INT248 PROJECT

CA-2

LOVELY PROFESSIONAL UNIVERSITY

NAME: Pratik **SEC**: KE176

REG: 11808492 **ROLL**: 27

SUBMITTED TO: MD. IMRAN HUSSAIN SIR

INTRODUCTOIN

Fake news, defined as a made-up story with an intention to deceive, has been widely cited as a contributing factor to the outcome of the 2016 United States presidential election. While Mark Zuckerberg, Facebook's CEO, made a public statement denying that Facebook had an effect on the outcome of the election, Facebook and other online media outlets have begun to develop strategies for identifying fake news and mitigating its spread. Zuckerberg admitted identifying fake news is difficult, writing," This is an area where I believe we must proceed very carefully though. Identifying the truth is complicated."

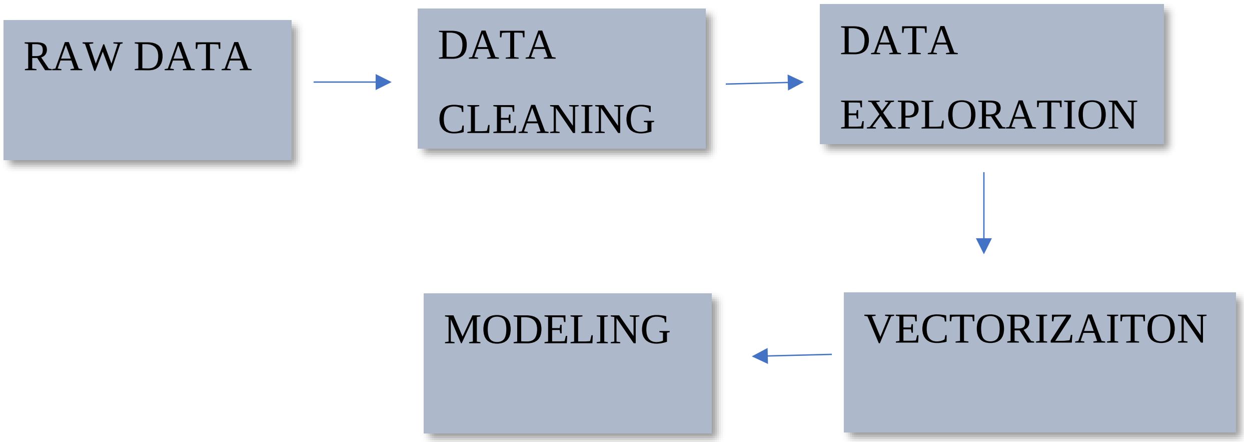
Fake news is increasingly becoming a menace to our society. It is typically generated for commercial interests to attract viewers and collect advertising revenue. However, people and groups with potentially malicious agendas have been known to initiate fake news in order to influence events and policies around the world. It is also believed that circulation of fake news had material impact on the outcome of the 2016 US Presidential Election.

Dataset

The datasets used for this project were drawn from Kaggle. The dataset has about 44898 rows of data from various articles on the internet. We had to do pre-processing of the data, in order to train our models.
A full training dataset has the following attributes:

1. title: the title of a news article
2. subject: type of the news article
3. text: the text of the article; incomplete in some cases
4.date: date of the news

Proposed Architecture



1. **Logistic Regression**

   Logistic regression is a fundamental method initially formulated by David Cox in 1958 that builds a logistic model (also known as the logit model). Its most significant advantage is that it can be used both for classification and class probability estimation, because it is tied with logistic data distribution. It takes a linear combination of features and applies to them a nonlinear sigmoidal function. In the basic version of logistic regression, the output variable is binary, however, it can be extended into multiple classes (then it is called multinomial logistic regression). The binary logistic model classifies specimen into two classes, whereas the multinomial logistic model extends this to an arbitrary number of classes without ordering them.

2. **Decision Tree Classifier**

   The principle of splitting criteria is behind the intelligence of any decision tree classifier. Decision trees are presented similar to a flow chart, with a tree structure wherein instances are classified according to their feature values. A node in a decision tree represents an instance, outcomes of the test represented by branch, and the leaf node epitomized the class label. Three variations of decision trees are explored here, viz., Best First Decision Tree (BFTree), ForestPA, and SysFor because of the fast model build time and processing speed.

3. Random Forest

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms. The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.

