FLIP ROBO

FAKE NEWS CLASSIFICATION PROJECT

Submitted by:

TAMALI SAHA

FlipRobo SME:

GULSHANA CHAUDHARY

ACKNOWLEDGMENT

I would like to express my special gratitude to Flip Robo Technologies team, who has given me this opportunity to deal with this dataset during my internship. It helped me to improve my analyzation skills. I want to express my gratitude to Ms. Gulshana Chaudhary (SME, Flip Robo) as she has helped me to get out of all the difficulties I faced while doing the project. I also want to give huge thanks to entire DataTrained team.

Bibliography:

Reference used in this project:

1. Hands on Machine learning with scikit learn and tensor flow by Aurelien Geron.
2. Andrew Ng Notes on Machine Learning (GitHub).
3. Different projects on Github and Kaggle.
4. Different conference papers on Recharchgate.
5. Fake News Detection on Social Media: A Data Mining Perspective, KaiShu, Amy Sliva
6. How to detect Fake news with Machine Learning by Matt Clarke, Practical data Science.
7. Fake News Detection Using Machine Learning Algorithms by Uma Sharma, Sidarth Saran, Shankar M. Patil

1. Introduction

1.1 Business Problem Framing

For both printed and digital media, the reliability of information has long since become a problem that has an impact on society and business. Because of how quickly and magnified information spreads on social networks, faulty or misleading material has a great chance of having a real-world influence on millions of users in a matter of minutes. Many public concerns regarding this issue, as well as proposed solutions, have recently been voiced.

1.2 Conceptual Background of the Domain Problem

One of the biggest issues of our time is fake news. It has a significant impact on both our offline and online conversation. One can even argue that false news currently represents a clear and present threat to the democracy and social stability of the West. Fake news has been around for a very long time, almost as long as it took for news to become widely disseminated following the invention of the printing press in 1439. There is no universally accepted definition of what constitutes "fake news," despite the fact that there has been a long history of fake news on social media. Appropriate clarifications are required in order to effectively direct the future paths of false news detection research.

1.3 Review of Literature

The widespread dissemination of false information may have a detrimental effect on both people and society. First, fake news has the potential to upset the ecosystem's delicate balance of authenticity. For instance, it is clear that during the U.S. 2016 presidential election, the most popular false news was even more widely disseminated on Facebook than the most popular mainstream news.

The extensive spread of fake news has the potential for extremely negative impacts on individuals and society. [5]

In the project **How to detect Fake news with Machine Learning by Matt Clarke**, apply NLP and machine learning techniques to see how hard it is to identify fake news from real news. It is really that hard for social networks to identify and flag disinformation with a high degree of accuracy. They use the same approach to create a sarcasm detection model. [6]

Fake news intentionally persuades consumers to accept biased or false beliefs. Fake news is usually manipulated by propagandists to convey political messages or influence. For example, some report shows that Russia has created fake accounts and social bots to spread false stories.

In the paper **Fake News Detection Using Machine Learning Algorithms by Uma Sharma, Sidarth Saran, Shankar M. Patil** explains the system which is developed in three parts. The first part is static which works on machine learning classifier. They studied and trained the model with 4 different classifiers and chose the best classifier for final execution. The second part is dynamic which takes the keyword/text from user and searches online for the truth probability of the news. The third part provides the authenticity of the URL input by user [7].

1.4 Motivation for the Problem Undertaken

The project is provided to me by Flip Robo Technologies as a part of the internship programme (Internship Batch No-31). This problem is a real world dataset. The exposure of this data gives me the opportunity to locate my skills in solving a real time problem. It is the primary motivation to solve this 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.

This study aims to analyse and predicting malignant comment when using **Classification Model** like Logistic Regression, Random Forest etc. Thus, the purpose of this study is to grow the knowledge of **Classification methods** in machine learning fields. Those are the different factors to undertaken the problem for studypurpose.

2. Analytical Problem Framing

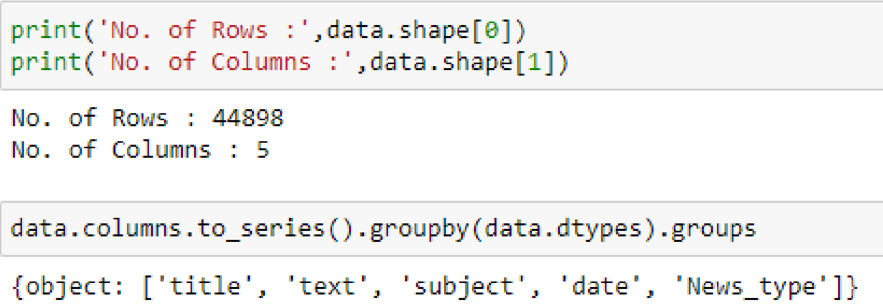
2.1 Mathematical/ Analytical Modelling of the Problem

The problem is of Binary Classification. We have to classify is 0 if the news in fake or 1 if the news is fake. We have 44898 news records in the dataset. There are some missing values in the dataset, but are very less compared to the density of the dataset. There are in total 5 features in the dataset. Create word dictionary and word cloud for further and future

Here 4 different algorithm are used and final model is chosen by best AUC-ROC score and accuracy score.

2.2 Data Sources and their formats

There are one set of data. The dataset has 44898 rows and 5 columns. Primarily it has 5 columns where some columns are unnecessary.
Later the dataset is divided into two parts, training and testing. After determine the proper model, the model is applied to predict the target variable for the test dataset.



```
print('No. of Rows :',data.shape[0])
print('No. of Columns :',data.shape[1])
```

```
No. of Rows : 44898
No. of Columns : 5
```

```
data.columns.to_series().groupby(data.dtypes).groups
```

```
{object: ['title', 'text', 'subject', 'date', 'News_type']}
```

2.3 Data Pre-processing Done:

2.3.1 Feature Engineering:

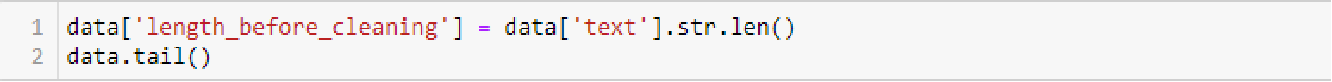
'--', 'null', 'NA', ' ' are present in the dataset. Need to remove those from the dataset.

2.3.2 Drop unnecessary columns:

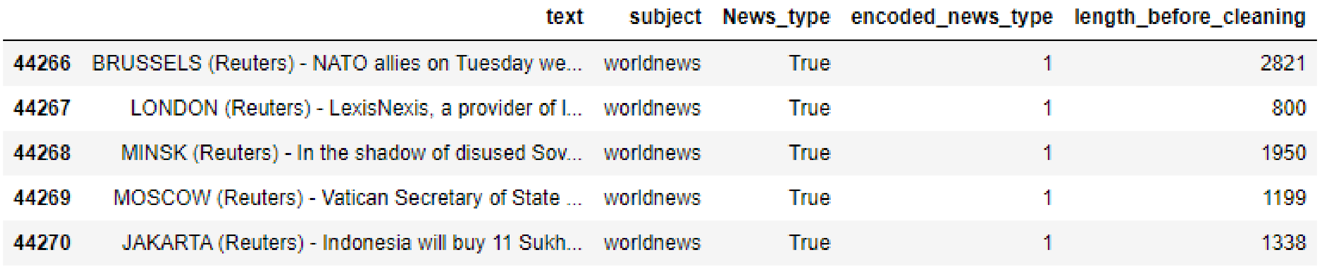
Drop the unnecessary column from the dataset.

2.3.3 Calculate length before cleaning of ‘massage_body’ column:

Let’s calculate the comment length before cleaning.



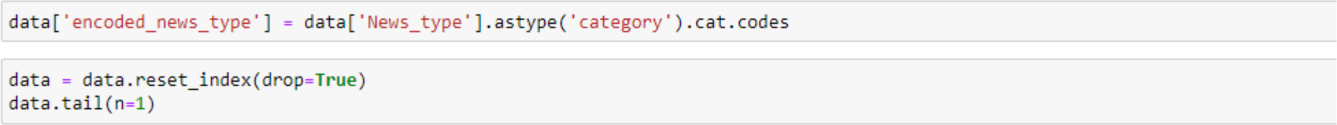
```
data['length_before_cleaning'] = data['text'].str.len()
data.tail()
```



	text	subject	News_type	encoded_news_type	length_before_cleaning
44266	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	True	1	2821
44267	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	True	1	800
44268	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	True	1	1950
44269	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	True	1	1199
44270	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	True	1	1338

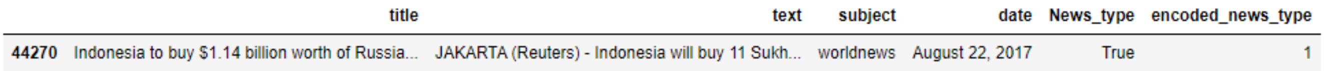
2.3.4 Encoding type dataset:

Here ‘news type’ dataset is object datatype. Let’s encoded the type columns as true=1 and fake=0 for further progress. Then reset the index.



```
data['encoded_news_type'] = data['News_type'].astype('category').cat.codes
```

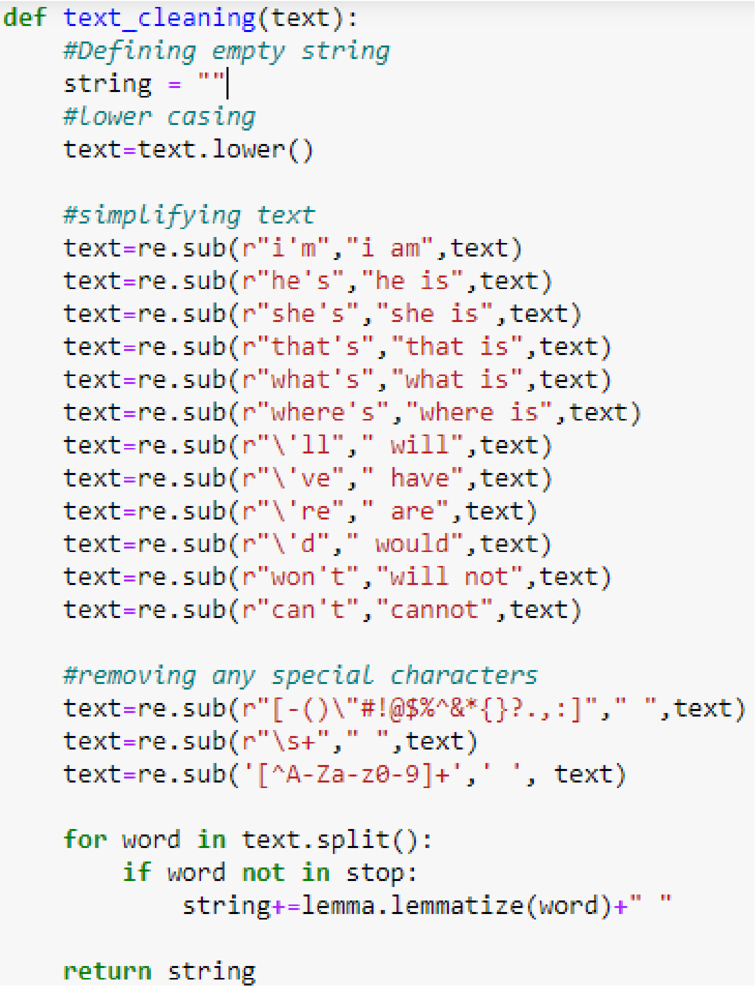
```
data = data.reset_index(drop=True)
data.tail(n=1)
```



	title	text	subject	date	News_type	encoded_news_type
44270	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	True	1

2.3.5 Natural Language Processing:

Import all necessary libraries for NLP. Now let’s remove all alphabets, numbers from the text body. Then apply Stemmer, Stop words and such necessary NLP steps on massage body for cleaning the dataset.



```python
def text_cleaning(text):
    #Defining empty string
    string = ""
    #lower casing
    text=text.lower()

    #simplifying text
    text=re.sub(r"i'm","i am",text)
    text=re.sub(r"he's","he is",text)
    text=re.sub(r"she's","she is",text)
    text=re.sub(r"that's","that is",text)
    text=re.sub(r"what's","what is",text)
    text=re.sub(r"where's","where is",text)
    text=re.sub(r"\'ll"," will",text)
    text=re.sub(r"\'ve"," have",text)
    text=re.sub(r"\'re"," are",text)
    text=re.sub(r"\'d"," would",text)
    text=re.sub(r"won't","will not",text)
    text=re.sub(r"can't","cannot",text)

    #removing any special characters
    text=re.sub(r"[-()\"#!@$%^&*{}?.,:]"," ",text)
    text=re.sub(r"\s+"," ",text)
    text=re.sub('[^A-Za-z0-9]+',' ', text)

    for word in text.split():
        if word not in stop:
            string+=lemma.lemmatize(word)+" "

    return string
```

Apply all necessary steps on massage body column for cleaning the dataset. The steps are as follows.

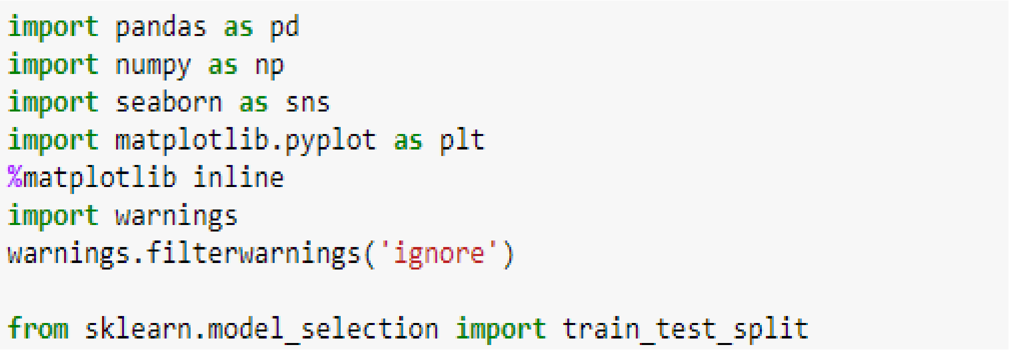
1. First lowercased strings from the given string by converting each uppercase character to lowercase for the full news text.
2. Now simplifying text by converting different short form of words (contractions) to the actual words.
3. Then removing any special characters.
4. By using WordNetLemmatizer, lemmatize the news text.

2.4 State the set of assumptions (if any) related to the problem under consideration

No such assumptions are taken for this case.

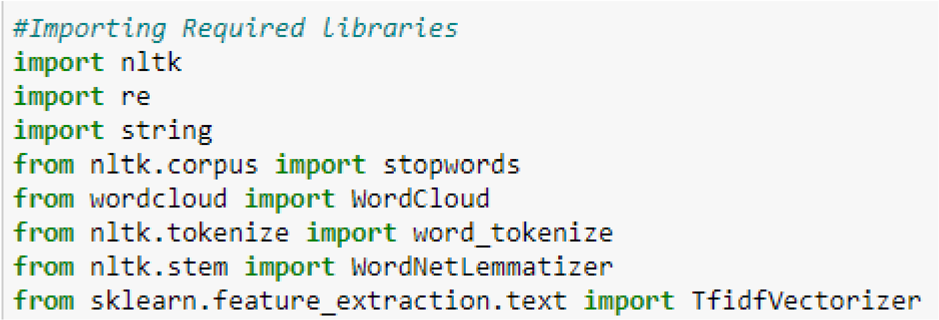
2.5 Hardware and Software Requirements and Tools Used

Processor: Intel(R) Core(TM) i3-5005U CPU @ 2.00GHz 2.00 GHz
RAM: 4.00 GB
System Type: 64-bit operating system, x64-based processor
Window: Windows 10 Pro
Anaconda – Jupyter Notebook
Libraries Used –



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

from sklearn.model_selection import train_test_split
```

For Word-Cloud the following libraries are used.



```
#Importing Required libraries
import nltk
import re
import string
from nltk.corpus import stopwords
from wordcloud import WordCloud
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import TfidfVectorizer
```

Except this, different libraries are used for machine learning model building from sklearn.

3. Model/s Development and Evaluation

3.1 Identification of possible problem-solving approaches (methods):

In this problem Classification-based machine learning algorithm like logistic regression can be used. Removed any excess spaces, changed the email addresses subject line to a phone number that is probably wise, etc. For building an appropriate ML model before implementing classification algorithms, data is split in training & test data using train_test_split. Then different statistical parameter like accuracy score, confusion matrix, classification report, precision, recall etc. are determined for every algorithm. Hyper parameter tuning is performed to get the accuracy score much higher and accurate than earlier.

Then the best model is chosen from 4 different algorithm.

3.2 Testing of Identified Approaches (Algorithms)

Total 4 algorithms used for the training and testing are:

1. Logistic Regression
2. Linear SVC
3. Decision Tree Classifier
4. Multinomial Naive Bayes

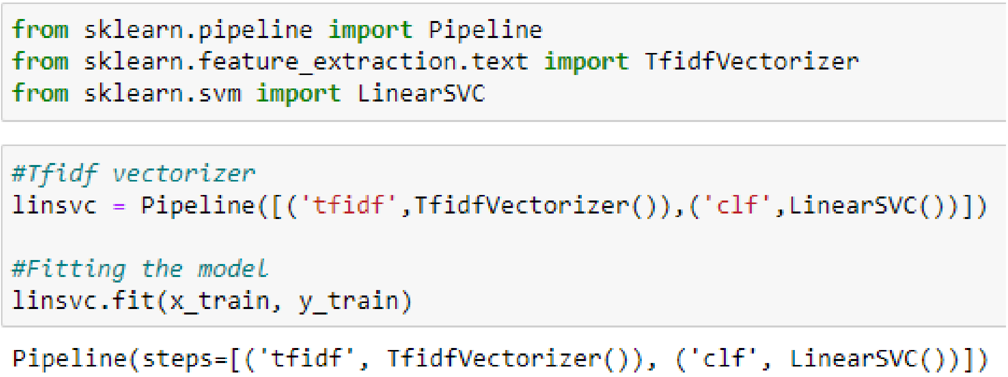
3.3 Key Metrics for success in solving problem under consideration:

From metrics module of sklearn library import classification_report, accuracy_score, confusion_matrix, classification_report and f1_score. From model_selection also, we use cross_val_score. Those are the matrices use to validate the model's quality. Let's discuss every metrics shortly.

- Classification report: It is a performance evaluation metric in machine learning which is used to show the precision, recall, F1 Score, and support score of your trained classification model
- Accuracy score: It is used when the True Positives and True negatives are more important. Accuracy can be used when the class distribution is similar.
- Confusion Matrix: It is a table that is used in classification problems to assess where errors in the model were made. The rows represent the actual classes the outcomes should have been. While the columns represent the predictions we have made. Using this table it is easy to see which predictions are wrong.
- Precision: It can be seen as a measure of quality. If the precision is high, an algorithm returns more relevant results than irrelevant ones.
- Recall: The recall is calculated as the ratio between the numbers of Positive samples correctly classified as Positive to the total number of Positive samples.
- F1 Score: F1 = 2 * (precision * recall) / (precision + recall)

3.4 Run and Evaluate selected models

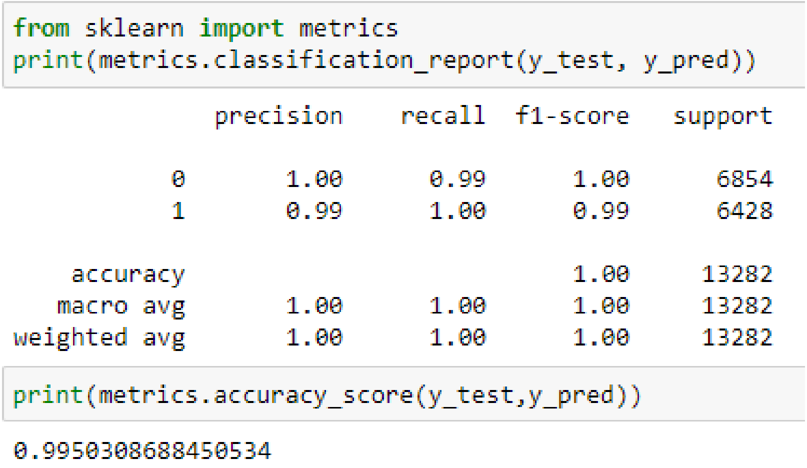
A Linear SVC:



```
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import LinearSVC
```

```
#Tfidf vectorizer
linsvc = Pipeline([('tfidf',TfidfVectorizer()),('clf',LinearSVC())])

#Fitting the model
linsvc.fit(x_train, y_train)
```

```
Pipeline(steps=[('tfidf', TfidfVectorizer()), ('clf', LinearSVC())])
```



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

```
              precision    recall  f1-score   support

           0       1.00      0.99      1.00      6854
           1       0.99      1.00      0.99      6428

    accuracy                           1.00     13282
   macro avg       1.00      1.00      1.00     13282
weighted avg       1.00      1.00      1.00     13282
```

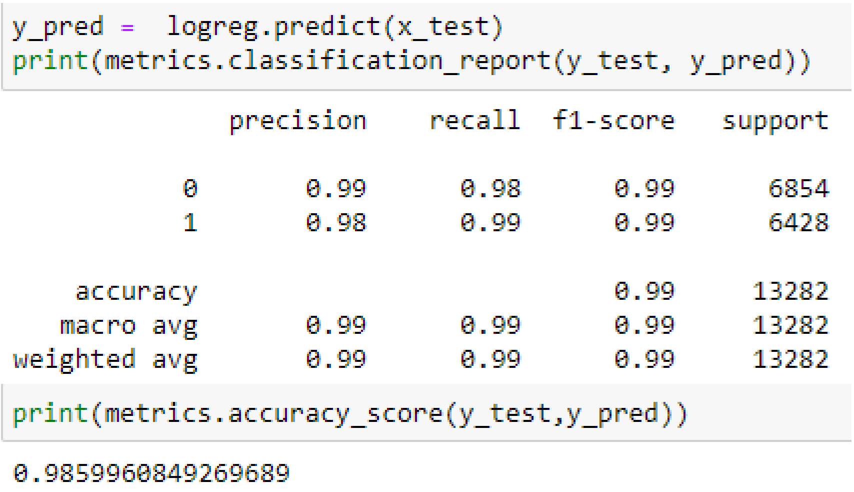
```
print(metrics.accuracy_score(y_test,y_pred))
```

```
0.9950308688450534
```

In this way accuracy score is determined for each 4 different classification model.

B Logistic Regression:

The accuracy score, confusion matrix and classification report after using logistic regression is as follows.



```
y_pred =  logreg.predict(x_test)
print(metrics.classification_report(y_test, y_pred))
```

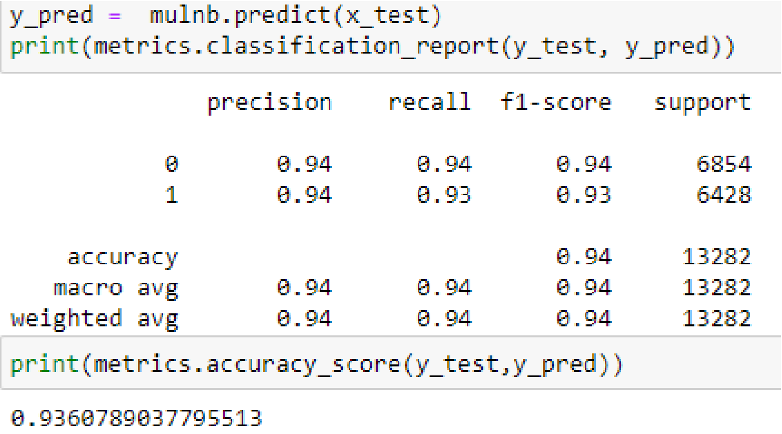
```
              precision    recall  f1-score   support

           0       0.99      0.98      0.99      6854
           1       0.98      0.99      0.99      6428

    accuracy                           0.99     13282
   macro avg       0.99      0.99      0.99     13282
weighted avg       0.99      0.99      0.99     13282
```

```
print(metrics.accuracy_score(y_test,y_pred))
```

```
0.9859960849269689
```

C MultinomialNB:

The accuracy score, confusion matrix and classification report after using Multinomial NB is as follows.



```
y_pred =  mulnb.predict(x_test)
print(metrics.classification_report(y_test, y_pred))
```

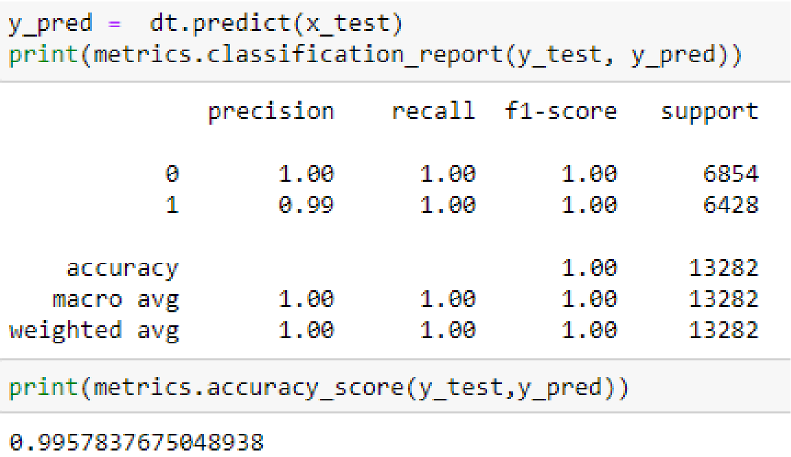
```
              precision    recall  f1-score   support

           0       0.94      0.94      0.94      6854
           1       0.94      0.93      0.93      6428

    accuracy                           0.94     13282
   macro avg       0.94      0.94      0.94     13282
weighted avg       0.94      0.94      0.94     13282
```

```
print(metrics.accuracy_score(y_test,y_pred))
```

```
0.9360789037795513
```

D DecisionTree Classifier:

The accuracy score, confusion matrix and classification report after using DecisionTreeClassifier is as follows.



```
y_pred =  dt.predict(x_test)
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      6854
           1       0.99      1.00      1.00      6428

    accuracy                           1.00     13282
   macro avg       1.00      1.00      1.00     13282
weighted avg       1.00      1.00      1.00     13282
```

```
print(metrics.accuracy_score(y_test,y_pred))
```

```
0.9957837675048938
```

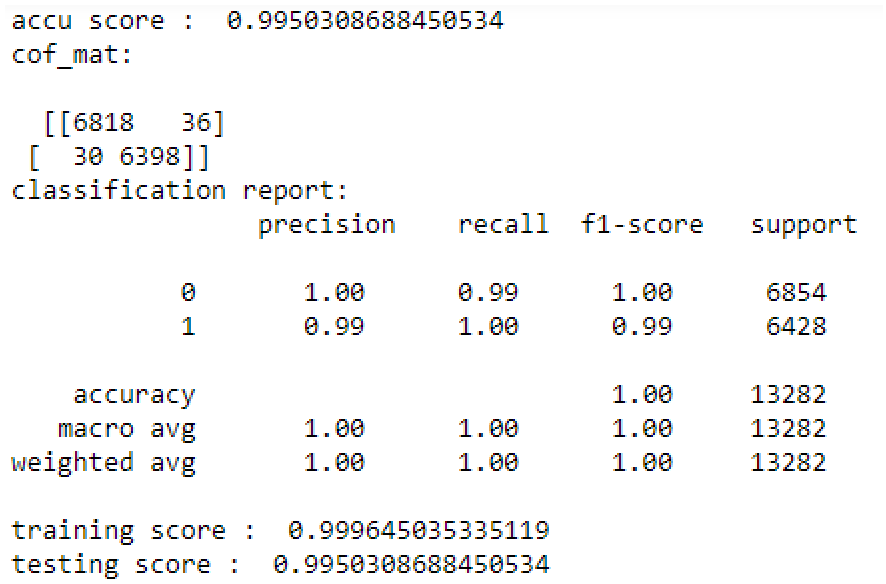
As per 4 different model, Both DecisionTreeClassifier and LinearSVC is good for this particular dataset.

3.5 AUC- ROC Curve:

By AUC- ROC curve also every 4 models are good. Here we take linearSVC as final one.

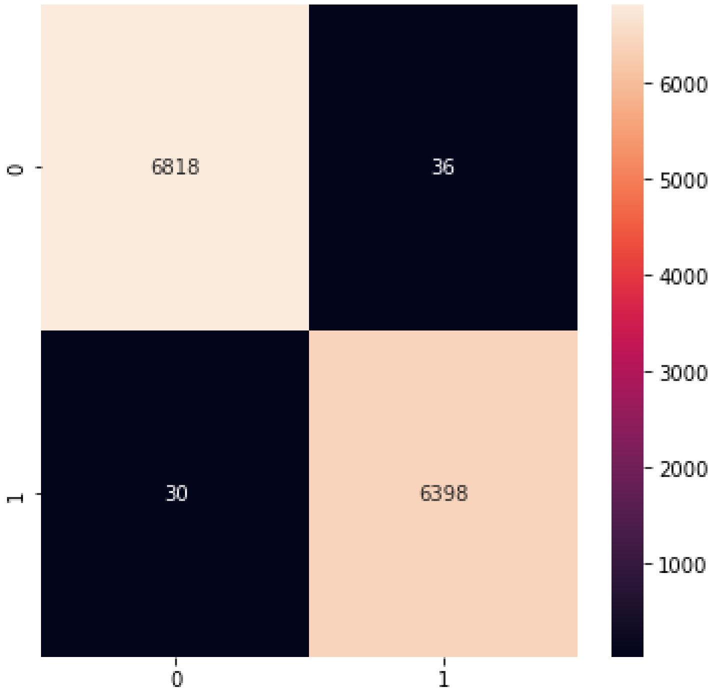
3.6 Final Model:

For final model the target variable is as follows after using linearSVC.



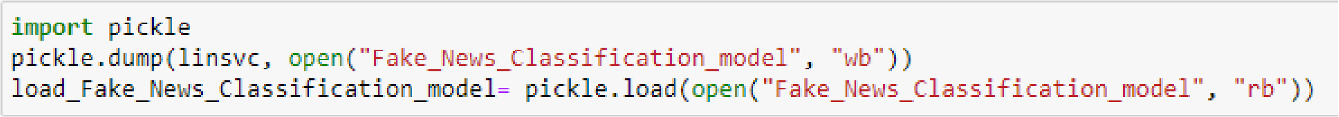
```
accu score :  0.9950308688450534
cof_mat:

  [[6818   36]
 [   30 6398]]
classification report:
                precision    recall  f1-score   support

           0       1.00      0.99      1.00      6854
           1       0.99      1.00      0.99      6428

    accuracy                           1.00     13282
   macro avg       1.00      1.00      1.00     13282
weighted avg       1.00      1.00      1.00     13282

