**PROJECT REPORT CS583**

**PROJECT TITLE:** Sentiment Analysis for Obama and Romney’s tweets.

**Input:** Tweets of both the politicians.

**Output:** Identify the tweet into either of three classes :

* Positive (1)
* Neutral (0)
* Negative (-1)

**Type of Model:**  Since, we are already given labelled datasets of both the politicians. We have performed supervised learning on the models.

**Methodologies used to train the Classifier:**

* **Logistic Regression 🡪 (Accuracy: 60% )**
* **SGD Classifier🡪 (Accuracy: 53% )**
* **Naïve Bayes🡪 (Accuracy: 52% )**
* **Random Forest🡪 (Accuracy: 55%)**
* **XgBoost🡪 (Accuracy: 55%)**
* **SVM🡪 (Accuracy: 55.5%)**
* **Gradient Boost🡪 (Accuracy: 50%)**
* **Grid Search🡪 (Accuracy: 59.6%)**
* **Deep Learning Models:**
  + **Feed Forward Sequential Network🡪 (Accuracy: 41%)**
  + **Bi-directional LSTM🡪 (Accuracy: 55%)**
  + **Transfer Learning Model🡪 (Accuracy: )**

**Data Pre-Processing:**

* **Data Cleaning**
  + **All letters converted to lower case**
  + **Remove HTML tags**
  + **Split Hashtags**
  + **Remove URLS, punctuation , emojis, hashtags, mentions.**
  + **Remove digits**
  + **Remove extra spaces**
  + **Remove stopwords**
* **Lemmatization**
* **Normalization**

Number of Input features to the model was 500 after the Data Cleaning and pre-processing.

**Conclusion:**

We split the input data into 8:2 ratio for training and testing.

We got the best accuracy with Logistic Regression with sufficient consistency near 60%.

Deep learning models could not deliver high accuracy for the dataset provided, with the Bi-directional LSTM giving the best accuracy, whereas Normal Feed Forward Neural Network giving accuracy as low as 41%.

All other methods gave accuracy in the range of 50-60 %, with the test data.