PROJECT REPORT

Maximum Accuracy Achieved

- 66.983% for the Binary Classification Task.

Achieving the result for the Binary Classification Task

- The data source used for this project contains 3 files (train2.tsv, val2.tsv, test2.tsv).
- These files have been converted to their csv formats for better handling.
- Converting 6 Labels to 2 Labels:
  - "mostly-true","half-true","true" → "TRUE"
  - "barely-true", "false", "pants-fire" → "FALSE"
- Used only the "Statements" and "Label" columns for the prediction.
- File Descriptions:
  - DataExtraction.py → This file does all the preprocessing of the data. Some tasks like tokenizing, stemming have been covered in this file.
  - FeatureExtraction.py → This file extracts the features with the help of sci-kit python libraries. I have used simple Bag-Of-Words method for feature extraction.
  - Classifiers.py → Extracted features in the previous file are fed into the classifiers to predict the fake news. I have used Naïve Bayes and Logistic Regression classifiers from sklearn. Comparing the f1 scores and confusion matrix to select the best classifier. Creating a final model for prediction.
  - Prediction.py → The best performing classifier was **Naïve Bayes using BOW** with 66.983% accuracy. The file takes the news as an input from the user and classify it as "TRUE" or "FALSE" along with probability of truth.

Ideas Tried out:

- Approached a PhD. Scholar and students under her who have worked upon Fake News Detection. They had used CNN and embedded Systems. I didn't have any experience over it so it was difficult for me to continue with that in the timeframe.
- I studied about different feature extraction algorithms and got comfortable with BOW(Bag of Words).
- I performed Exploratory Data Analysis on {(Speaker- Label),(Party- Label) and (Subject-Label)} to find relation between different attributes.

Ideas To be Tried Out:

- Six-way classification task.
- Extracting features from "Justification" column.
- Choosing more Classifiers and NLP algorithms for better performance.
- Training on more data.

CITATIONS

- (Loper, 2009)
- (Oliphant, 2006)
- (https://github.com/python/cpython/tree/3.7/Lib/pickle.py, n.d.)
- (Fabian Pedregosa, 2011)
- (https://github.com/nishitpatel01/Fake_News_Detection)
- (https://github.com/nishitpatel01/Fake_News_Detection/blob/master/DataPrep.py)
- (McKinney, 2010)
- (https://github.com/nishitpatel01/Fake_News_Detection/blob/master/FeatureSelection.py)
- (Hunter, 2007)
- (https://github.com/nishitpatel01/Fake_News_Detection/blob/master/classifier.py)
- ({TensorFlow}: Large-Scale Machine Learning on Heterogeneous Systems)
- (https://github.com/nishitpatel01/Fake_News_Detection/blob/master/final_model.sav)
- (https://github.com/nishitpatel01/Fake_News_Detection/blob/master/prediction.py)