4. LSTM

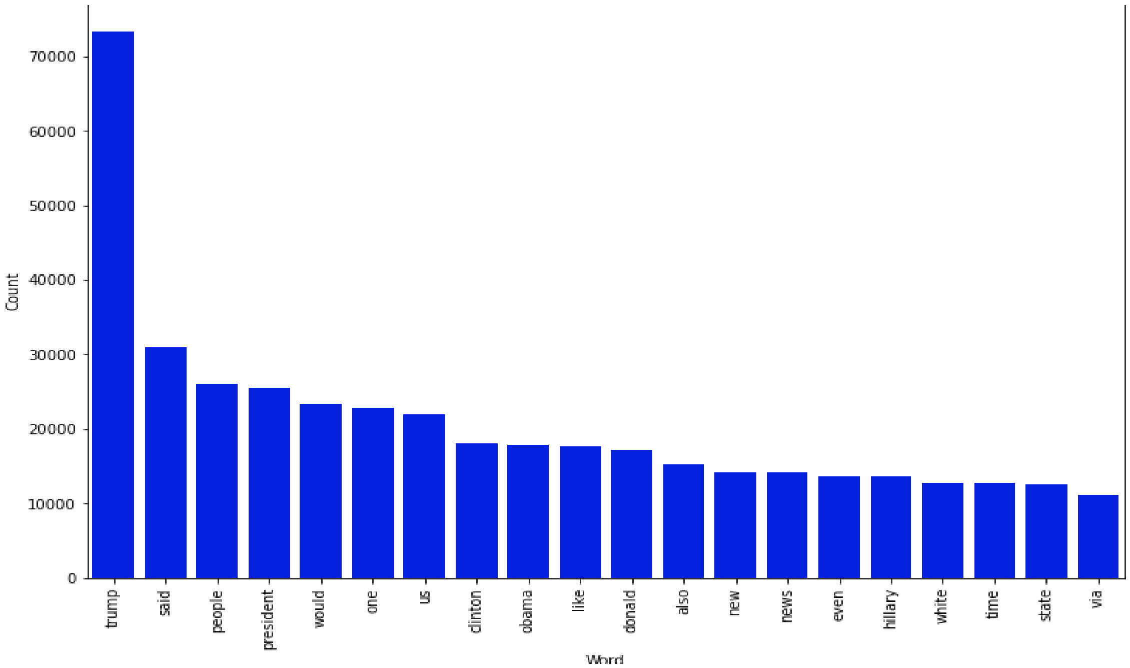
Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can process not only single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegmented, connected handwriting recognition, speech recognition and anomaly detection in network traffic or IDSs (intrusion detection systems).

A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three *gates* regulate the flow of information into and out of the cell.

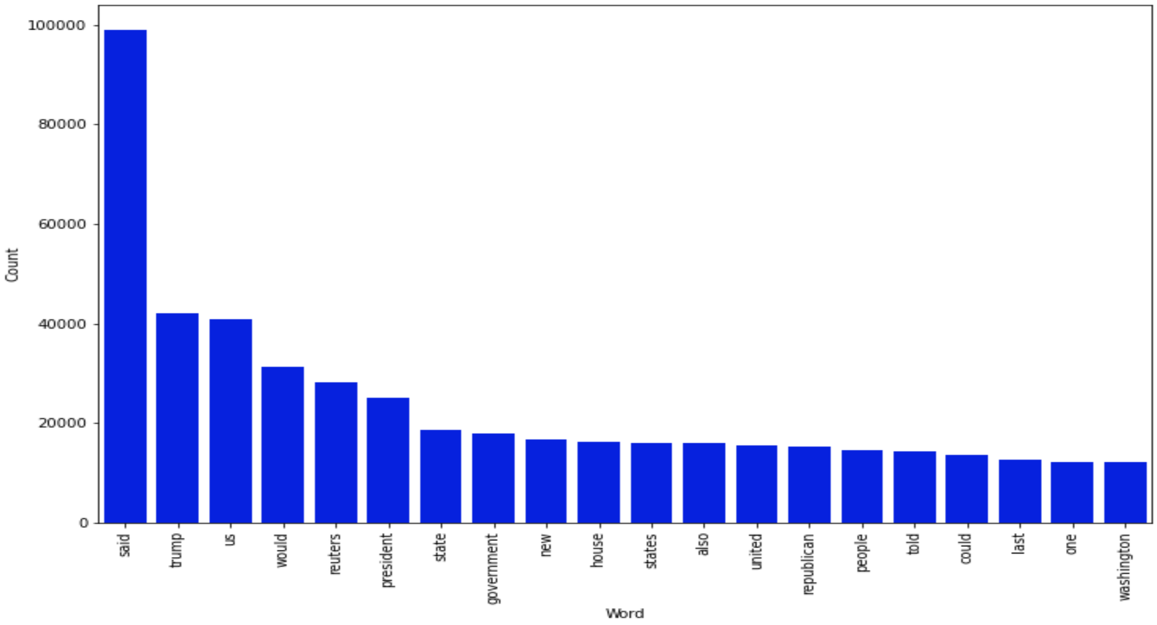
Result And Experiment Analysis

Model	Accuracy	Precision	Recall
Logistic Regression	0.989	0.987	0.991
Decision Tree	0.997	0.998	0.995
Random Foreset	0.988	0.956	0.992
LSTM	1.00	1.00	1.00

Frequency of most common words in fake news.



