Fake News Detection Project

Submitted by:

Sabuj Ghosh

ACKNOWLEDGMENT

Aurelien Geron’s book titled ‘Hands-on Machine Learning with Scikit-Learn, Keras & Tensorflow’ was really helpful.

INTRODUCTION

- ## Business Problem Framing

  Nowadays political and economic agendas are being pushed like anything on news media. Back when Noam Chomsky theorized 'Manufacturing consent' to this volatile socio-political setup where we can see social media flooded with fake news and allegations, it is very important to distinguish the fake news from the real news.

- ## Conceptual Background of the Domain Problem

  Understanding regarding Natural language processing and how fake news is perpetuated is very important for this problem.

- ## Motivation for the Problem Undertaken

  To build a model that can detect fake news without human intervention.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  This is a classification problem so we focused on classifier models.

- Data Sources and their formats

  Provided by Fliprobo.

- Data Preprocessing Done

  Null value handling, outlier detection and removal.

  Presence of null value almost certainly denotes a fake news, so that information was stored in a separate column.

  Ratio of length of headline to length of news is very high when it is a fake news; that ratio was saved in a different column.

- Data Inputs- Logic- Output Relationships

  Relative frequency of various tokenized words can predict if a news is fake or not.

- Hardware and Software Requirements and Tools Used

  Hardware: Personal PC.

  Software: Jupyter for python coding, python modules include Numpy, Pandas, Matplotlib, Scikit-Learn, NLTK among many others.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  After removal of stopwords, individual news were tokenize, their frequency with respect to all other news was vectorized to feed into various models.

- Testing of Identified Approaches (Algorithms)
  - ◆Null value handled.
  - ◆Outliers removed.
  - ◆Stopwords removed.
  - ◆News, headlines and ‘Written by’ tokenized.
  - ◆Corpus of news, headlines and ‘Written by’ built separately.
  - ◆News, headlines and ‘Written by’ vectorized.
  - ◆Those vectors stacked horizontally.
  - ◆Train data set and test data set created from the vectors and labels.
  - ◆Various models trained with train data and tested using test data using various evaluation metrics.
  - ◆The highest performing model got hypertuned.
  - ◆The highest performing model got cross validated.
  - ◆After successful cross validation a model is selected.

- Run and Evaluate selected models

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(x,y,test_size=0.25)
```

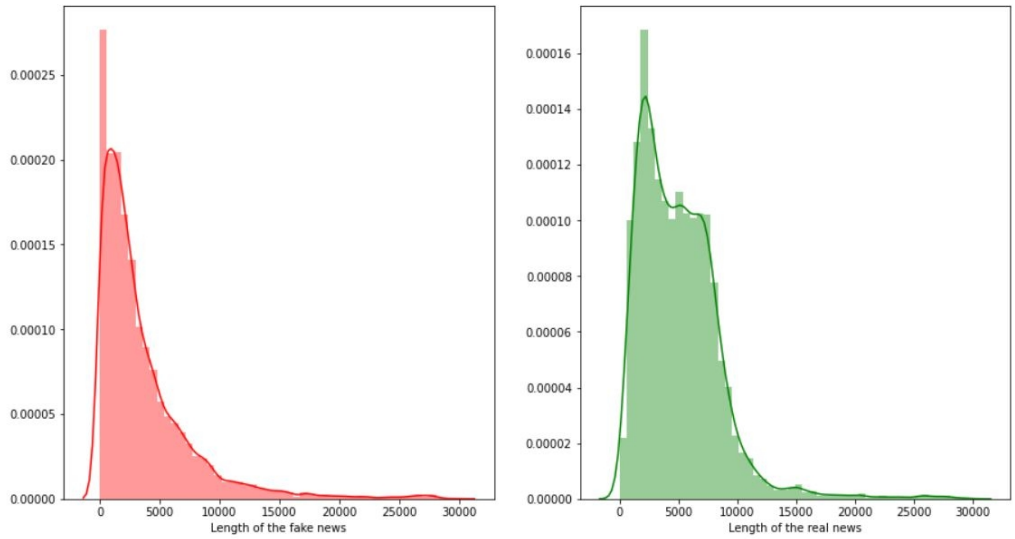
```
from sklearn.naive_bayes import MultinomialNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, classification_report,f1_score,log_loss

