Machine Learning Project - Fake News Detection

Created by Aditya Thakare

Fake News Detection using Classification Models

In this project I am using Classification models such as Logistic Regression, Decision Tree, Random Forest and Ensemble Learning

OVERVIEW:-

In recent years, the dissemination of fake news has been brought more and more into the spotlight as it has been massively used to disseminate political propaganda, influence the outcome of elections or harm a person or a group of people.

Highly sophisticated applications (bots) are organised in networks and massively spread to amplify fake news over social media in the form of text, images, audio or video files. Often, these bot nets happened to be organised by foreign state actors, trying to obscure the originator.

Fighting fake news is extremely challenging, as:

in a democracy, freedom of speech is a fundamental right fostering media independence and pluralism; however, sometimes there is a very subtle line between separating unconventional personal views and claims of truth from fake news;
fake news can be detected by checking consistency of the news with different domains, such as technical background to discover the real sender or social and/or judicial background (for example: what is the intention of the fake message, e.g. putting harm on a person or a group); therefore, fact-checking requires having awareness on different contexts and the availability of reliable sources;
the sheer mass of fake news spread over social media cannot be handled manually.
Manual fact checking can address some of these challenges, for example when checking the consistency of news in different contexts. However, manual fact-checking is too slow to cover big information spreaders such as social media platforms. This is where automation comes into play. Automated fact-checking tools often combine different methods, for example artificial intelligence, natural language processing (analysing the language used) and blockchain. As regards to fake news embedded in images and videos, the tools often combine metadata; social interactions; visual cues; the profile of the source; and other contextual information surrounding an image or video to increase accuracy.

Algorithms are trained to verify news content; detect amplification (excessive and/or targeted dissemination); spot fake accounts and detect campaigns. Often, the fake news analysis process applies several algorithms sequentially. However, effectiveness of these algorithms is yet to be improved.

Even if fake news is spread heavily on social media, research has found that human behaviour (“word of mouth” marketing) contributes more to the spread of fake news than automated bots do. This shows that fighting the fake news sender is not the only approach. It also makes sense to increase the resilience to fake news on the side of the recipient and our society. Therefore, another important pillar of fake news detection is to increase citizens’ awareness and media literacy.

GOAL OF THIS PROJECT

The goal of a fake news detection ML project is to develop a machine learning model that can automatically identify and classify news articles or other pieces of information as either "real" or "fake".

The proliferation of fake news has become a significant problem in recent years, with many people relying on social media as their primary source of news. Fake news can be used to spread propaganda, mislead people, and influence public opinion in a negative way. The detection of fake news is essential to ensure that people receive accurate and truthful information, especially when it comes to critical issues like politics, health, and safety.

By developing an ML model that can automatically detect fake news, it can help to combat the spread of fake news, reduce the impact of misinformation, and protect people from being misled by false information. The project involves using various techniques like natural language processing, data mining, and machine learning algorithms to analyze the text and other features of news articles and determine if they are fake or genuine.

ANALYSING THE DATASET

The Dataset is consist of
1) index number
2) Title - It is a title of the news article
3) Text - The text is the news article
4) Subject - Subject is the type of the news
5) date - date is the publishing date of the news article

The fake news and True news dataset combined has a 44878 rows and 5 columns

STEPS TO BE PERFORMED

1) we neeed to import all the necessary liabraries
2) we need to import the necessary datasets
3) we need to give the necessary class to the fake and true news
4) we need to merge the datasets
5) we need to drop all the unnecessary columns
6) we need to create a function to remove urls and special symbols from the text/news articles
7) we need to put class in x and text in y
8) create the machine learning models and test the dataset with the different machine learning models in order to get the best accuracy.

