FLIP ROBO

Fake News Classifier

Submitted by:

ASHUTOSH MISHRA

ACKNOWLEDGMENT

I'd like to extend my gratitude to my mentor Mr. Md. Kashif for giving me this opportunity to work upon this project. Below are all the details of the project which I consumed while preparing and drafting the project –

The references, research papers, data sources, professionals, etc are majorly referred from "https://www.educative.io/answers/preprocessing-steps-in-natural-language-processing-nlp" study collaterals and data repository. Also, some of the other resources like "Ineuron" were explored to gain a deep understanding on the assigned subject.

Above stated details stands correct to the best of my knowledge and I hereby acknowledge the same.

INTRODUCTION

The issue of fake news has been a dominant theme in the headlines for several years. Fake news are those news stories that are false and the story itself is fabricated, with no verifiable facts, sources or quotes. Sometimes these stories may be propaganda that is intentionally designed to mislead the reader, or may be designed as click-bait written for economic incentives (the writer profits on the number of people who click on the story).

However, it's important to acknowledge that fake news is a complex and nuanced problem, one that is far greater than the narrow definition above. The term itself has become politicized, and is widely used to discredit any opposing viewpoint. Some people use it to cast doubt on their opponents, controversial issues or the credibility of some media organizations. In addition, technological advances such as the advent of social media enable fake news stories to proliferate quickly and easily as people share more and more information online. Increasingly, we rely on online information to understand what is happening in our world.

The universe of fake news is much larger than simply false news stories. Some stories may have a nugget of truth, but lack any contextualizing details. They may not include any verifiable facts or sources. Some stories may include basic verifiable facts, but are written using language that is deliberately inflammatory, leaves out pertinent details or only presents one viewpoint. "Fake news" exists within a larger ecosystem of misinformation and disinformation.

Business Problem Framing

According to the some researchers, Fake news has become one of the biggest problems of our age. It has serious impact on our online as well as offline discourse. 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.

Here we are looking to build a Fake News detection model as having said that for different media outlets, the ability to attract viewers to their websites is necessary to generate online advertising revenue. So it is really necessary to detect the fake news, so that a person can let themselves aware about the same.

See, the demand to build such models is increasing rapidly as almost most of the industries, individuals and entities are facing a big challenge regarding the same and to tackle with it they want to be proactive by identifying such highly fake utterances and in-order to do so NLP and Deep Learning could be that uprising thing to predict and classify these fake news.

Our goal is to build a prototype of Fake News Classifier which can be used to classify fake and true news, so that it can get controlled and restricted from spreading any antipathy or hostile news.

Conceptual Background of the Domain Problem

Natural Language Process (NLP) and Deep Learning comes as a vital tool to solve almost any type of business problems to help companies optimize the standards resulting in overall revenue and profits growth. Moreover, it also improves their marketing strategies and demands focus on changing trends.

Machine learning data only works with numerical features so we have to convert text data into numerical columns. So we have to preprocess the text and that is called natural language processing.

Now, if we talk about the Fake News Classifier project what I feel is that LSTM RNN would be the key recurrent neural network in identifying these fake news as the news title would be having too many words in each documents of its corpus and definitely it will help us to preserve the Long Term- Short Term Memory of each words in the system, so that we can get genuine prediction.

Also, implementation of Deep Learning and Deep NLP techniques like- Stemming, Stopwords, Tensorflow, Keras, Embedding, LSTM, Dense, Dropout, One_Hot encoder, Pad_Sequence, Sequential and Regular Expression would be really helpful while predicting the target variable. So, here in this project I'll be using NLP with Deep Learning.

Review of Literature

There has been a remarkable increase in the cases of Fake news on various social media platforms due to the forefront technology advancement. Many celebrities and politicians are facing backlashes from these hoaxes and have to come across hostile and offensive comments. This can take a toll on anyone and affect them mentally leading to depression, mental illness, self-hatred and even suicidal thoughts. Our goal is to build a prototype of fake news classifier which can be used to classify fake and hoax news so that it can be controlled and restricted from spreading hatred and antipathy.

Being a Data Scientist, I'll be using here Long term-Short term Memory Recurrent Neural Network (LSTM RNN) model to design and optimize the result going forward. Also the target variable is in the form of categorical label.

Initially two dataset were provided to me; one was **true news** dataset and other was **fake news** dataset. **True** news dataset consists of 21417 rows and 4 different attributes name as-title, text, subject, and date. Whereas the **fake** news dataset consist of 23481 rows and only 4 columns name as- title, text, subject, and date.

Eventually, the target variable was not given to us in the both the dataset. So, I've added a target variable name as **label** in both the datasets. For the True news I've put **1** in the label whereas **0** for Fake news label.

Once I have got our features and target variable I've appended the fake dataset with that of true dataset. So in overall our resultant dataset has 44898 rows and 5 columns as- title, text, subject, date, and label.

Also, I have observed that all the features were in object data type and two of them (title and text column) consists of long documents inside it and there are no null values are present in our dataset.

I'll be using here stemming technique not lemmatization as we are not making project for any particular text summarization or language translation where the output need to be well organized and meaningful.

