PROJECT REPORT ON:

“Fake News Prediction”

SUBMITTED BY

Ajit Madame

ACKNOWLEDGMENT

I would like to express my special gratitude to "Flip Robo" team, who has given me this opportunity to deal with a beautiful dataset and it has helped me to improve my analyzation skills. And I want to express my huge gratitude to Ms.Gulshana Chaudhary (SME Flip Robo), she is the person who has helped me to get out of all the difficulties I faced while doing the project.

A huge thanks to my academic team "Data trained" who are the reason behind what I am today. Last but not least my parents who have been my backbone in every step of my life. And also thank you for many other persons who has helped me directly or indirectly to complete the project.

Contents:

1.INTRODUCTION

1.1Business Problem Framing:

Fake news has become one of the biggest problems of our age. It has serious impact on our online as well as offline discourse. One can even go as far as saying that, to date, fake news poses a clear and present danger to western democracy and stability of the society.

What's inside is more than just rows and columns. Make it easy for others to get started by describing how you acquired the data and what time period it represents, too.

1.2 Conceptual Background of the Domain Problem

Fake news's simple meaning is to incorporate information that leads people to the wrong path. Nowadays fake news spreading like water and people share this information without verifying it. This is often done to further or impose certain ideas and is often achieved with political agendas.

For media outlets, the ability to attract viewers to their websites is necessary to generate online advertising revenue. So, it is necessary to detect fake news.

1.3 Review of Literature

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. By conducting a systematic literature review, we identify the main approaches currently available to identify fake news and how these approaches can be applied in different situations. Some approaches are illustrated with a relevant example as well as the challenges and the appropriate context in which the specific approach can be used.

1.4 Motivation for the Problem Undertaken

Misinformation refers to misrepresented information in a macro aspect, including a series of fabricated, misleading, false, fake, deceptive or distorted information. It is usually created by information creators with malicious intentions for achieving certain purposes. As such, the credibility of the information is usually undermined. Under the common umbrella of conveying misrepresented information, it closely relates to several similar concepts, such as **fake news**, **rumor**, **deception**, **hoaxes**, **spam opinion** etc. Despite being similar, there exists salient differences among them in terms of the degrees of wrongness, the contexts of usage and the functions of serving for different propagation purposes. Below will address the main concepts of the several varieties of misinformation.

2.Analytical Problem Framing

2.1 Mathematical/ Analytical Modelling of the Problem

The libraries/dependencies imported for this project are shown below:

```
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
```

```
import nltk
nltk.download('stopwords', quiet=True)
nltk.download('punkt', quiet=True)
from wordcloud import WordCloud
from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer
from nltk.tokenize import word_tokenize, regexp_tokenize

from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.model_selection import train_test_split, cross_val_score, GridSearchCV,
from scipy.sparse import csr_matrix
```

Here in this project, we have been provided with two datasets namely fake and ture CSV files. I will be combined both CSV file to make one dataset. I will build a machine learning model using train dataset. And using this model we will make predictions for our test dataset.

2.2 Data Sources and their formats

We have been provided with two datasets namely train and test CSV filers. Train datasets contains 44898 rows and 5 columns.

	title	text	subject	date	class
0	Donald Trump Sends Out Embarrassing New Yearâ...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
3	Trump Is So Obsessed He Even Has Obamaâ□□s Na...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0
5	Racist Alabama Cops Brutalize Black Boy While...	The number of cases of cops brutalizing and ki...	News	December 25, 2017	0

2.3 Data Pre-processing Done

- First step I have imported required libraries and I have imported the dataset which was in csv format.
- Then I did all the statistical analysis like checking shape, nunique, value counts, info etc.
- Apply Stemming using SnowballStemmer.
- Here I convert text data into vector form by using TfidfVectorizer.
- Then doing some EDA and Building Models.