training score :  0.999645035335119
testing score :  0.9950308688450534
```

3.7 Confusion Matrix:

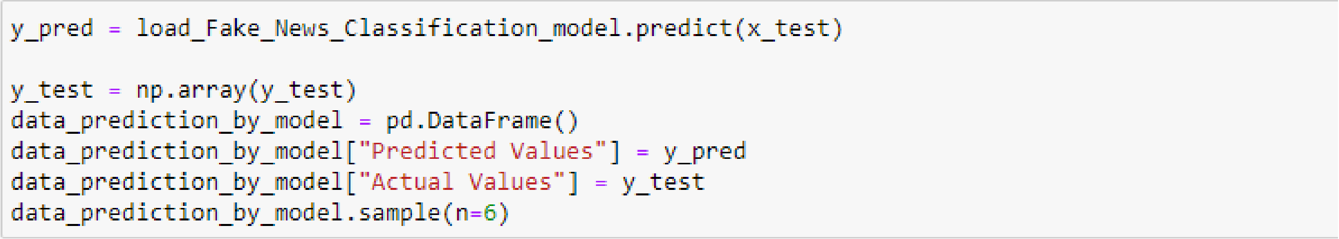


3.8 Load the model:

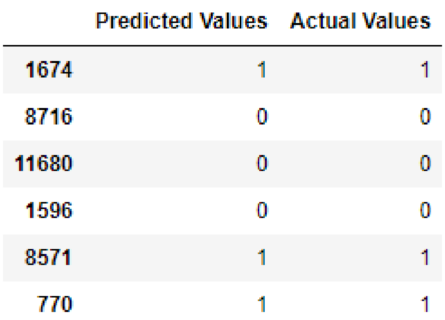
Load the model for further use using pickle. After final modelling let’s see the 6 random dataset of actual and predicted target.



```
import pickle
pickle.dump(linsvc, open("Fake_News_Classification_model", "wb"))
load_Fake_News_Classification_model= pickle.load(open("Fake_News_Classification_model", "rb"))
```



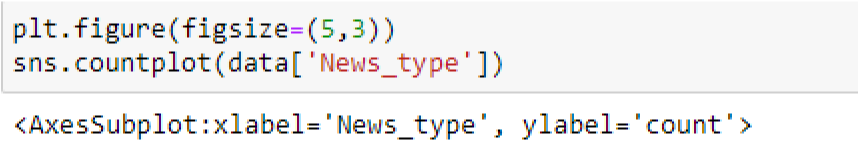
```
y_pred = load_Fake_News_Classification_model.predict(x_test)

y_test = np.array(y_test)
data_prediction_by_model = pd.DataFrame()
data_prediction_by_model["Predicted Values"] = y_pred
data_prediction_by_model["Actual Values"] = y_test
data_prediction_by_model.sample(n=6)
```



	Predicted Values	Actual Values
1674	1	1
8716	0	0
11680	0	0
1596	0	0
8571	1	1
770	1	1

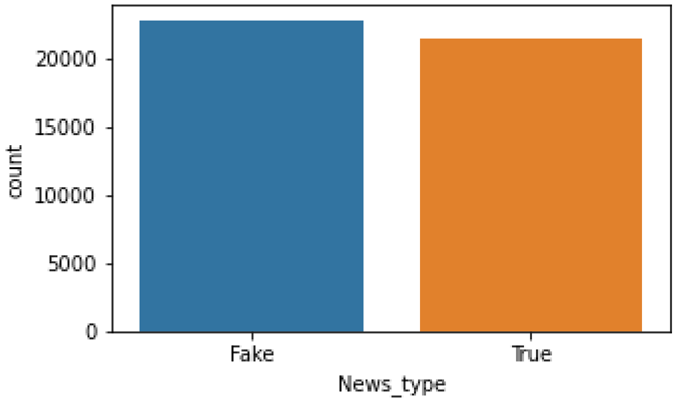
3.9 Visualizations:

Let’s start the observation exploration of feature analysis. Visualize the dataset before cleaning for fake and true news.

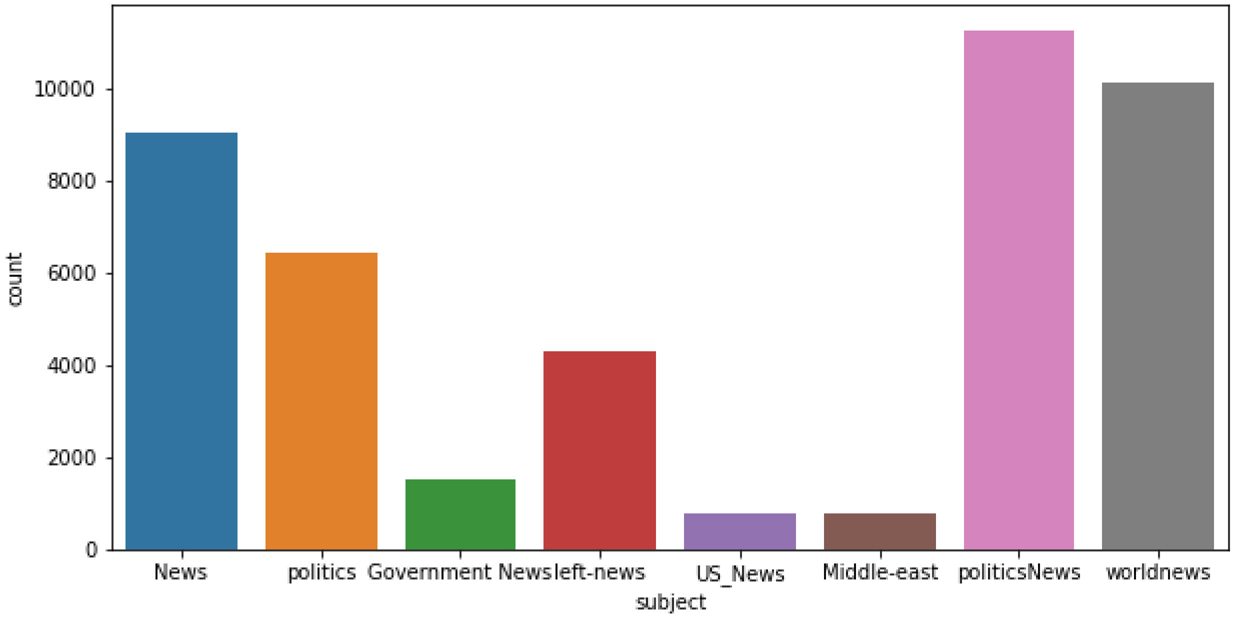


