INFO 7390 – ADVANCES IN DATA SCIENCE AND ARCHITECTURE

Hate Speech Detection on Tweets

Multiclass classification of tweets using different modelling techniques

Group No. 24

Shika Shyam	002194543
Deepika Balasubramaniam	002194564

Goals and Objectives:

In this project our goal is to classify tweets as hate speech or related and non-hate speech. In a more specific sense, we will be classifying tweets into one of four categories - Normal, Spam, Abusive and Hateful.
Below is the objective of this project detailed:

- We will be cleaning the tweets by removing special characters and non-alphanumeric characters, and then using NLP Text preprocessing techniques.
- For text preprocessing we will be using Stemming or Lemmatization, and see how these perform against each other.
- We will be making use of 4 different techniques to do feature engineering - i.e. converting the tweets into feature sets using embedding techniques, namely - Count Vectorizer, Term Frequency - Inverse Document Frequency (TF-IDF), Word to Vector (W2V), Global Vectors for Word Representation (GLOVE).
- We will then be training Gaussian Naive Bayes, Multinomial Naive Bayes, Support Vector Machine and Long Short-Term Memory (LSTM) models on each of these ablation settings.
- We will choose the best performing combination of preprocessing method + embedding method, and then use that setting on a hyperparameter tuned version of NB, SVM and LSTM to better our model performance even further.
- Since ultimately, this is a classification task, we will take a look at how each of these settings and each of these models performed with respect to three main metrics - Precision, Recall and F-1 Score.
- We will compare each of our models (with each ablation setting) 27 different settings in total, and identify our best models for our project goal of detecting and classifying hate speech.

Impact of the solution:

With the ever-expanding social media world, it has become increasingly important to be able to identify and flag abusive content ahead of time. As social media grows a tweet has the power to reach billions of people within seconds and affect their lives adversely with the content of the speech. For this very reason, our project aims to model and classify tweets into one of four classes namely – Abusive, Hateful, Spam and Normal.

We are then trying different ablation settings (Text preprocessing and Embedding techniques) with three different models – Naïve Bayes, Support Vector Machines and LSTMs and their respective hyperparameter tuned versions to pick a model with high recall and agreeably compromising on Precision and F-1 Score.

The reason for choosing high recall as our model evaluation metric is, we would rather have our model correctly classify an abusive tweet as abusive, and incorrectly classify a normal tweet as abusive THAN miss classifying an abusive tweet altogether.