hate speech, while these lexical detection methods tend to have low precision due to the fact that they classify all messages containing particular terms as hate speech. Previous work using supervised learning has failed to distinguish between the two categories. This project aims to analyze and compare posts within the Twitter platform in order to improve the accuracy rate of the model, particularly related to false positives as prior attempts to do this utilizing other NLP methods had shown to have a high false-positive rate of identifying hate speech. In order to accomplish this comparative analysis, this project will be utilizing a Twitter dataset containing labeled hate speech from a previous paper in 2017 that used older NLP methods, this project will seek to improve accuracy and speed with iterations of BERT. BERT will give us an advantage in this effort with it’s capability of working cross framework and with different pre-trained models. This helps as the pre-trained representations can either be contextual or context free, these contextual representations can also either be unidirectional or bidirectional.

**Dataset:**

The dataset for this project will be the dataset gathered by the authors of "Automated Hate Speech Detection and the Problem of Offensive Language." (2017). This dataset contains ~25000 mutliclassed labeled offensive comments and tweets from social media. This data is labeled as offensive but not racist, racist, and not offensive. These tweets were searched on the basis of the lexicon, yielding tweets from over 33,000 unique Twitter users and the timeline of each user was extracted, allowing us to create a sample of over 25,000 tweets. This project does not seek to address issues related to gathering hate speech, which is heavily influenced by bias of data collectors. This issue would need to be addressed in future work.

**Proposed Model**:

A key focus of this project is to utilize and fine tune the BERT algorithm (Bidirectional Encoder Representations from Transformers) to encapsulate hateful content within the Twitters comments due to its features being the most poignant for predicting hateful content. The model would help as it has its advantages of applying the bidirectional training of Transformer models to language modeling BERT models have also been proven in the past with its application to offensive language identification in prior projects. In the effort of this project, BERT will be utilized with the predescribed dataset from Davidson, et. al. to apply newer models to accepted benchmark datasets. As the focus is on improving the accuracy and efficiency of the model accuracy, experiments will focus on multiple iterations of BERT and deep learning frameworks to identify the most efficient and accurate models with this dataset.

**Project Github:**

[**https://github.com/armandheydarian/Hate-Speech-Detection-in-Social-Media/**](https://github.com/armandheydarian/Hate-Speech-Detection-in-Social-Media/)

**Sources:**

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., … Kourtellis, N. (2018). *Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior*.

Gomez, R., Gibert, J., Gomez, L., & Karatzas, D. (2019). *Exploring Hate Speech Detection in Multimodal Publications*.

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

Qian, J., Bethke, A., Liu, Y., Belding, E., & Wang, W. (2019). *A Benchmark Dataset for Learning to Intervene in Online Hate Speech*.

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

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).

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the Type and Target of Offensive Posts in Social Media.