FAKE NEWS DETECTION PROJECT

Submitted by:
NIPAM GOGOI

ACKNOWLEDGMENT

I would like to thank Flip Robo Technologies for providing me with the opportunity to work on this project from which I have learned a lot. I am also grateful to Miss Khushboo Garg for her constant guidance and support.

INTRODUCTION

BUSINESS PROBLEM FRAMING

The authenticity of Information has become a longstanding issue affecting businesses and society, both for printed and digital media. On social networks, the reach and effects of information spread occur at such a fast pace and so amplified that distorted, inaccurate, or false information acquires a tremendous potential to cause real-world impacts, within minutes, for millions of users. Recently, several public concerns about this problem and some approaches to mitigate the problem were expressed.

Therefore, It is important to identify the fake news from the real true news. The problem has been taken over and resolved with the help of Natural Language Processing tools which help us identify fake or true news based on historical data.

CONCEPTUAL BACKGROUND OF THE DOMAIN PROBLEM

Internet is one of the important inventions and a large number of persons are its users. These persons use this for different purposes. There are different social media platforms that are accessible to these users. Any user can make a post or spread the news through these online platforms. These platforms do not verify the users or their posts. So some of the users try to spread fake news through these platforms. This fake news can be a propaganda against an individual, society, organization or political party. A human being is unable to detect all these fake news. So there is a need for machine learning classifiers that can detect these fake news automatically. Use of machine learning classifiers for detecting the fake news is described in this project report.

REVIEW OF LITERATURE

The widespread problem of fake news is very difficult to tackle in today's digital world where there are thousands of information sharing platforms through which fake news or misinformation may propagate. It has become a greater issue because of the advancements in AI which brings along artificial bots that may be used to create and spread fake news. The situation is dire because many people believe anything they read on the internet and the ones who are amateur or are new to the digital technology may be easily fooled. A similar problem is fraud that may happen due to spam or malicious emails and messages. So, it is compelling enough to acknowledge this problem take on this challenge to control the rates of crime, political unrest, grief, and thwart the attempts of spreading fake news. Text, or natural language, is one form that is difficult to process simply because of various linguistic features and styles like sarcasm, metaphors, etc. Moreover, there are thousands of spoken languages and every language has its grammar, script and syntax. Natural language processing is a branch of artificial intelligence and it encompasses techniques that can utilize text, create models and produce predictions. This work aims to create a system or model that can use the data of past news reports and predict the chances of a news report being fake or not. Fake news is not a new concept. Before the era of digital technology, it was spread through mainly yellow journalism with a focus on sensational news such as crime, gossip, disasters and satirical news. With the widespread dissemination of information via digital media platforms, it is of utmost importance for individuals and societies to be able to judge the credibility of it. Fake news is not a recent concept, but it is a commonly occurring phenomenon in current times. The consequence of fake news can range from being merely annoying to influencing and misleading societies or even nations. A variety of approaches exist to identify fake news.

ANALYTICAL PROBLEM FRAMING

Dataset description

There are 6 columns in the dataset provided:

The description of each of the column is given below:

- “id”: Unique id of each news article
- “headline”: It is the title of the news.
- “news”: It contains the full text of the news article
- “Unnamed:0”: It is a serial number
- “written_by”: It represents the author of the news article
- “label”: It tells whether the news is fake (1) or not fake (0).

MODEL DEVELOPMENT AND EVALUATION

Exploratory Data Analysis

```
1 df.head()
```

	Unnamed: 0	id	headline	written_by	news	label
0	0	9653	Ethics Questions Dogged Agriculture Nominee as...	Eric Lipton and Steve Eder	WASHINGTON — In Sonny Perdue's telling, Geo...	0
1	1	10041	U.S. Must Dig Deep to Stop Argentina's Lionel ...	David Waldstein	HOUSTON — Venezuela had a plan. It was a ta...	0
2	2	19113	Cotton to House: 'Do Not Walk the Plank and Vo...	Pam Key	Sunday on ABC's “This Week,” while discussing ...	0
3	3	6868	Paul LePage, Besieged Maine Governor, Sends Co...	Jess Bidgood	AUGUSTA, Me. — The beleaguered Republican g...	0
4	4	7596	A Digital 9/11 If Trump Wins	Finian Cunningham	Finian Cunningham has written extensively on...	1

```
1 df.shape  # Total 20800 records
```

```
(20800, 6)
```

```
1 df.written_by.nunique()   # 4201 unique writers
```

```
4201
```

```
1 df.nunique()  # Unique counts of every columns
```

```
Unnamed: 0    20800
id            20800
headline      19803
written_by     4201
news          20386
label             2
dtype: int64
```