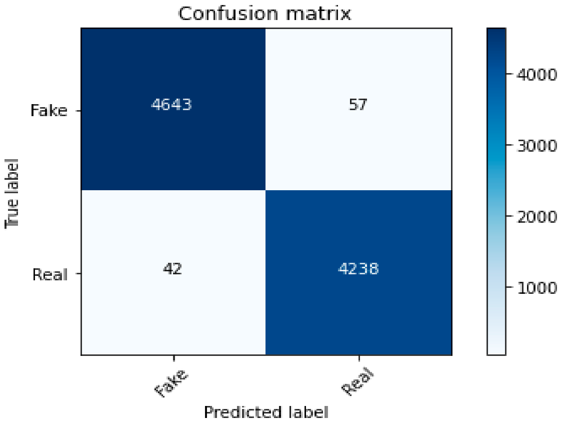
Frequency of most common words in real news.



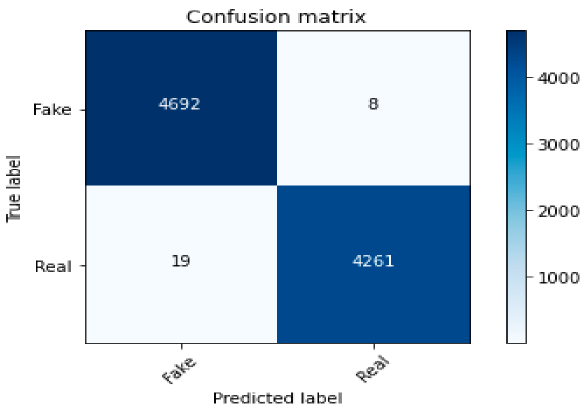
The LSTM comes out to be the best with highest values in accuracy,precision and recall.

Output Screenshots

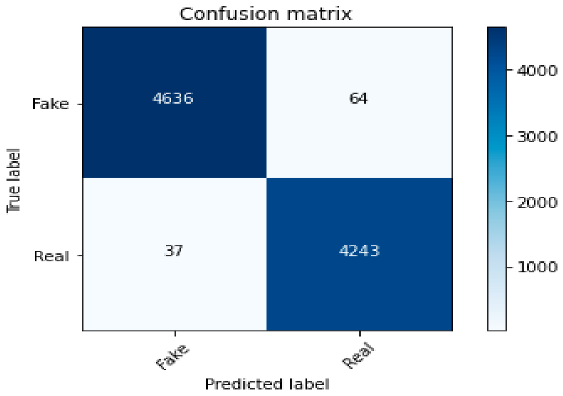
Logistic Regression



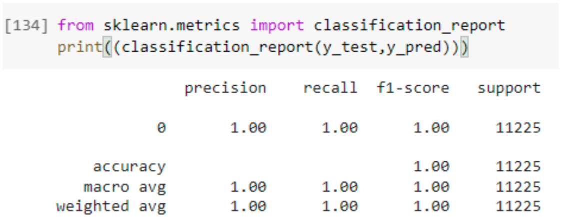
Decision Tree



Random Forest



LSTM



```
[134] from sklearn.metrics import classification_report
      print((classification_report(y_test,y_pred)))
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     11225

    accuracy                           1.00     11225
   macro avg       1.00      1.00      1.00     11225
weighted avg       1.00      1.00      1.00     11225
```

Conclusion and Future Scope

A complete, production-quality classifier will incorporate many different features beyond the vectors corresponding to the words in the text. For fake news detection, we can add as features the source of the news, including any associated URLs, the topic (e.g., science, politics, sports, etc.), publishing medium (blog, print, social media), country or geographic region of origin, publication year, as well as linguistic features not exploited in this exercise use of capitalization, fraction of words that are proper nouns (using gazetteers), and others. Besides, we can also aggregate the well-performed classifiers to achieve better accuracy. For example, using bootstrap aggregating for models to get better prediction result. An ambitious work would be to search the news on the Internet and compare the search results with the original news. Since the search result is usually reliable, this method should be more accurate, but also involves natural language understanding because the search results will not be exactly the same as the original news. So we will need to compare the meaning of two contents and decide whether they mean the same thing.

References

- Dataset from Kaggle **https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset**

- Wikipedia
- Github
- http://cs229.stanford.edu/proj2017/final-reports/5244348.pdf