**CS583 – Research Project**

*Rishabh Mehta () & Pramodh Acharya (pachar7)*

This project aims to analyze the sentiments of tweets about the presidential candidates and classify them as positive, neutral, or negative sentiments. The data is from Twitter during a presidential debate between Obama and Romney. The resultant classification is a probable measure of the election outcome.

We created a sentiment analyzer to classify tweets in text format as positive, negative, or neutral. We followed the following set of steps to achieve the same.

**The Dataset**

The training data is a set of labeled twitter data with the tweet text, date, and timestamp of the tweet.

The test data is a set of unlabeled data with each row containing a tweet id and the text.

**Text Preprocessing**

Several steps of standard processing for Twitter data was performed on each row of raw data. Following are the sequential steps of preprocessing used.

* **Cleaning Raw Tweets**
* *Lowering Case:* Lowering the case of all the words helps to reduce the dimensions by decreasing the size of the vocabulary.
* *Removal of Html tags:* Removed anything enclosing a ‘<’ and ‘>’ inclusive. This removes any Html DOM elements as these are not useful in training the model.
* *Removal of Hyperlinks:* We remove URLs by searching for any part of the sentence starting with an ‘http’ and removing the complete URL string.
* *Removal of mentions:* We remove mentions from tweets as they are not useful. Any word or part of the sentence starting with ‘@’ is removed completely.
* *Removal of Special Characters:* All remaining special characters like ‘#’ or punctuations like ‘,’, ‘.’ are removed.
* *Removal of Numbers:* Numbers in most cases do not provide any meaningful opinion on the sentiment and hence we removed it.
* *Removal of Stopwords:* Stopwords were removed with help from nltk.corpus English library’s ‘stopwords’ as they don’t provide valuable information towards the analysis or sometimes skew the model.
* **Tokenization**

This is the process of splitting the processed raw tweets (sentences) into smaller chunks such as a list of words.

We initially tried the spacy module to tokenize and lemmatize the sentences, but splitting it manually (without any library) and using wordnet to lemmatize provided better results.

A sub-step while tokenizing is splitting any joined words, which is mostly the case in hashtags. This is done by using a utility function we wrote to split any possible word to the least dimension possible.

* **Lemmatization**

This is the process of fetching the lemma’s or the simplest form or root form of the words as present in a dictionary. We lemmatize all the verbs and nouns/plurals. This reduces the dimensionality of the dataset by a large value.

**Models And Techniques Used**

**Experimental Results**

**Conclusion**