2.4 Data Inputs - Logic - Output Relationships

I have analysed the input output logic with word cloud and I have word clouded the sentenced that as classified as foul language in every category. A tag/word cloud is a novelty visual representation of text data, typically used to depict keyword metadata on websites, or to visualize free form text. It's an image composed of words used in a particular text or subject, in which the size of each word indicates its frequency or importance.

```
c_0 = df['text'][df['class']==0]
spam_cloud = WordCloud(width=600,height=400,background_color='black',max_words=50).generate(' '.join(c_0))
plt.figure(figsize=(4,4),facecolor='k')
plt.imshow(spam_cloud)
plt.axis('off')
plt.tight_layout(pad=0)
plt.show()
```

2.5 Hardware, Software and Tool Used

Hardware Used:

Processor – Intel core i3

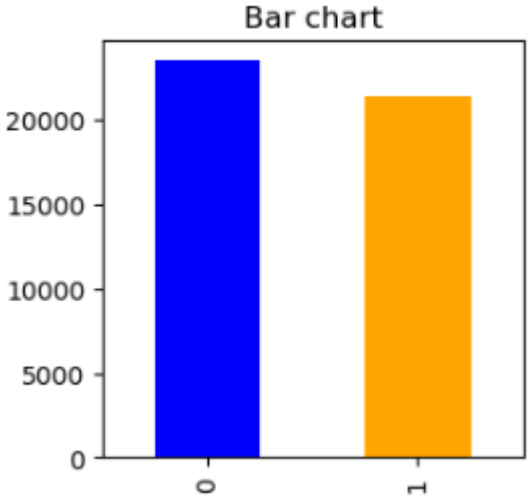
Physical Memory – 8 GB

Software Used:

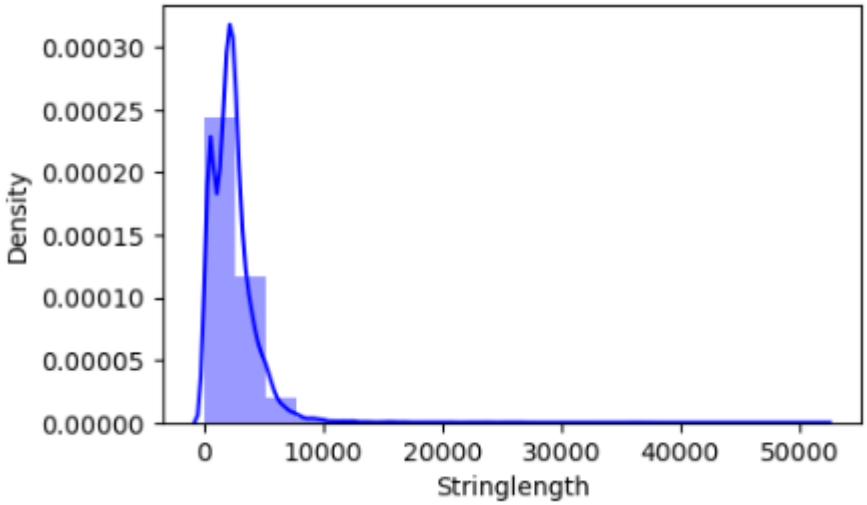
- Windows 10 Operating System
- Anaconda Package and Environment Manager
- Jupyter Notebook
- Python Libraries used: In Which Pandas, Seaborn, Matplotlib, Numpy and Scipy
- sklearn for Modelling Machine learning algorithms, Data Encoding, Evaluation metrics, Data Transformation, Data Scaling, Component analysis, Feature selection etc.

3.Data Analysis and Visualization

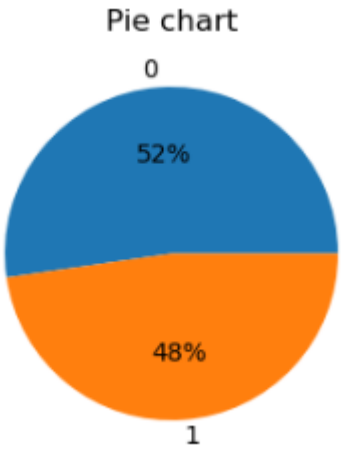
3.1 Univariate Visualization



- We can see, in fake news and true news in which have a less difference.
- Data imbalanced so we need balance it but difference is less so it will not impact that much so I take as it is.

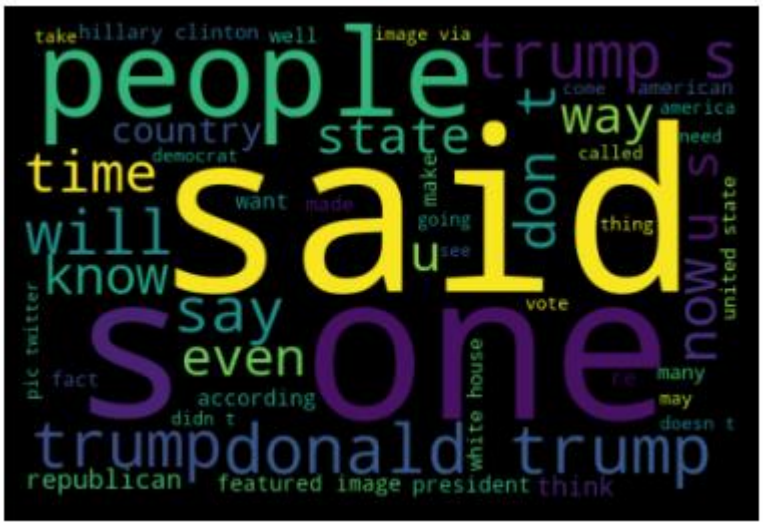


- we can see most of the news are lies between 0 to 10000 words.

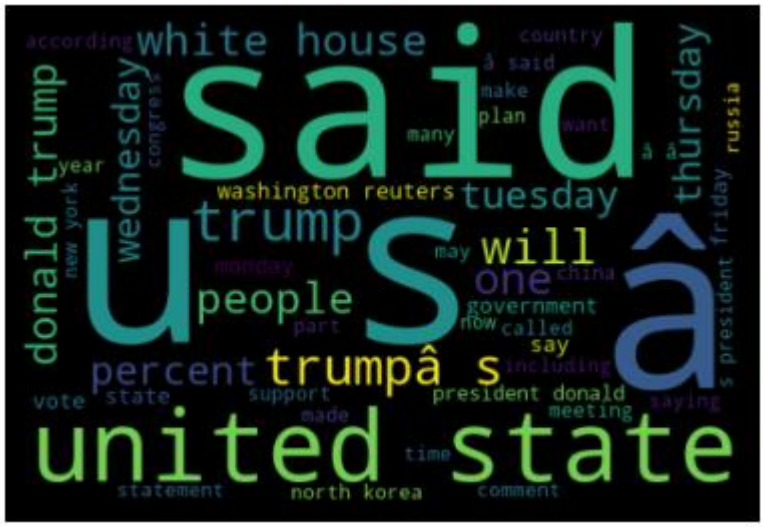


- We can see, in fake news and true news are having only 4% difference.

3.2 Word Cloud Visualization



- People, time, said, one and trump these are the word are mostly used class 0 (fake news).



- Said, USA, percent and Trump these are the words are in True news. But few words are also in fake news.

4. Models Development and Evaluation

```
def wordopt(text):
    text = text.lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub("\\W"," ",text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    return text
```

```
df["text"] = df["text"].apply(wordopt)
```

```
x = df["text"]
y = df["class"]
```

```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
```

Splitting Training and Testing

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25)
```

Convert text to vectors

```
from sklearn.feature_extraction.text import TfidfVectorizer

vectorization = TfidfVectorizer()
xv_train = vectorization.fit_transform(x_train)
xv_test = vectorization.transform(x_test)
```

Logistic Regression

```
from sklearn.linear_model import LogisticRegression

LR = LogisticRegression()
LR.fit(xv_train,y_train)
```

```
LogisticRegression()
```

```
pred_lr=LR.predict(xv_test)
```

```
LR.score(xv_test, y_test)
```

```
0.9884135472370766
```

```
print(classification_report(y_test, pred_lr))
```

```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5808
           1       0.99      0.99      0.99      5412

    accuracy                           0.99     11220
   macro avg       0.99      0.99      0.99     11220
weighted avg       0.99      0.99      0.99     11220
```

Decision Tree Classifier

```
from sklearn.tree import DecisionTreeClassifier

DT = DecisionTreeClassifier()
DT.fit(xv_train, y_train)
```

```
DecisionTreeClassifier()
```

```
pred_dt = DT.predict(xv_test)
```

```
DT.score(xv_test, y_test)
```

```
0.9963458110516934
```

```
print(classification_report(y_test, pred_dt))
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      5808
           1       1.00      1.00      1.00      5412

    accuracy                           1.00     11220
   macro avg       1.00      1.00      1.00     11220
weighted avg       1.00      1.00      1.00     11220
```

Gradient Boosting Classifier

