21100138 Taha Razzaq

Automated Hate Speech Detection

One of the biggest problems faced during automated hate speech detection is separating hateful speech from offensive language. Although no formal definition of the terms exists, many countries, like United Kingdom and France classify hateful speech as the speech that promotes violence and disorder while targeting minorities. Building upon the variants of the definition, hateful speech can holistically be defined as the language used to express hatred towards a certain group or the language used to humiliate the members of the group.
The paper aims to classify a set of tweets into 3 distinct groups - hateful speech, Offensive language and neither - and highlight the key challenges faced in automated hate speech classification.

Classifying text gathered from social media as hateful speech is not a novel idea and many researchers have worked on this problem. A common approach used in tweet classification is Bag of Words however, since it has a high recall, the number of false positives present are alarmingly significant. Another common problem faced by researchers is the high prevalence of offensive language in hateful speech. Since the hateful text contains a great number of offensive words, the model ends up incorrectly classifying them as offensive language. The workaround for the problem is to use syntactic features that depict the intensity of hate speech and non-linguistic features like gender or ethnicity. The latter option is rarely used since information on social media is either missing or incorrect.

The paper documents results of classification performed on a set of tweets collected using the Twitter API. The tweets used contained words present in the hate speech lexicon available at Hatebase.org. From a corpus of 85.4 million tweets, 25k tweets were randomly selected and manually coded by CrowdFlower (CF) workers. The tweets were labelled by 3 workers and an inter agreement core of 92% was achieved. The workers were given set definitions for the 3 classes and the tweets which did not have a majority class were discarded. Eventually, 24,802 tweets were used. Out of 24,802 tweets only 5% were labelled as hate speech while 76% were offensive and 16.6% were neither. The number of tweets labelled as hate speech by the CF workers was considerably lower than the percentage flagged by twitter since a strict criteria was used.

The NLTK and Scikit toolkit was used for feature extraction and model implementation. The Porter stemmer produced bigram, unigram and trigram features of the lowercase tweets, along with their corresponding part of speech tags. Furthermore, a sentiment score was assigned to each tweet according to the sentiment lexicon designed for social media.
Initially Logistic regression with L1 regularization was applied on the data in order to reduce its dimensionality. Multiple models were tested and the grid search showed that Logistic regression and linear SVM out performed the others. Logistic regression with L2 regularization was the final model used, since it enabled predicted probabilities to be examined easily. Being a multiclass classification problem, One-Vs-Rest framework was employed and the model was trained using a part of the data collected (train data).

The predictions made by the model showed that it had a 0.91 precision score, with a 0.90 recall and 0.90 F1 score. However 40% of the hate speech was misclassified. The model had biases and it classified less intense hate speech as an offensive language. Tweets containing words related to strong hate speech like multiple racial and homophobic words were correctly classified, however other tweets with lower occurrences of hate words were misclassified. Other instances of misclassification of hate speech occur since only the words are considered and the context is ignored. Tweets with a hate context are misclassified since they contain very less hate words. Although, the classifier works well for tweets highly populated with hate words, however, it tends to fail otherwise. Tweets originally belonging to the Offensive language and neither class were also misclassified due to their context being ignored.

Differentiating between hateful speech and offensive language is an important and a crucial matter since governments tend to take serious action against hateful speakers. Although, lexical methods are able to differentiate between these two types of speech to in extreme cases, however they do not work well for other (medium) cases. The results showed that the hate speech can be used to target a specific group or minorities, can be general and also be a conversation between people. Since, major researches and projects are being carried out in this regard, it is important that future researchers understand inclusion of the human bias in the model. Hateful speech is subjective and labellers label tweets according to their definitions.Future researches need to ensure that they identify and remove the human bias before classifying texts.