MNB=MultinomialNB()
DTC=DecisionTreeClassifier()
SVMR=SVC()
KNN=KNeighborsClassifier()
LR=LogisticRegression()
models=[MNB,DTC,SVMR,KNN,LR]

```

```
for model in models:
    model.fit(X_train,y_train)
    y_pred=model.predict(X_test)
    print('---'*15)
    print(model)
    print(confusion_matrix(y_test,y_pred))
    print(classification_report(y_test,y_pred))
    print(f1_score(y_test,y_pred))
    print(log_loss(y_test,y_pred))
```

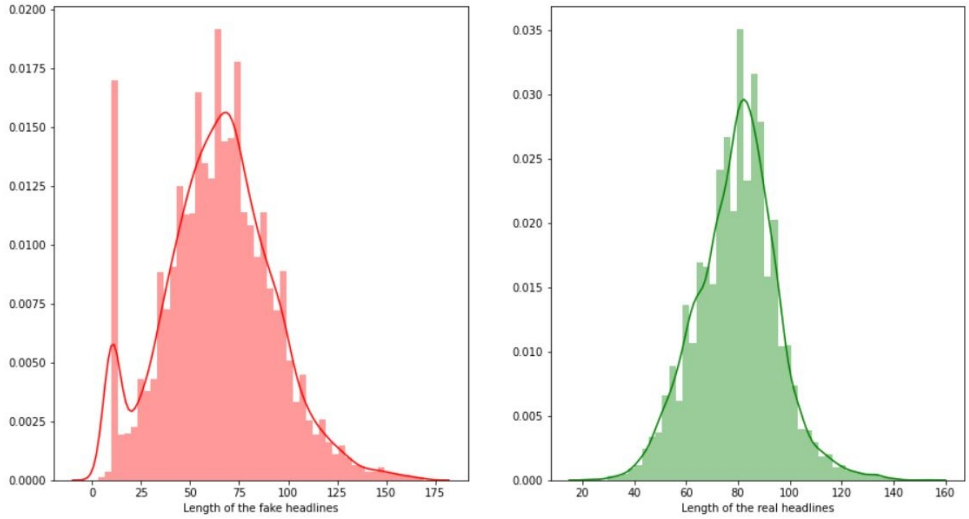
```
---------------------------------------------
MultinomialNB()
[[2483   60]
 [ 247 2371]]
              precision    recall  f1-score   support

           0       0.91      0.98      0.94      2543
           1       0.98      0.91      0.94      2618

    accuracy                           0.94      5161
   macro avg       0.94      0.94      0.94      5161
weighted avg       0.94      0.94      0.94      5161

0.9391958803723509
2.054534456323076
---------------------------------------------
DecisionTreeClassifier()
[[2449   94]
 [  79 2539]]
              precision    recall  f1-score   support

           0       0.97      0.96      0.97      2543
           1       0.96      0.97      0.97      2618

    accuracy                           0.97      5161
   macro avg       0.97      0.97      0.97      5161
weighted avg       0.97      0.97      0.97      5161

0.9670538944962864
1.1577762988719222
```

```
---------------------------------------------
SVC()
[[2466   77]
 [  13 2605]]
              precision    recall  f1-score   support

           0       0.99      0.97      0.98      2543
           1       0.97      1.00      0.98      2618

    accuracy                           0.98      5161
   macro avg       0.98      0.98      0.98      5161
weighted avg       0.98      0.98      0.98      5161

0.9830188679245284
0.6023157226398161
---------------------------------------------
KNeighborsClassifier()
[[2148  395]
 [  55 2563]]
              precision    recall  f1-score   support

           0       0.98      0.84      0.91      2543
           1       0.87      0.98      0.92      2618

    accuracy                           0.91      5161
   macro avg       0.92      0.91      0.91      5161
weighted avg       0.92      0.91      0.91      5161

0.919296987087518
3.0115801625062466
---------------------------------------------


 ---------------------------------------------
 LogisticRegression()
 [[2472   71]
  [  22 2596]]
               precision    recall  f1-score   support

            0       0.99      0.97      0.98      2543
            1       0.97      0.99      0.98      2618

     accuracy                           0.98      5161
    macro avg       0.98      0.98      0.98      5161
 weighted avg       0.98      0.98      0.98      5161