```
from sklearn.ensemble import GradientBoostingClassifier

GBC = GradientBoostingClassifier(random_state=0)
GBC.fit(xv_train, y_train)
```

```
GradientBoostingClassifier(random_state=0)
```

```
pred_gbc = GBC.predict(xv_test)
```

```
GBC.score(xv_test, y_test)
```

```
0.995632798573975
```

```
print(classification_report(y_test, pred_gbc))
```

```
              precision    recall  f1-score   support

           0       1.00      0.99      1.00      5808
           1       0.99      1.00      1.00      5412

    accuracy                           1.00     11220
   macro avg       1.00      1.00      1.00     11220
weighted avg       1.00      1.00      1.00     11220
```

Random Forest Classifier

```
from sklearn.ensemble import RandomForestClassifier

RFC = RandomForestClassifier(random_state=0)
RFC.fit(xv_train, y_train)
```

```
RandomForestClassifier(random_state=0)
```

```
pred_rfc = RFC.predict(xv_test)
```

```
RFC.score(xv_test, y_test)
```

```
0.9933155080213903
```

```
print(classification_report(y_test, pred_rfc))
```

```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5808
           1       0.99      0.99      0.99      5412

    accuracy                           0.99     11220
   macro avg       0.99      0.99      0.99     11220
weighted avg       0.99      0.99      0.99     11220
```

Model Evaluation

```
def output_lable(n):
    if n == 0:
        return "Fake News"
    elif n == 1:
        return "Not A Fake News"

def manual_testing(news):
    testing_news = {"text":[news]}
    new_def_test = pd.DataFrame(testing_news)
    new_def_test["text"] = new_def_test["text"].apply(wordopt)
    new_x_test = new_def_test["text"]
    new_xv_test = vectorization.transform(new_x_test)
    pred_LR = LR.predict(new_xv_test)
    pred_DT = DT.predict(new_xv_test)
    pred_GBC = GBC.predict(new_xv_test)
    pred_RFC = RFC.predict(new_xv_test)

    return print("\n\nLR Prediction: {} \nDT Prediction: {} \nGBC Prediction: {} \nRFC Prediction: {}".format(output_lable(pred_L
                                                                                                              output_lable(pred_D
                                                                                                              output_lable(pred_G
                                                                                                              output_lable(pred_R
```

```
news = str(input())
manual_testing(news)
```

```
Vic Bishop Waking TimesOur reality is carefully constructed by powerful corporate, political and special interest sources in or
der to covertly sway public opinion. Blatant lies are often televised regarding terrorism, food, war, health, etc. They are fas
hioned to sway public opinion and condition viewers to accept what have become destructive societal norms.The practice of manip
ulating and controlling public opinion with distorted media messages has become so common that there is a whole industry formed
around this. The entire role of this brainwashing industry is to figure out how to spin information to journalists, similar to
the lobbying of government. It is never really clear just how much truth the journalists receive because the news industry has
become complacent. The messages that it presents are shaped by corporate powers who often spend millions on advertising with th
e six conglomerates that own 90% of the media:General Electric (GE), News-Corp, Disney, Viacom, Time Warner, and CBS. Yet, thes
e corporations function under many different brands, such as FOX, ABC, CNN, Comcast, Wall Street Journal, etc, giving people th
e perception of choice  As Tavistock s researchers showed, it was important that the victims of mass brainwashing not be aware
that their environment was being controlled; there should thus be a vast number of sources for information, whose messages coul
d be varied slightly, so as to mask the sense of external control. ~ Specialist of mass brainwashing, L. WolfeNew Brainwashing
Tactic Called AstroturfWith alternative media on the rise, the propaganda machine continues to expand. Below is a video of Shar
yl Attkisson, investigative reporter with CBS, during which she explains how  astroturf,  or fake grassroots movements, are use
d to spin information not only to influence journalists but to sway public opinion. Astroturf is a perversion of grassroots. As
troturf is when political, corporate or other special interests disguise themselves and publish blogs, start facebook and twitt
er accounts, publish ads, letters to the editor, or simply post comments online, to try to fool you into thinking an independen
t or grassroots movement is speaking. ~ Sharyl Attkisson, Investigative ReporterHow do you separate fact from fiction? Sharyl A
ttkisson finishes her talk with some insights on how to identify signs of propaganda and astroturfing  These methods are used t
o give people the impression that there is widespread support for an agenda, when, in reality, one may not exist. Astroturf tac
tics are also used to discredit or criticize those that disagree with certain agendas, using stereotypical names such as conspi
racy theorist or quack. When in fact when someone dares to reveal the truth or questions the  official  story, it should spark
a deeper curiosity and encourage further scrutiny of the information.This article (Journalist Reveals Tactics Brainwashing Indu
stry Uses to Manipulate the Public) was originally created and published by Waking Times and is published here under a Creative
Commons license with attribution to Vic Bishop and WakingTimes.com. It may be re-posted freely with proper attribution, author
bio, and this copyright statement. READ MORE MSM PROPAGANDA NEWS AT: 21st Century Wire MSM Watch Files
```