- There are 20800 rows and 6 columns in the entire dataset.
- There were 4201 unique writers of news in the entire dataset.
- Na values existed in the Dataset.

```
df.isna().sum()   # na values exist in the Dataset
```

```
Unnamed: 0       0
id               0
headline       558
written_by    1957
news            39
label            0
dtype: int64
```

- The na values in news columns were dropped. These rows had Headlines but they were not in English language.

```
# Data with na values in News
df[df.news.isna() == True]
```

	Unnamed: 0	id	headline	written_by	news	label
556	556	9454	Il saoule tout le monde avec son analyse polit...	NaN	NaN	1
1642	1642	11486	Pour booster les ventes, Dassault offre un por...	NaN	NaN	1
1765	1765	573	Le top des recherches Google passe en top des ...	NaN	NaN	1
1968	1968	9446	Trop essoufflé après avoir cherché ses affaire...	NaN	NaN	1
2200	2200	3729	Les Américains ne sont plus qu'à quelques heur...	NaN	NaN	1
3183	3183	13107	Les gardes-frontières se mettent en alerte pou...	NaN	NaN	1
3927	3927	4358	Ne supportant plus l'ambiance de la campagne é...	NaN	NaN	1
4333	4333	14499	Primaire – François Fillon se désiste au profi...	NaN	NaN	1
4746	4746	2148	Gorafi Magazine: Barack Obama « Je vous ai déj...	NaN	NaN	1
4747	4747	8649	Donald Trump s'excuse pour toutes les minorité...	NaN	NaN	1
4942	4942	10867	Live Soirée présidentielle US 2016 >> Le Gorafi	NaN	NaN	1
6849	6849	6215	New-York – Le lâcher de confettis prévu à Time...	NaN	NaN	1
7100	7100	3329	GuinessBook : 100 millions d'Américains batten...	NaN	NaN	1

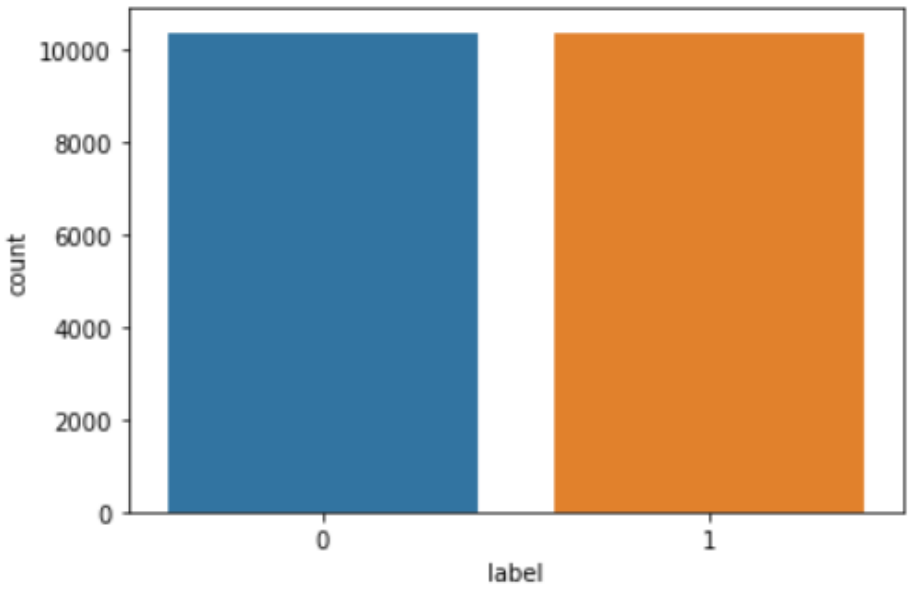
- It was a balanced Dataset even after the removal of na values.

```
: 1 df.label.value_counts()   # It's a Balanced Dataset

: 0    10387
  1    10374
  Name: label, dtype: int64
```

```
: 1 sns.countplot(df["label"])

: <matplotlib.axes._subplots.AxesSubplot at 0x2704cb23fa0>
```



