SLP REPORT GROUP 22

Fake News Detector Using Natural Language Processing

TEAM MEMBERS

Aakash Mahadevan (16UCC001)

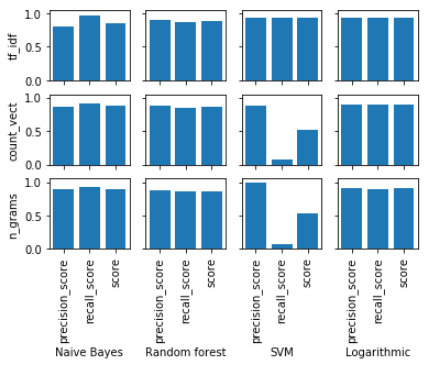
Anushka Prakash (16UCS223)

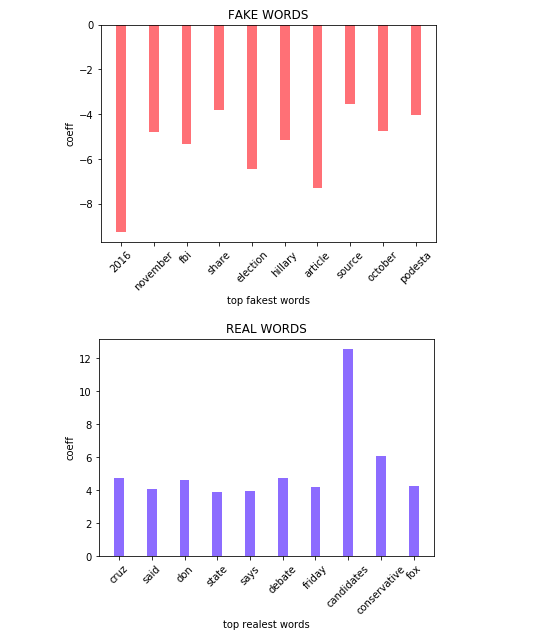
INTRODUCTION

The purpose of this project is to successfully determine whether a given news is fake or not by

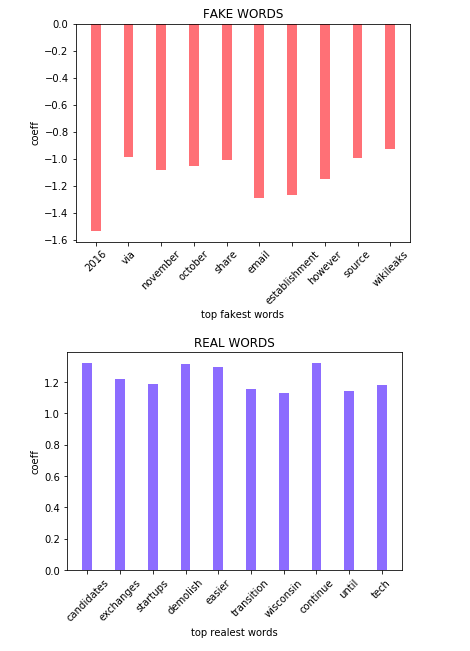
Employing the principles of Machine Learning and Natural Language Processing. We use 3 types of vectorizers and 4 types of classification models with each one, and then finally compare the results.

SCREENSHOTS

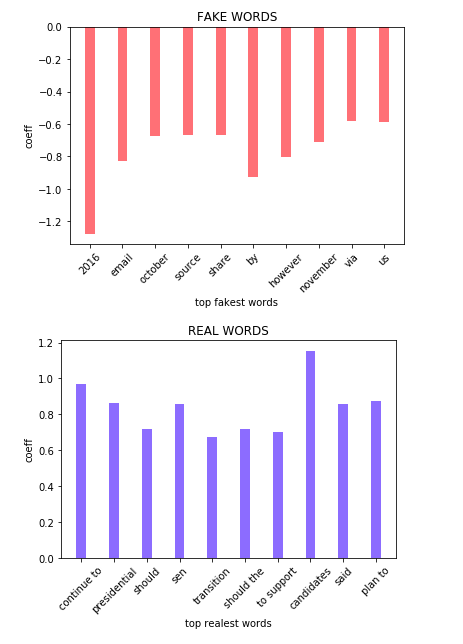




***logreg using tfidf vectorizer***



***logreg using count vectorizer***



***logreg using n-grams***

INSTALLATION PROCESS

We have used the Jupyter Notebook on Anaconda Navigator to write and execute the code. The code is in both .ipynb and .py format. To run .py file, libraries like pandas, numpy, scikit learn, matplotlib need to be preinstalled on the system. To run .ipynb Anaconda Navigator is required.

TECHNOLOGIES

Anaconda Navigator

Jupyter Notebook

DESCRIPTION

**The Data Set**

The dataset we are going to use was released by Kaggle, comprising of articles published during the 2016 US Presidential elections. We also used articles from AllSlides.

The final dataset contains nearly 15000 articles, with their headlines and bodies separated out, and labeled as well (fake or real). It is in .csv format.

**Libraries**

1. Numpy

2. Scikit-learn

3. Matplotlib

4. Pandas

**Vectorizers**

We use 3 types of vectorizers:

1. Tf-idf vectorizer
2. Count vectorizer
3. N-grams

**Classifiers**

We use 4 types of classifiers:

1. Multinomial Naive Bayes
2. Random Forests
3. Support Vector Machine Radial Basis Function classifier
4. Logistic Regression

After importing the dataset, we cleaned it, removing any NaNs and stray characters. We vectorized the data in each of the three ways, and then trained each one of them using all the four classifiers.

We took the score, precision score and recall score for every combination and compared them.

We used confusion matrices and bar charts to visualize our results.

Logistic Regression consistently turned out to be the best one. We hit the 93% accuracy mark.

Defining fake news with simple bag-of-words or TF-IDF vectors is an oversimplified approach. Especially with a multilingual dataset full of noisy tokens. Taking a look at what actually the classifiers have learnt reveals that it may not be give desired results with other types of news articles. Also, there may be certain trends in the data that our classifiers might have missed. There is always a possibility to increase to increase the accuracy of models, but we got a chance to actually see how different vectorizers and classifiers work as compared to each other.