1) Importing the Libraries

In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
import warnings
warnings.filterwarnings("ignore")
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
import re
import string
```

Importing the Dataset

In [2]:

```
data_fake = pd.read_csv("fake.csv")
data_true = pd.read_csv("true.csv")
```

In [3]:

```
data_fake
```

Out[3]:

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year’...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama’s Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017
...	...	...	...	...
23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016
23478	Sunnistan: US and Allied ‘Safe Zone’ Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016
23480	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016

23481 rows × 4 columns

In [4]:

```
data_true
```

Out[4]:

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017
...	...	...	...	...
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017
21415	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017

21417 rows × 4 columns

3) EDA

In [47]:

```
data_true.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21407 entries, 0 to 21406
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   title    21407 non-null  object
 1   text     21407 non-null  object
 2   subject  21407 non-null  object
 3   date     21407 non-null  object
 4   class    21407 non-null  int64 
dtypes: int64(1), object(4)
memory usage: 836.3+ KB
```

The True values dataset is consist of 21407 rows and 5 columns and there is no null values in the dataset and the datatypes are int64 and object datatypes

In [48]:

```
data_fake.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23471 entries, 0 to 23470
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   title    23471 non-null  object
 1   text     23471 non-null  object
 2   subject  23471 non-null  object
 3   date     23471 non-null  object
 4   class    23471 non-null  int64 
dtypes: int64(1), object(4)
memory usage: 917.0+ KB
```

The False values dataset is consist of 23471 rows and 5 columns and there is no null values in the dataset and the datatypes are int64 and object datatypes

In [49]:

```
data_true.describe()
```

Out[49]:

	class
count	21407.0
mean	1.0
std	0.0
min	1.0
25%	1.0
50%	1.0
75%	1.0
max	1.0

In [*]:

```
sns.pairplot(data)
plt.show()
```

In [50]:

```
data_fake.describe()
```

Out[50]:

	class
count	23471.0
mean	0.0
std	0.0
min	0.0
25%	0.0
50%	0.0
75%	0.0
max	0.0

3) Merging two Datasets

In [5]:

```
data_fake["class"] = 0
data_true["class"] = 1
```

In [6]:

```
data_fake.shape, data_true.shape
```

Out[6]:

```
((23481, 5), (21417, 5))
```

In [7]:

```
data_fake_manual_testing = data_fake.tail(10)
for i in range(23480,23470,-1):
    data_fake.drop([i], axis = 0, inplace =True)

data_true_manual_testing = data_true.tail(10)
for i in range(21416,21406,-1):
    data_true.drop([i], axis =0, inplace=True )
```

In [8]:

```
data_fake.shape, data_true.shape
```

Out[8]:

```
((23471, 5), (21407, 5))
```

In [9]:

```
data_fake_manual_testing['class']=0
data_true_manual_testing['class']=1
```

In [10]:

```
data_fake_manual_testing.head(10)
```

Out[10]:

	title	text	subject	date	class
23471	Seven Iranians freed in the prisoner swap have...	21st Century Wire says This week, the historic...	Middle-east	January 20, 2016	0
23472	#Hashtag Hell & The Fake Left	By Dady Chery and Gilbert MercierAll writers ...	Middle-east	January 19, 2016	0
23473	Astroturfing: Journalist Reveals Brainwashing ...	Vic Bishop Waking TimesOur reality is carefull...	Middle-east	January 19, 2016	0
23474	The New American Century: An Era of Fraud	Paul Craig RobertsIn the last years of the 20t...	Middle-east	January 19, 2016	0
23475	Hillary Clinton: ‘Israel First’ (and no peace ...	Robert Fantina CounterpunchAlthough the United...	Middle-east	January 18, 2016	0
23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	0
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016	0
23478	Sunnistan: US and Allied ‘Safe Zone’ Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016	0
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016	0
23480	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016	0

In [11]:

```
data_true_manual_testing.head(10)
```

Out[11]:

	title	text	subject	date	class
21407	Mata Pires, owner of embattled Brazil builder ...	SAO PAULO (Reuters) - Cesar Mata Pires, the ow...	worldnews	August 22, 2017	1
21408	U.S., North Korea clash at U.N. forum over nuc...	GENEVA (Reuters) - North Korea and the United ...	worldnews	August 22, 2017	1
21409	U.S., North Korea clash at U.N. arms forum on ...	GENEVA (Reuters) - North Korea and the United ...	worldnews	August 22, 2017	1
21410	Headless torso could belong to submarine journ...	COPENHAGEN (Reuters) - Danish police said on T...	worldnews	August 22, 2017	1
21411	North Korea shipments to Syria chemical arms a...	UNITED NATIONS (Reuters) - Two North Korean sh...	worldnews	August 21, 2017	1
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	1
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	1
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	1
21415	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	1
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	1

In [12]:

