NLA Project: Hate Speech Detection

Ravsimar Singh(20161117)
Vaibhav Bajaj(20161179)

Our project is *Hate Speech Detection* on social media. Detecting hate speech is a difficult task, since there is very subtle difference between a data sample that is ‘offensive’ and something that can be classified as ‘hate speech’.
Hate speech is generally considered as a negative sentiment and emotions towards a particular community, whether it is a minority or a simply different based on ethnicity, gender or race etc. Social media has ample usage of curse words, but not all data containing curse words can be considered as hate speech.

Overview

We propose to compare various models already available and build our own model for hate speech detection. A thorough comparison of results between various methods, and various approaches to better analyse a good approach towards hate speech detection.

Data

Since annotated and labelled data is not available for Gab, Reddit, and Voat, we will use the available twitter data, and train the models. Multiple data sources exist for hate speech on Twitter.
For testing, we may use data from the above sites and analyse the results. However, more concrete results would be found if labelled data from the above sites could be found.

Preprocessing

Preprocessing of the data using known NLP techniques. Some of the methods we are planning to use:
Given a tweet, the following preprocessing procedure:

- Removing excess useless characters (e.g. ‘hellooooooo’ to ‘hello’)
- Stemming, to reduce word inflections.
- Splitting text into tokens.
- Removing URLs, mentions(@s) (more relevant to twitter, other social media may have different way of mentioning links)
- Removing stop words (except maybe for negative words such as ‘not’) to better identify the words used in hate speech.

Comparing various approaches such as:

- Char n-grams, word n-grams
- TF-IDF (Term Frequency - Inverse Document Frequency)

Model

Word embeddings have been shown to be more practical than one-hot vector encodings. We will compare and use various methods such as:

- GLoVE
- Bag of words
- Word2Vec

Comparing various machine learning models:

- SVMs
- Logistic regression
- CNNs
- LSTMs

Reference Papers

- Hate Speech Detection: A Solved Problem?The Challenging Case of Long Tail on Twitter
- Hate Speech Detection Using Natural Language Processing Techniques
- Improving Hate Speech Detection with Deep Learning Ensembles
- Deep Learning for Hate Speech Detection in Tweets
- Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter