

PROJECT

Fake News Detection

Submitted by:

Robinson P

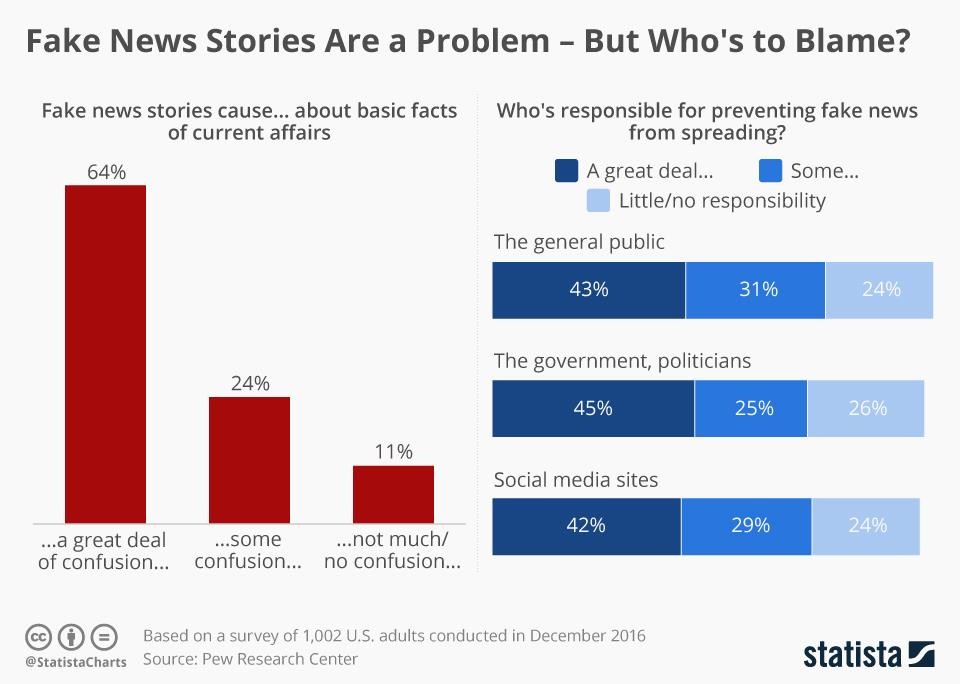
**ACKNOWLEDGMENT**

This project consumed huge amount of work, research and dedication. Still, implementation would not have been possible if we did not have support of many individuals from flip robo technologies, therefore Iwould like to extend our sincere gratitude to all of them.

My thanks and appreciations also go to my datatrained for providing me oppurtunities to carry out project

**INTRODUCTION**

* Business Problem Framing
* As an increasing amount of our lives is spent interacting online over the internet, more and more people tend to seek and consume news from social media, news agency homepages, search engines. On the other hand, it enables the proliferation of “fake news", i.e., low quality news with intentionally false information. Popular social media platforms such as Facebook, twitter have proven to be an effective means of channels for spreading these false news due to their wide reach and the speed in which information is spread.
* The plague of fake news not only poses serious threats to the integrity of journalism, but has also created turmoil in the political world. The term ‘fake news’ became common parlance for the issue, particularly to describe factually incorrect and misleading articles published in social media feeds, news blogs and online newspapers, mostly for the purpose of making money through page views. The extensive spread of fake news has the potential for extremely negative impacts on individuals and society. Fake news is intentionally written to mislead readers to believe false information, which makes it difficult and nontrivial to detect based on news content. Therefore, the issue of fake new detection is both challenging and crucial. Hence, we decide to take up this challenge and find solution in an efficient way.



* Conceptual Background of the Domain Problem

Describe The Given data set consist of election campaing comments the idea is to detect the fake news detection for which classification approach and TFIDF vectorizers can be used

* Motivation for the Problem Undertaken

Describe your objective behind to make this project, this domain and what is the motivation behind.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

**Evaluation Metrics Selection**

During the modeling process, we choose multiple different evaluation metrics to evaluate the performance of models based on the nature of our data:

* Recall
* F Score

**Basic Model Comparison**

Using Multinomial Naive Bayes as our baseline model, we first used k-fold cross validation and compared the performance of the followingi three models without any hyperparameter tuning: Logistic Regression, and Linear SVC. Word2vec

* Data Sources and their formats

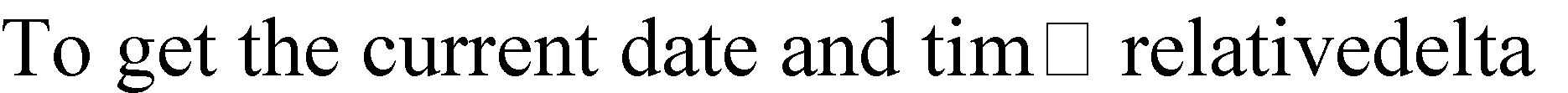
The data used for this project was drawn from fliprobo technologies, the data was in csv format, with null values the original dataset had around 23000 rows

* Data Preprocessing Done

Data had several of uncleaned data and that was removed at the initial stage before exporting it into the notebook .columns of headlines and news were combined to obtain detail news and to avoid data loss the null values were replaced with mode values before deployment

* Data Inputs- Logic- Output Relationships

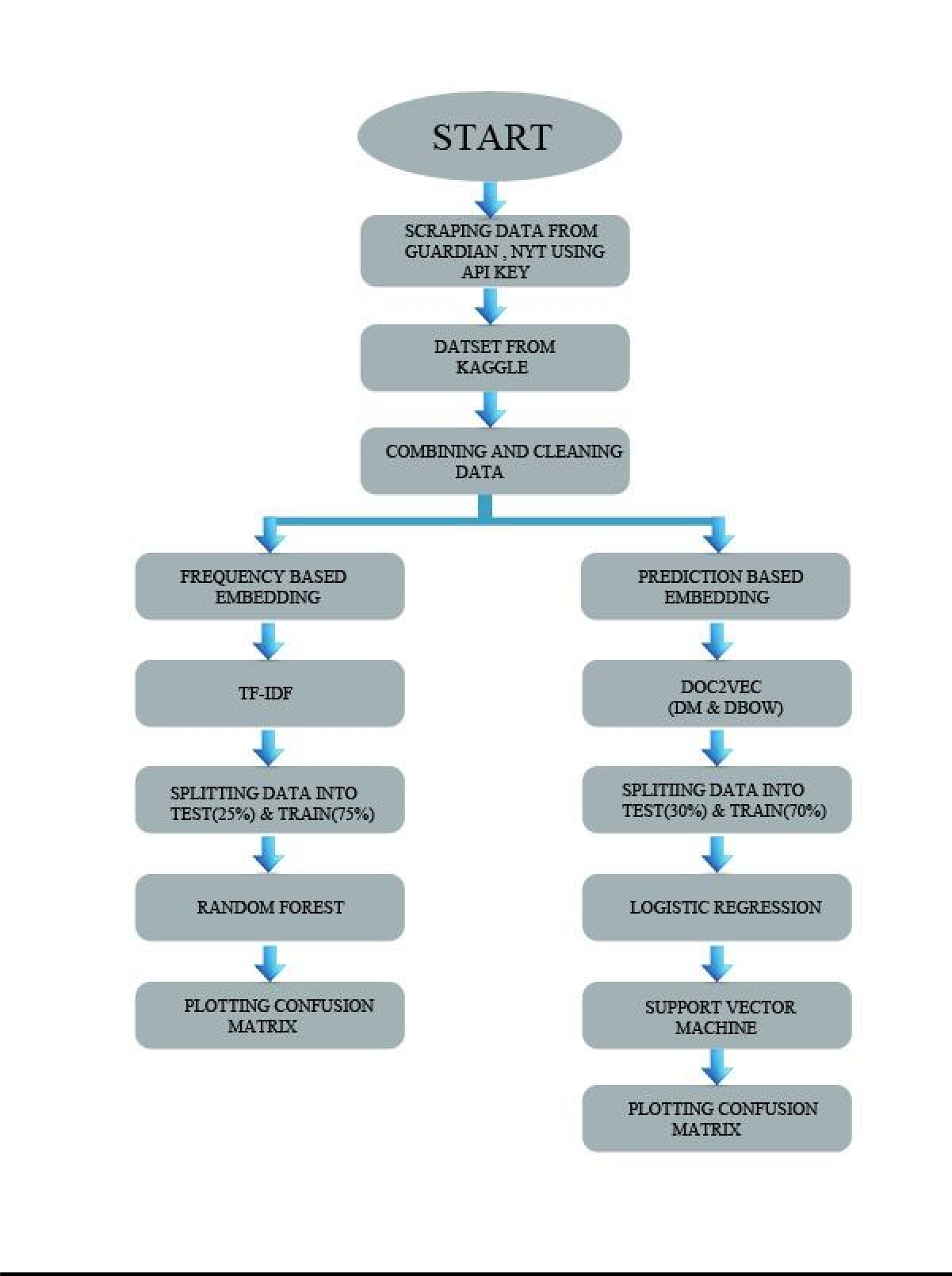
TFIDF vectorizers ,wordembedding were adopted the dataset is crosschecked to get more weightage words

* Hardware and Software Requirements and Tools Used
* BeautifulSoup – a python library to for pulling data out of HTML and XML files. To scrape or extract all the data from a particular website
* pymongo - To establish connection between python code and mongo database. It is used to retrieve the data from MongoDB and store locally.
* json - it is used to retrieve json data from the website
* requests - to get the requested URL for fetching the contents from the URL.
* time and datetime - - to calculate

the absolute year, months, days, time between two given dates.

* pandas - for reading the csv and converting it to dataframe
* numpy - for array computation
* re- for regular expression
* sklearn- scikit learn for shuffling the data, performing cross validation, performance metrics
* seaborn & matplotlib - for plotting graphs
* nltk (Natural Language Toolkit) –suite of libraries and programs for symbolic and statistical natural language processing in python.
* Gensims-python library for topic modelling, document indexing and similarity retrieval for large corpora.
* Tqdm-shows the progress of any iteration of epochs.
* SVM –python library used for multiclass classification

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods) 
* Testing of Identified Approaches (Algorithms)

**Word Embedding:-**

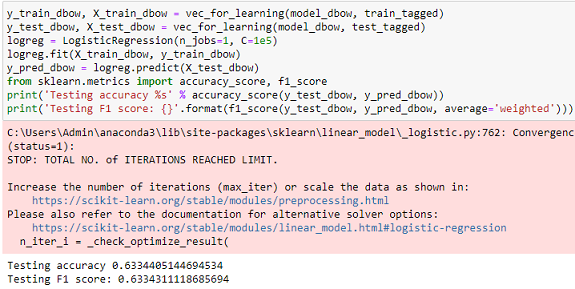
1. Frequency based Embedding
2. Prediction based Embedding

Logistic Regression

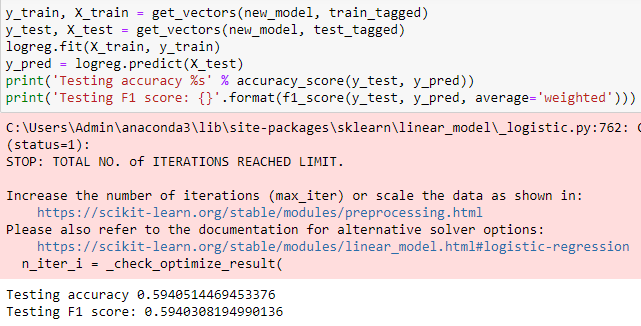
SVC

* Run and Evaluate selected models

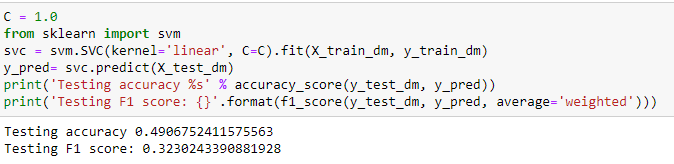
Logistic Regression



Logistic\_Regression with Doc2vec



Svc

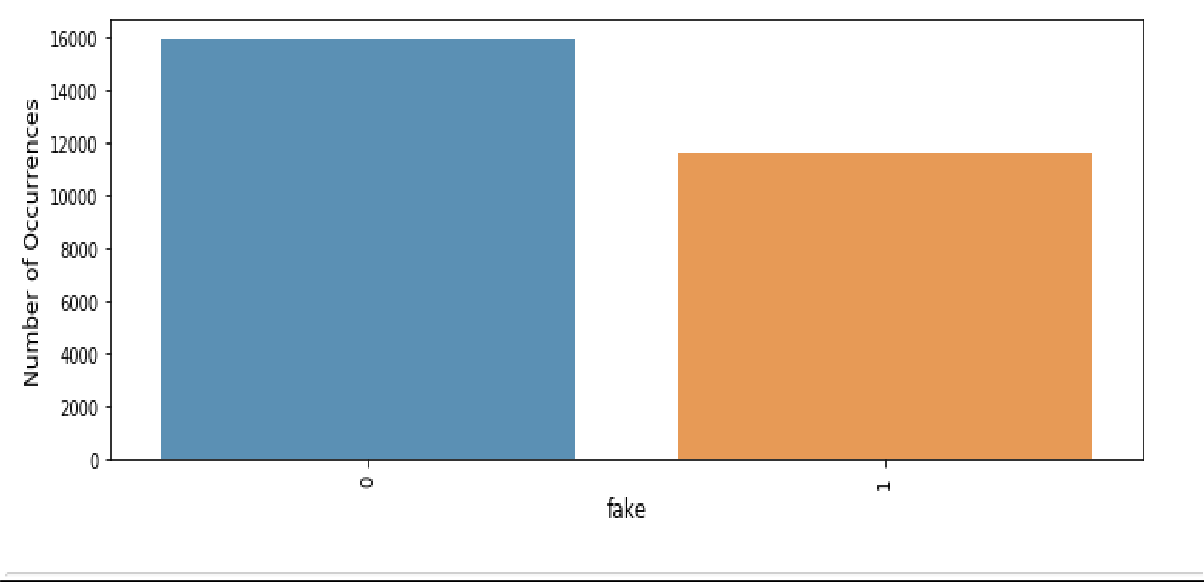


The logistic model performed better

* Key Metrics for success in solving problem under consideration
* Confusion metrics –used for identifying true positives and negatives
* Recall-
* F1 Score-

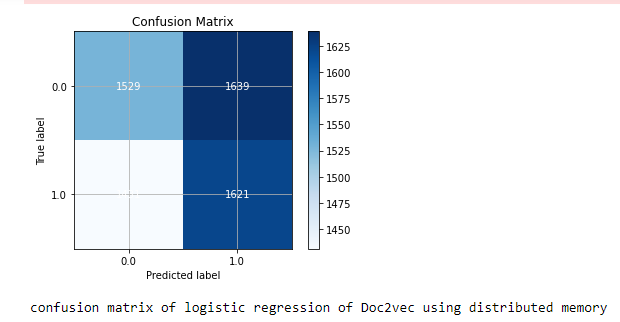
.

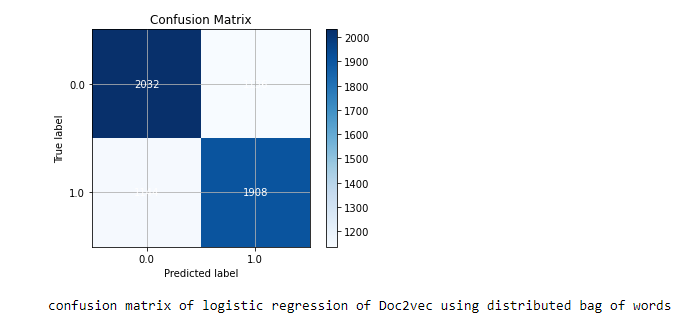
* Visualizations

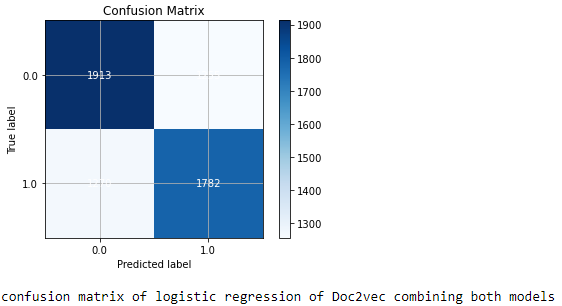


* Interpretation of the Results

Based on the observations made







The model of logistic regression of doc2vec using distributed memory performs better

**CONCLUSION**

* Key Findings and Conclusions of the Study

The data set had a lot of unwanted data which were cleaned and the SVC,logistic Regression Doc2Word vectorizer models were deployed in which logistic regression of doc2vec using distributed memory performs better. Decision tree and random forest classifier models werenot able to perform because of issue in PC performance.To efficiently execute this task.The optimal parameters for count vectorizer are no lowercasing, two-word phrases not single words, and to only use words that appear at least three times in the corpus.

            This model’s maximum accuracy score is 63% for logistic model with distributed memory

* Learning Outcomes of the Study in respect of Data Science

Understood tokenizing, lemmatizing application of F1score Recall and models used for NLP classification

* Limitations of this work and Scope for Future Work

Different models applying LSTM and neural networks can be tested

Accuracy was satisfactory