```
LR Prediction: Fake News
DT Prediction: Fake News
GBC Prediction: Fake News
RFC Prediction: Fake News
```

```
news = str(input())
manual_testing(news)
```

```
SAO PAULO (Reuters) - Cesar Mata Pires, the owner and co-founder of Brazilian engineering conglomerate OAS SA, one of the large
st companies involved in Brazil s corruption scandal, died on Tuesday. He was 68. Mata Pires died of a heart attack while takin
g a morning walk in an upscale district of S o Paulo, where OAS is based, a person with direct knowledge of the matter said. Ef
forts to contact his family were unsuccessful. OAS declined to comment. The son of a wealthy cattle rancher in the northeastern
state of Bahia, Mata Pires  links to politicians were central to the expansion of OAS, which became Brazil s No. 4 builder earl
ier this decade, people familiar with his career told Reuters last year. His big break came when he befriended Antonio Carlos M
agalh es, a popular politician who was Bahia governor several times, and eventually married his daughter Tereza. Brazilians jok
ed that OAS stood for  Obras Arranjadas pelo Sogro  - or  Work Arranged by the Father-In-Law.  After years of steady growth tr
iggered by a flurry of massive government contracts, OAS was ensnared in Operation Car Wash which unearthed an illegal contract
ing ring between state firms and builders. The ensuing scandal helped topple former Brazilian President Dilma Rousseff last yea
r. Trained as an engineer, Mata Pires founded OAS with two colleagues in 1976 to do sub-contracting work for larger rival Odebr
echt SA - the biggest of the builders involved in the probe.  Before the scandal, Forbes magazine estimated Mata Pires  fortune
at $1.6 billion. He dropped off the magazine s billionaire list in 2015, months after OAS sought bankruptcy protection after th
e Car Wash scandal. While Mata Pires was never accused of wrongdoing in the investigations, creditors demanded he and his famil
y stay away from the builder s day-to-day operations, people directly involved in the negotiations told Reuters at the time. He
is survived by his wife and his two sons.


LR Prediction: Not A Fake News
DT Prediction: Not A Fake News
GBC Prediction: Not A Fake News
RFC Prediction: Not A Fake News
```

5. Conclusions

5.1 Key Finding and Conclusions

The finding of the study is that only few users over online use unparliamentary language. And most of these sentences have more stop words and are being quite long. As discussed before few motivated disrespectful crowds use these foul languages in the online forum to bully the people around and to stop them from doing these things that they are not supposed to do. Our study helps the online forums and social media to induce a ban to profanity or usage of profanity over these forums.

Due to fake news Distrust in the media, Undermining the democratic process, Platforms for harmful conspiracy theories and hate speech

5.2 Limitation of this works and scope for future works

Problems faced while working in this project:

- More computational power was required as it took more than 2 hours
- Dataset quite balanced but too large texts
- Good parameters could not be obtained using hyperparameter tuning as time was consumed more, Areas of improvement:
- Could be provided with a good dataset which does not take more time.
- Less time complexity

4.2 Interpretation of the results

Based on comparing the above Precision, Recall, Accuracy Scores with all models, it is determined that Complement Decision Tree Classifier is the best model for the dataset.