 0.98240302743614
 0.622391586154668
```

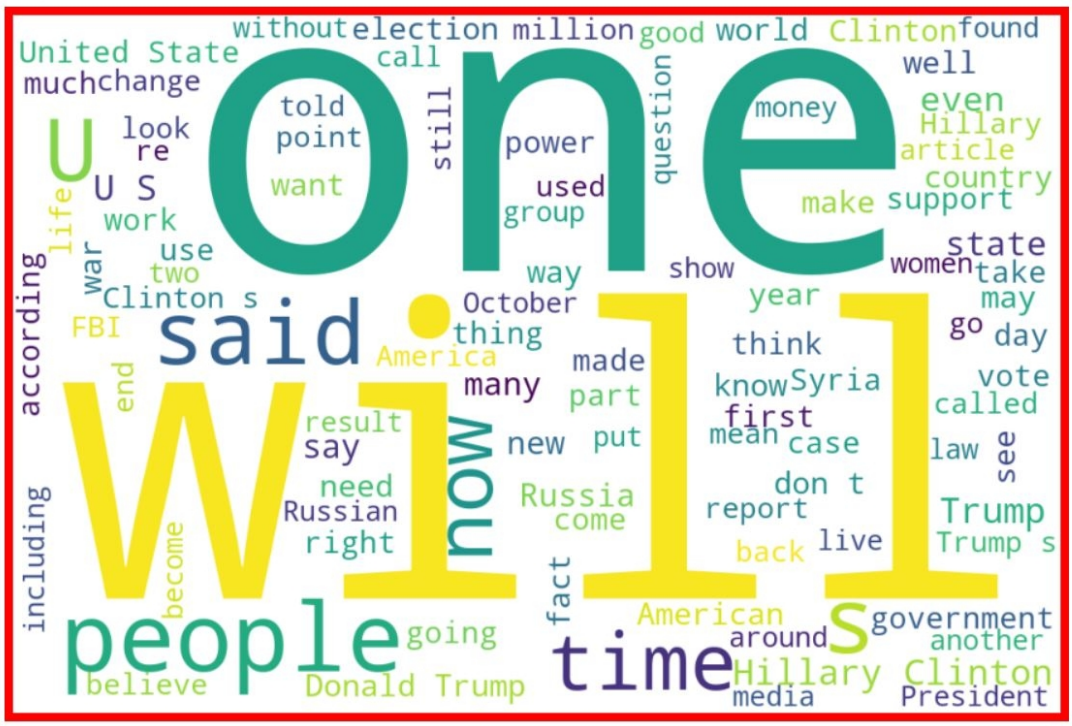
- Key Metrics for success in solving problem under consideration
    - f1 score.
    - Recall.
    - Precision.
    - Log-loss.
- Visualizations



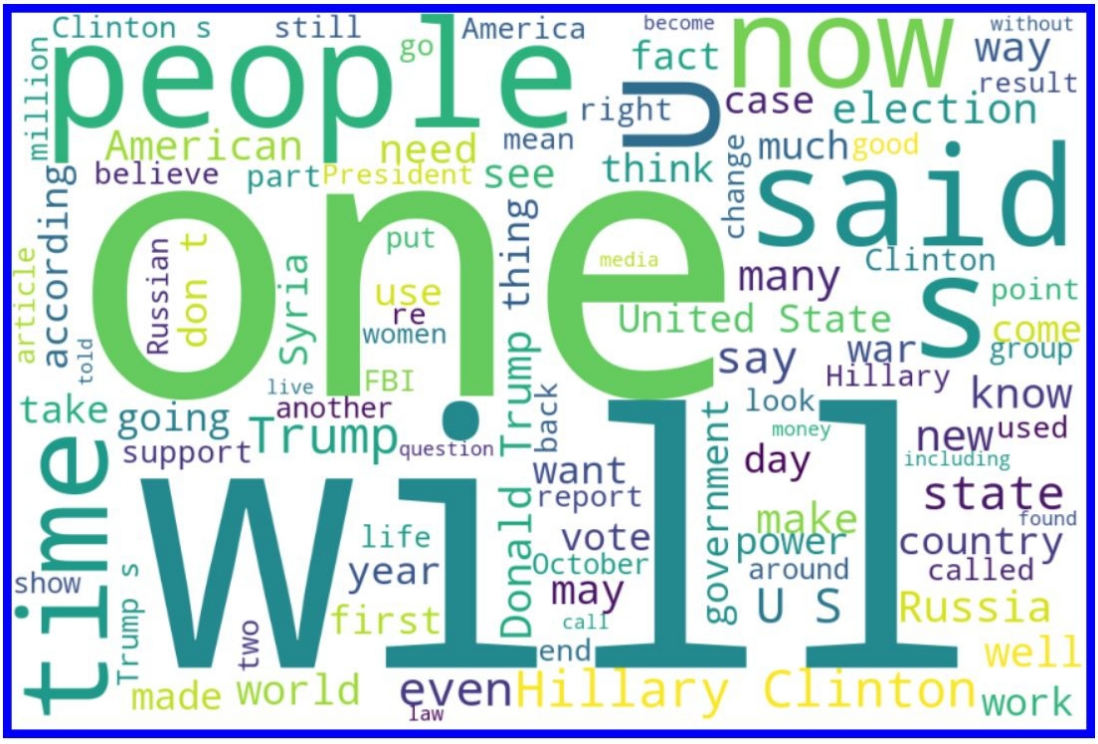
Distributions of length of fake news is more skewed and lengths of real news are significantly higher.



Fake news headlines are longer in comparison to real news.

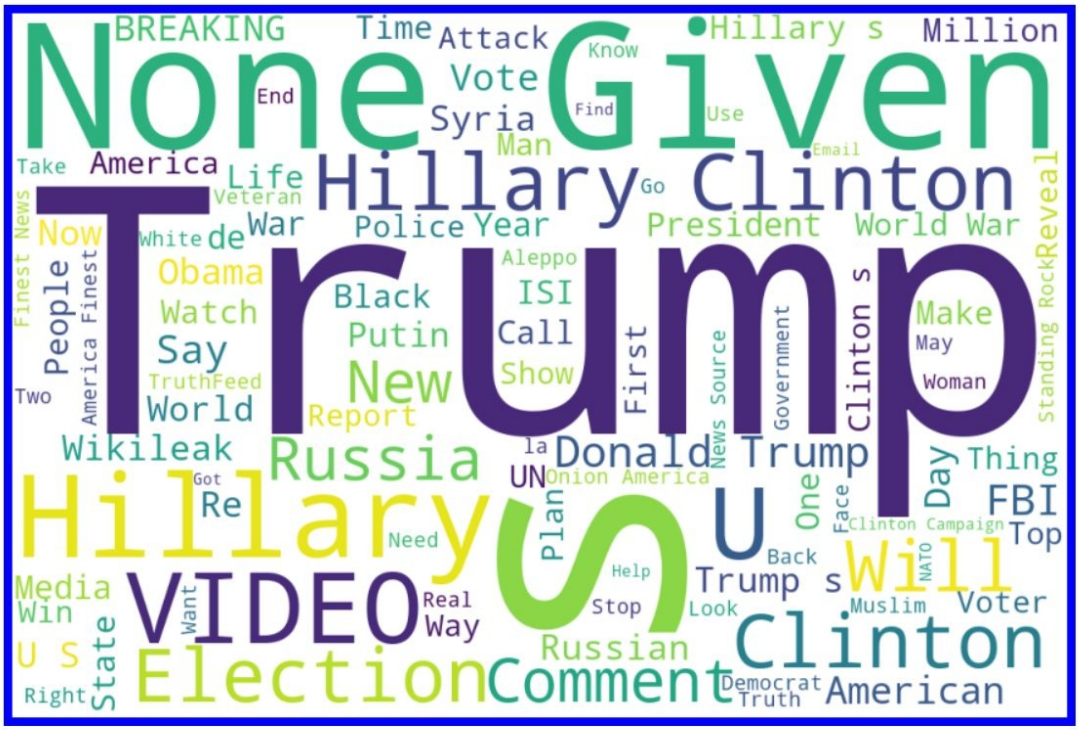


Word Cloud for fake news



Word Cloud for real news

Word Cloud for fake headlines



Word Cloud for real headlines

- Interpretation of the Results

  Visualizations: Fake news have longer headlines and shorter news and hence a feature column was created which contained length of headline/length of news.

  Preprocessing: Instances which contained news with length more than 30000 and headlines with length more than 170 was removed. Presence of null value in any column denoted a high probability of the news being fake so that information was saved in a separate column.

  Modelling: Support Vector classifier performs best for the dataset, close contending models include logistic regression and decision tree.

CONCLUSION

- Key Findings and Conclusions of the Study

  Fake news can be predicted with more than 98% accuracy using SVC.

- Learning Outcomes of the Study in respect of Data Science

  A good quality data goes a long way and is even important than efficient models too in some cases.