Automated hate Speech Detection and the Problem of Offensive Language

The goal of this study was to highlight the challenges faced in differentiating hate speech from offensive language. A definition for hate speech has not been officially formulized even though it is an agreed upon notion that hate speech is a language used to spread hatred and derogatory remarks which are targeted towards a specific group. They tried to construct their own definition of hate speech by making use of previous definitions. However, they mentioned that hate speech cannot be restricted to language that instigates violence as it could narrow down its definition. People often used phrases that would be considered as hate speech to a specific group while it could be innocuous if used in conjunction with some other group. This was a major issue that hindered scientists in designing a usable hate speech detection system.

Bag-of-words was tried to solve this issue but due to the presence of high frequency of offensive words in a tweet it misclassified it as hate speech even though it was not targeted against any group. Subtle linguistic variations were used for differentiations purposes but contextual differences between the phrases can alter the sentences meaning. The placement of verbs and nouns in a sentence has also been made use of while identifying the intensity of hate speech. The authors state that other supervised approaches have not been successful in their attempt of identification of hate speech. Apart from linguistic features, features such as ethnicity or gender may help us.

The researchers first extracted hate speech phrases from Hatebase.org and scraped sample tweets from the Twitter which contained the hate speech phrases. The tweets were then asked to be labeled by workers into three categories; hate speech, offensive and neither according to a definition provided by the researchers. Each tweet was analyzed by three workers and the majority decision was assigned to each tweet. The researchers used a stricter criterion for hate speech so only 5 percent of the entire tweets were labeled as hate speech by majority of the people while only 1.3 percent were labeled after a majority decision.

The initial step was to test the data with logistic regression with L1 regularization. The researchers also tested several other classifiers like naïve Bayes, decision trees and random forests. However, they found Logistic Regression and Linear SVM to perform the best. The final model used L2 regularization and was trained on the entire dataset.

Of all the model, the best performing model was 0.91 precise, had a recall of 0.90 and an F1 score of 0.90. However, 40 percent of the hate speech was misclassified. Most of the offensive and neither offensive nor hate speech were correctly classified with 0.91 percent and 0.95 percent accuracy. Around 0.31 percent of hate speech was classified as offensive by the model. The reason behind misclassification was that the true hate speech identified by the model contained racist and homophobic slurs while the true hate speech classified by the system was actually less hateful and mislabeled by the coders indicating that the coders might not have been aware of the context. Another reason for misclassification of hate speech was that some hate speech tweets did not contain the words usually associated with hate speech and did not contain any slurs or curse words. Other rarer forms of hate speech that did not occur often were also misclassified.

Highlighting the flaws of the previous work, the authors explains that the definition of offensive language was very broad, but this issue was by overcome by them by having a stricter definition of hate speech. The analysis of the misclassified tweets revealed that they contained song lyrics and slurs used in everyday language by some individuals so the extent of the prevalence of hate speech was overestimated by the classifier. Tweets classified as neither generally contained positive sentiments but had words from the Hatebase lexicon. It was also revealed that the classifier was able to classify some hate speech tweets correctly that were wrongly labeled by the coders.

Distinguishing between hate speech and offensive language is vital due to the social and moral implications. Lexical methods and identification of certain terms can be useful for differentiating. Contextual background of the tweet could be analyzed to identify if they were indeed hateful or just offensive. With human having their own biases about what is considered hateful and offensive, it is essential to identify and rectify our social biases.