PROJECT REPORT

CS583 Data mining and Text mining

Twitter Sentiment Analysis for Obama and Romney’s tweets

By Giri Madhav and Kumar chandrasekhar

Input: Tweets of both the politicians.

Output: Identify the tweet into either of three classes :

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

Task : Task is to build multi class classifier that classifies obama and romney’s tweets

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

Methodology:

1. Data Pre-Processing
2. Modelling
3. Performance Evaluation

Data Pre-Processing:

- **Data Cleaning**

  Data has a lot of unnecessary noise that needs to be cleaned to avoid models being under performed**.** Data has emojis, hashtags, mentions, encosings, URLs, stop words, punctuations and more unnecessary noise.

  Data cleaning techniques used are:

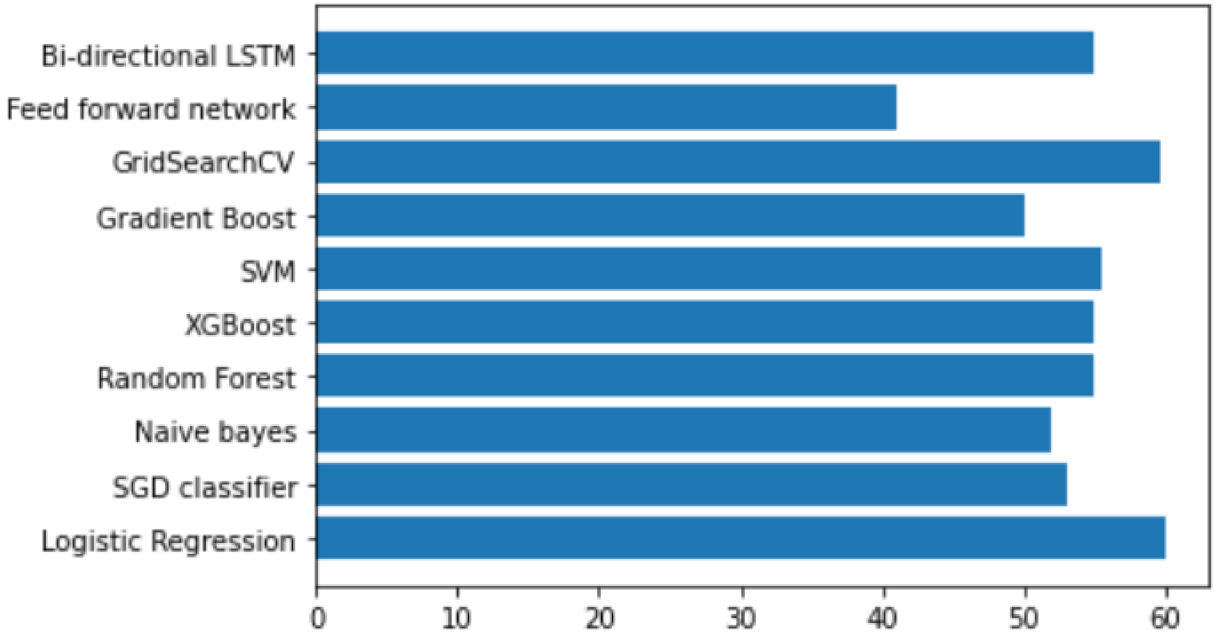
  - All letters converted to lowercase
  - Expand Contractions
  - Remove HTML tags
  - Split Hashtags
  - Remove URLS, punctuations , emojis, hashtags, mentions.
  - Remove digits and words containing any digits
  - Remove extra spaces
  - Remove stopwords - Removed most frequent and rare words in the data

    - Remove unnecessary columns and null values
- Tokenization
- Lemmatization
- Feature Extraction - TfidfVectorizer is used to convert text data to numeric vectors.

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 CV (Accuracy: 59.6%)
- Deep Learning Models:
    - Feed Forward Sequential Network (Accuracy: 41%)
    - Bi-directional LSTM (Accuracy: 55%)

Results Comparison:



Number of Input features to the model was 500 after the Data Cleaning and preprocessing. Both Obama and Romney data joined together and applied modeling. Machine learning models gave better performance results than deep learning models.

Conclusion:

- We split the input data into 8:2 ratio for training and testing.
- sklearn cross_validate function is used to perform 5 cross validation on the train data.
- 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.

While implementing this project, we learnt different text data handling techniques, feature extraction techniques, implementing different supervised models and deep learning models for text data.

In the course, we learnt different supervised, unsupervised and semi-supervised algorithms. We learnt new topics like life long continual learning and aspect based opinion mining .