**Midpoint Report for Twitter Sentiment Analysis**

**Project Objective:**

Sentiment analysis comes under natural language processing and it is an automated process that can be used to understand the “sense” or the “tone” of the sample text. The sentiment in a text can be positive, negative, or neutral.

**Project Approach:**

1. Get the data - Train Data

The data will be retrieved from Kaggle for our project research. Using this data we are planning to train our project model.

1. Connection with Twitter (APIs)

To fetch the data from Twitter we will be using twitter API by registering on Twitter app along with getting all the necessary keys and tokens.Tweepy is the python client for the official Twitter API.

1. Data categorization

The data will be categorized into groups according to the needs.

Example: The user needs to know tweets sentiments on the movie “Harry Potter”. All the tweets related to the Harry Potter movie will be grouped together and then the data will be passed to the model for further exploration.

1. Remove unwanted words from Twitter Data

Unwanted words are the words, which we do not require in our analyzing process. For example, “and”, “for”, “etc.”. These words can be removed from the data and then proceed with the analysis.

1. Apply classification algo.

To classify data we are planning to use 3 classifiers. (We are planning to use deep learning instead of one of the algorithms, but that is to be decided.)

1. Naive Bayes
2. Support Vector Machine
3. XGBoost
4. Visualize and Compare

For visualization of data we are planning to create graphs and input the graphs with the performance of each algorithm we have taken for classification.

**Team Structure:**

Larrisa Carvalho

Siri Koduru

Rucha Kulkarni

We are a team of 3 so we are planning to work on each classifier algorithm. Data Cleaning, Connection with API, Data Categorization we will do together by collaborating on zoom.

**Progress:**

1. Get the data - Train Data

We have retrieved data from Kaggle, and trained our model based on this Kaggle data.

1. Connection with Twitter (APIs)

We are working on to Twitter API (we will use “Tweepy” mostly.), The work is in progress and we are planning to get it done by Sunday July 25th.

1. Data categorization

We are working on the list of negative words and positive words and it is almost completed, which will help to categorise data into positive or negative sentiments.

1. Remove unwanted words from Training Data

We completed this step and tried removing unwanted words from the data which we collected from the Kaggle. For example, if a tweet contains “and”, “for”, “etc.”, these words are removed from the data and then we will proceed with the analysis.

We are working in collaboration, so we are using Google Colab for now. We will create a GitHub repository and will push our code there.

**Date / time of scheduled midpoint meeting with Professor**

Thursday, 22nd July 2021

11:10 am - 11:30 am