```
plt.figure(figsize=(5,3))
sns.countplot(data['News_type'])
```

```
<AxesSubplot:xlabel='News_type', ylabel='count'>
```

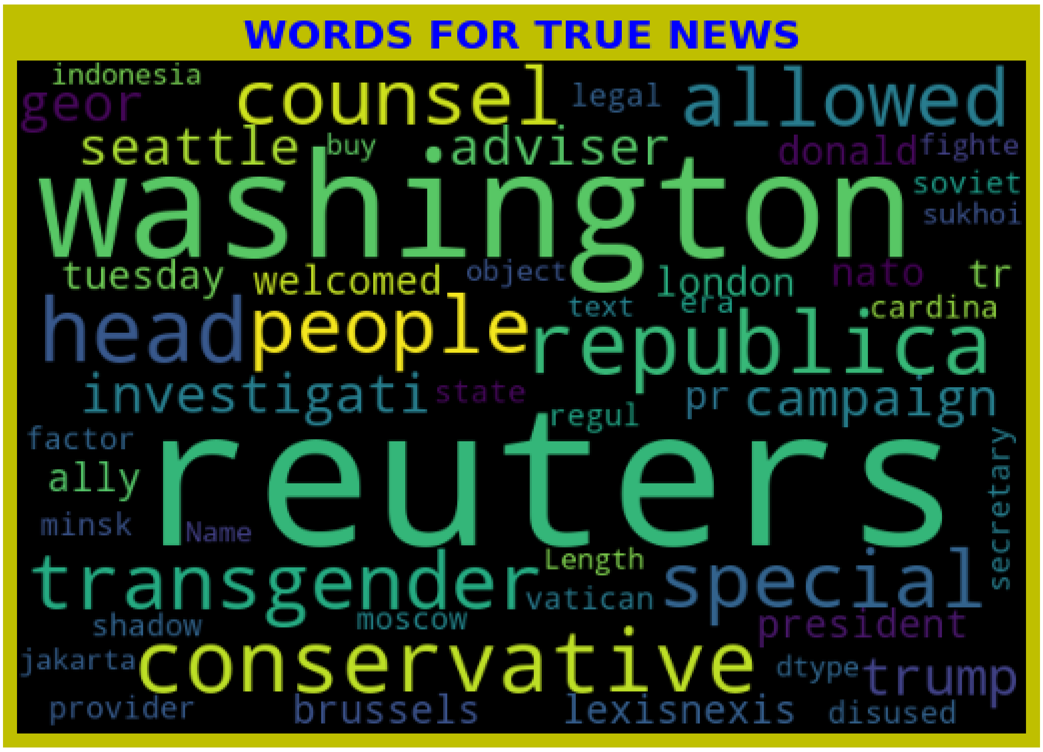


There are two type of News_type, True and False. The percentage of two different type is around 50%. We can say that, the dataset is totally balanced.



There are total 8 types of subject. Of them , politicsNews type is maximum(25%) and Middle-east is minimum (just 1%).

3.10 Word Cloud for Spam sms:



WORDS FOR FAKE NEWS
21st century
christmas
21wire
wire
say
donald
trump
day
intelligence
new
american
house
friday
revealed
leav
former
milwaukee
committee
happy
wish
devin
david
chairman
year
sheriff
announced
annual
used
patrick
nu
francis
Length
henningsen
wireremember
earlier
dtype
whenever
text
go
jazeera
object
hi
Name
pope
yea
theme
reported
bac
message
america
predicted
al
familiar

3.11 Interpretation of the Results

After all the pre-processing steps, the dataset is ready to train machine learning models. All unnecessary words from comment text are deleted as they might give overfitting problem as well as it also could increase the time complexity. Now apply this dataset on different ML Classification Model (as discussed on part 3.4 - ‘Run and Evaluate selected models’) and check the best model for this particular dataset.

4. CONCLUSION

4.1 Key Findings and Conclusions of the Study

Here, we observed the various fake and true news. We have also seen how easily accessible algorithms can be used in this way to handle this difficulty. It was shown in our particular investigation that a logistic regression solution offers a significant improvement in classification compared to any other approach.

4.2 Learning Outcomes of the Study in respect of Data Science

Data cleansing is one of the most crucial phases; I attempted to make comments shorter and included all the relevant keywords in it. The power of visualization is beneficial for converting data into a graphical representation; it helps me to comprehend what the data is trying to communicate.

In this dataset, I utilized a variety of methods to find the best result and preserve that model. The use of NLP method is very helpful to determine the Fake news in different aspects.

4.3 Limitations of this work and Scope for Future Work

Additionally, the following studies are examples that might be taken into account for future work in this field:

- We offer the following strategy to enhance NLP classifiers: Convolutional neural networks (CNN) and Support vector clustering (SVC) are two additional algorithms that can be used to enhance the performance of existing classifiers. In the present study, the issue was reduced to two classes, although it is worthwhile to pursue the primary objective of six classes of remarks.
- For text processing and classification, we also advocate the use of better NLP methods.