```
data_merge = pd.concat([data_fake,data_true],axis=0)
data_merge.tail(10)
```

Out[12]:

	title	text	subject	date	class
0	Donald Trump Sends Out Embarrassing New Year’...	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 Obama’s Name...	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
6	Fresh Off The Golf Course, Trump Lashes Out A...	Donald Trump spent a good portion of his day a...	News	December 23, 2017	0
7	Trump Said Some INSANELY Racist Stuff Inside ...	In the wake of yet another court decision that...	News	December 23, 2017	0
8	Former CIA Director Slams Trump Over UN Bully...	Many people have raised the alarm regarding th...	News	December 22, 2017	0
9	WATCH: Brand-New Pro-Trump Ad Features So Muc...	Just when you might have thought we d get a br...	News	December 21, 2017	0

In [13]:

```
data_merge.columns
```

Out[13]:

```
Index(['title', 'text', 'subject', 'date', 'class'], dtype='object')
```

In [45]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 44878 entries, 0 to 44877
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   text    44878 non-null  object
 1   class   44878 non-null  int64 
dtypes: int64(1), object(1)
memory usage: 701.3+ KB
```

In [44]:

```
data.describe()
```

Out[44]:

	class
count	44878.000000
mean	0.477004
std	0.499476
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	1.000000

In [14]:

```
data = data_merge.drop(["title","subject","date"],axis=1)
```

In [15]:

```
data.isnull().sum()
```

Out[15]:

```
text     0
class    0
dtype: int64
```

After merging the dataset and dropping of unnecessary columns the dataset is consist of 44878 rows and 2 columns which will be useful for the purpose of model building and there is no null value present in the dataset

In [16]:

```
data = data.sample(frac=1)
```

In [17]:

```
data.head()
```

Out[17]:

	text	class
13440	GENEVA (Reuters) - In a thinly veiled referenc...	1
10043	Rebel Pundit decided to follow a local activis...	0
12264	A Wisconsin judge has refused to order local o...	0
20747	Was O Reilly out-of-bounds? Most Americans wou...	0
7179	Across the nation, mainly the South, there hav...	0

Dropping the index column

In [18]:

```
data.reset_index(inplace=True)
data.drop(["index"], axis=1, inplace=True)
```

In [19]:

```
data.columns
```

Out[19]:

```
Index(['text', 'class'], dtype='object')
```

In [20]:

```
data.head()
```

Out[20]:

	text	class
0	GENEVA (Reuters) - In a thinly veiled referenc...	1
1	Rebel Pundit decided to follow a local activis...	0
2	A Wisconsin judge has refused to order local o...	0
3	Was O Reilly out-of-bounds? Most Americans wou...	0
4	Across the nation, mainly the South, there hav...	0

Create the function to remove special symbols https address url from the text

In [21]:

```
import re
import string

def wordopt(text):
    text = text.lower()
    text = re.sub('\[.*?\]','',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)
    text = re.sub('[^a-zA-Z]', ' ', text)  # remove all non-alphabetic characters
    return text
```

In [22]:

```
'''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'''
```

Out[22]:

```
'def wordopt(text):\n    text = text.lower()\n    text = re.sub(\'\\[.*?
\\]\',\'\',text)\n    text = re.sub("\\w"," ",text)\n    text = re.sub(\'h
ttps?"//\\S+|www\\.\\S+\',\'\',text)\n    text = re.sub(\'<.*?>+\', \'\',t
ext)\n    text = re.sub(\'[%s]\' % re.escape(string.punctuation), \'\',tex
t)\n    text = re.sub(\'\n\', \'\',text)\n    text = re.sub(\'\\w*\\d\\w*
\',\'\',text)\n    return text'
```

In [23]:

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

In [24]:

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

Assigning the values for the model training purpose

In [25]:

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

We need to convert the data from text form to numeric form in order to get the model to understand it

In [26]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorization = TfidfVectorizer()
xv_train = vectorization.fit_transform(x_train)
xv_test = vectorization.transform(x_test)
```

Creating the Logistic Regression Algorithm

In [27]:

```
from sklearn.linear_model import LogisticRegression
LR =LogisticRegression()
LR.fit(xv_train,y_train)
```

Out[27]:

```
LogisticRegression()
```

In [28]:

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

In [29]:

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

Out[29]:

```
0.9868092691622103
```

In [30]:

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

```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5823
           1       0.99      0.99      0.99      5397

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

Logistic Regression is giving the accuracy of 98% while testing the dataset and 99% accuracy in classification report

Creating the Decision Tree algorithm

In [31]:

```
from sklearn.tree import DecisionTreeClassifier

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

Out[31]:

```
DecisionTreeClassifier()
```

In [32]:

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

In [33]:

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

Out[33]:

```
0.9958110516934047
```

In [34]:

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

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      5823
           1       1.00      1.00      1.00      5397

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

Decision Tree is giving the accuracy of 99.5% while testing the dataset and 100% accuracy in classification report which is very good

Creating the Gradient Boosting algorithm

In [35]:

```
from sklearn.ensemble import GradientBoostingClassifier

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

Out[35]:

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

In [36]:

```
pred_gb = GB.predict(xv_test)
```

In [37]:

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

Out[37]:

```
0.9951871657754011
```

In [38]:

```
print(classification_report(y_test,pred_gb))
```

```
              precision    recall  f1-score   support

           0       1.00      0.99      1.00      5823
           1       0.99      1.00      1.00      5397

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

Gradient Boosting algorithm is giving the accuracy of 99.5% while testing the dataset and 99% accuracy in classification report which is very good

Creating the Random Forest algorithm

In [39]:

```
from sklearn.ensemble import RandomForestClassifier

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

Out[39]:

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

In [40]:

```
pred_rf = RF.predict(xv_test)
```

In [41]:

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

Out[41]:

```
0.9836007130124778
```

In [42]:

```
print(classification_report(y_test,pred_rf))
```

```
              precision    recall  f1-score   support

           0       0.98      0.99      0.98      5823
           1       0.99      0.98      0.98      5397

    accuracy                           0.98     11220
   macro avg       0.98      0.98      0.98     11220
weighted avg       0.98      0.98      0.98     11220
```

Random Forest algorithm is giving the accuracy of 98% while testing the dataset and 98% accuracy in classification report

Overall, All the models are performing very well this dataset and giving very high accuracy as the dataset does not have any null values and looks well.

The Decision Tree is giving teh highest accuracy amongst all algorithms hence, we can use this algorithm for the model`

Testing the model with new observations

In [59]:

```
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_GB = GB.predict (new_x_test)
    pred_RF = RF.predict (new_xv_test)
    return print("\n\nLR Prediction: {} \nDT Prediction: {} \nGBC Prediction: {} \nRFC Pr
```

In [62]:

```
data
```

Out[62]:

	text	class
0	geneva reuters in a thinly veiled reference t...	1
1	rebel pundit decided to follow a local activis...	0
2	a wisconsin judge has refused to order local o...	0
3	was o reilly outofbounds most americans would ...	0
4	across the nation mainly the south there have ...	0
...	...	...
44873	clearly forgetting that we re in the year and...	0
44874	bill o reilly isn t my favorite and can be bel...	0
44875	washington reuters republican presidentelect ...	1
44876	isis stormed into mosul iraq last summer much ...	0
44877	united nations reuters united nations secreta...	1

44878 rows × 2 columns

In [*]:

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

The model is detecting the fake news accurately most of the times as expected

In [ ]:

```
```