- After that the dataset was filtered for rows with label = 1. i.e. fake news data.
- All the writers who used to write fake news are found from the dataset.

```
# Grouping the Fake News Data by Written By and adding labels to find total count of fake news per writer
new_df = df2.groupby(['written_by']).sum().reset_index()
```

```
# Creating a dataframe with only written by and count of Fake news
fakenews_writer = new_df[["written_by","label"]].sort_values(by=['label'], ascending= [False])
```

```
# The List of all Writers who write Fake News
list(fakenews_writer.written_by)
```

```
'BareNakedIslam',
'Activist Post',
'The Doc',
'EdJenner',
'Henry Wolff',
'Mac Slavo',
'Iron Sheik',
'Kaitlyn Stegall',
'Jason Ditz',
'noreply@blogger.com (Der Postillon)',
'Heather Callaghan',
'David Stockman',
'Geoffrey Grider',
'King World News',
'shorty',
'Consortiumnews.com',
'The Saker',
'-NO AUTHOR-',
'Dikran Arakelian (noreply@blogger.com)',
```

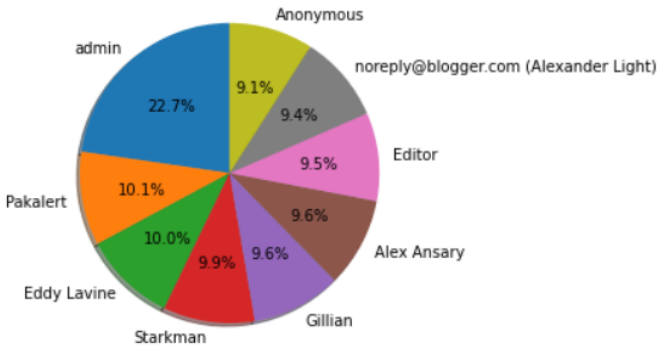
```
# Fake News writers with more than 50 fake news
fakenews_writer[fakenews_writer.label > 50]
```

	written_by	label
1710	admin	193
1213	Pakalert	86
526	Eddy Lavine	85
1474	Starkman	84
634	Gillian	82
59	Alex Ansary	82
527	Editor	81
1861	noreply@blogger.com (Alexander Light)	80
129	Anonymous	77
437	Dave Hodges	77

- Top 10 writers of fake news are plotted.

```
1 dict_2_pie_chart(dict(fakenews_writer.sort_values(by=['label'], ascending= [False])[:9].values))
```

```
<Figure size 1440x1440 with 0 Axes>
```

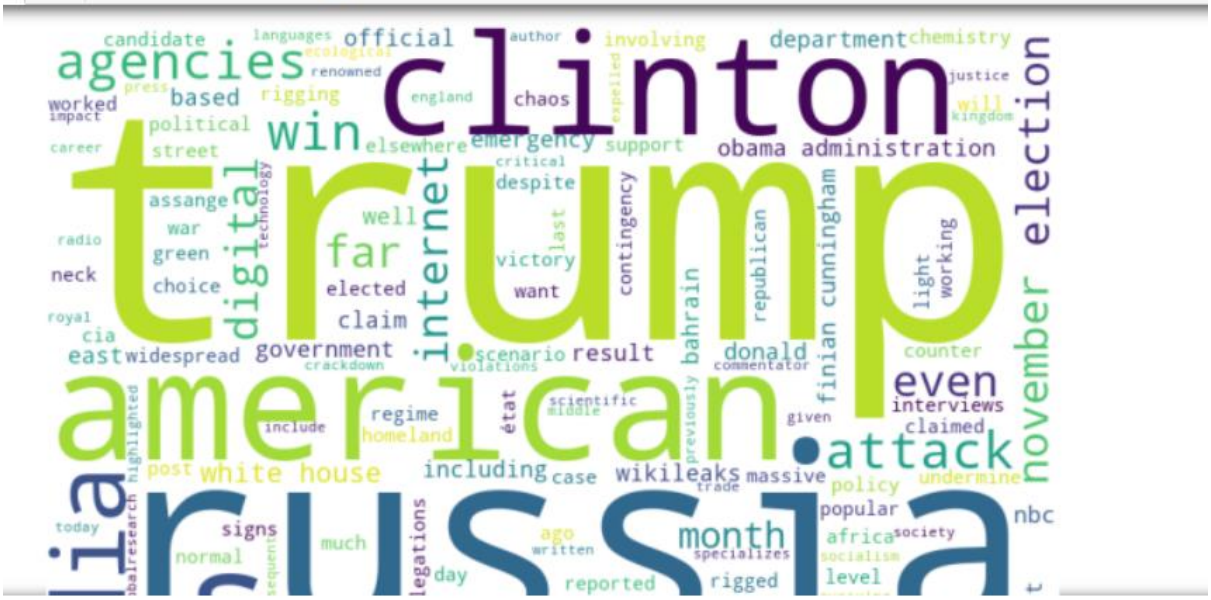


The writers with Highest number of Fake news are : admin, Pakalert, Eddy Lavine, Starkman etc to name a few.

- Word Clouds are plotted for 5 samples of Fake news as well as 5 samples of real news.

```
:
1 # Word Cloud of 5 Samples of Real News (Label = 0)
2 sample = list(df[df.label == 0].news[:5])
3 for i in sample:
4     Word_Cloud(i.split(" "))
```

```
:  1  # Word Cloud of 5 Samples of Fake News (Label = 1)
   2  sample = list(df[df.label == 1].news[:5])
   3  for i in sample:
   4      Word_Cloud(i.split(" "))
```



Here we can see that the word clouds of both fake as well as real news are hard to distinguish

KEY METRICS FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION

- When it comes to the evaluation of a data science model's performance, sometimes accuracy may not be the best indicator.
- So, we have used f1 score as well as recall, precision to check the performance of the models.

MODEL TRAINING

- Both the Headlines and news were combined to make the inputs for ML models.

Machine Learning Models

```
# Combining both Headlines and News to be Inputs for the ML Models
inputs = []
for i,j in zip(list(df.headline),list(df.news)):
    inputs.append(i + " " + j)

len(inputs)
```

```
20761
```

```
y = df.label
```

```
len(y)
```

```
20761
```

```
# Splitting into training and Test dataset
X_train,X_test,y_train,y_test=train_test_split(inputs,y,test_size=0.2,random_state=88)
```

Using TF-IDF to convert text data into numerical format for ML models

```
vectorizer = TfidfVectorizer(min_df =1,stop_words='english',use_idf=True,analyzer='word',
                             ngram_range=(1,1),max_features=15000)
x_train = vectorizer.fit_transform(X_train)
x_test  = vectorizer.transform(X_test)
```

- After that the Text data was converted into numerical format using TF_IDF vectorizer so that Machine Learning models can be trained.

- Four ML models were used to train the dataset using Sklearn, out of which the best was the Random Forest Model.

Random Forest

```
rand = RandomForestClassifier()
rand.fit(x_train, y_train)
```

```
RandomForestClassifier()
```

```
prediction2 = rand.predict(x_test)
```

```
print('\n','CONFUSION MATRIX','\n',confusion_matrix(y_test, prediction2))
print('\n','ACCURACY','\n',accuracy_score(y_test, prediction2))
print('\n','REPORT','\n',classification_report(y_test,prediction2))
```

```
CONFUSION MATRIX
[[2003   62]
[  80 2008]]

ACCURACY
0.9658078497471707

REPORT
              precision    recall  f1-score   support

           0       0.96      0.97      0.97      2065
           1       0.97      0.96      0.97      2088

    accuracy                           0.97      4153
   macro avg       0.97      0.97      0.97      4153
weighted avg       0.97      0.97      0.97      4153
```

CONCLUSION

Most of the Fake News were associated with US election campaigns with Keywords like Trump, US, Senate etc showing up in the entire dataset.

Top 10 writers generating highest number of fake news are:

'admin', 'Pakalert', 'Eddy Lavine', 'Starkman', 'Gillian', 'Alex Ansary', 'Editor', 'noreply@blogger.com (Alexander Light)', 'Anonymous', 'Dave Hodges'.

It was a balanced Dataset. All the models trained using Sklearn performed well with best performance by Random Forest with 97 percent accuracy as well as 97 percent f1 score.