Also, keras provides the one_hot() function that creates a hash/index of each word of a document as an efficient integer encoding. The Embedding has a vocabulary of 5000 and an input length of 50. We will choose an embedding space i.e. our output dimension as of 100 dimensions. The model is a simple binary classification model. The output from the embedding layer will be 50 vectors of 100 dimension each, one for each word.

As a result of the above LSTM-RNN modelling, I have managed to achieve below top model, which are given below as following–

Recurrent Neural Network	**Testing Accuracy (in %)**
LSTM RNN	98.4

Motivation for the Problem Undertaken

See, in my previous projects I've worked more upon Machine learning and NLP techniques but first time what I feel like that I can perform Deep NLP i.e. with the help of neural networks we can find out the accuracy of the model as well and this really excites and stimulate me as a zeal learner.

Also, this project guided me to baseline each aspect carefully and be concrete on the decision making process regardless it's an individual or an entity like a company.

In order to cater to the above project, my current knowledge and skill set has aided me a lot which I'd explore on this exponentially further.

Analytical Problem Framing

Here in this project I've to make prediction about Fake news and since we all know that the news are basically written in the form of text and the text size used to be very large in nature thereby here in this project I will be using LSTM-RNN Deep NLP technique for predicting the fake news.

I'll be using here stemming technique not lemmatization as we are not making project for any particular text summarization or language translation where the output need to be well organized.

TensorFlow is an open source framework developed by Google researchers to run machine learning, deep learning and other statistical and predictive analytics workloads. It develops neural networks faster and easier.

Being an Open-Source library for deep learning and machine learning, TensorFlow plays a role in text-based applications, image recognition, voice search, and many more.

For instance- DeepFace, Facebook's image recognition system, uses TensorFlow for image recognition. It is used by Apple's Siri for voice recognition.

06-Jan-23 7:58 PM Ashutosh Mishra

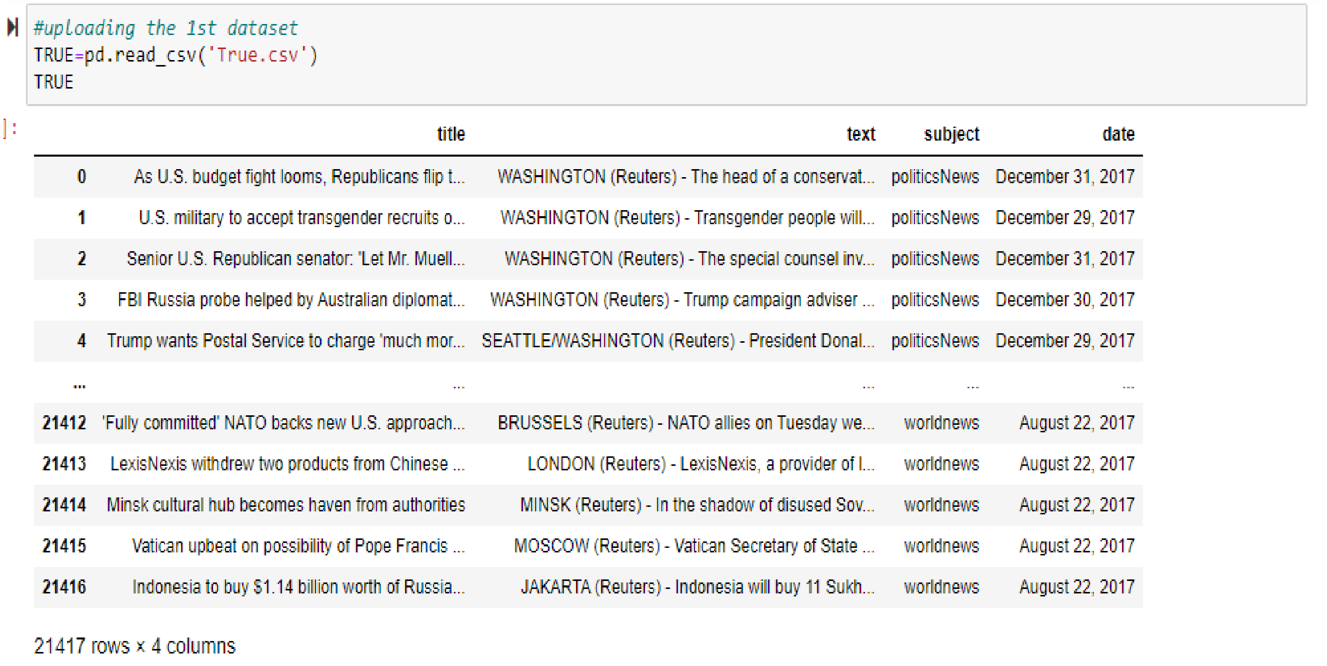
Keras is a high-level, deep learning API developed by Google for implementing neural networks.

Keras is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the TensorFlow library. It is written in Python and is used to make the implementation of neural networks easy. Tensorflow and Keras provide high-level APIs used for easily building and training models, but Keras is more user-friendly because it's built-in Python.

Data Sources and their formats

Eventually, I’ve been provided with two CSV datasets, but later I appended these into one.

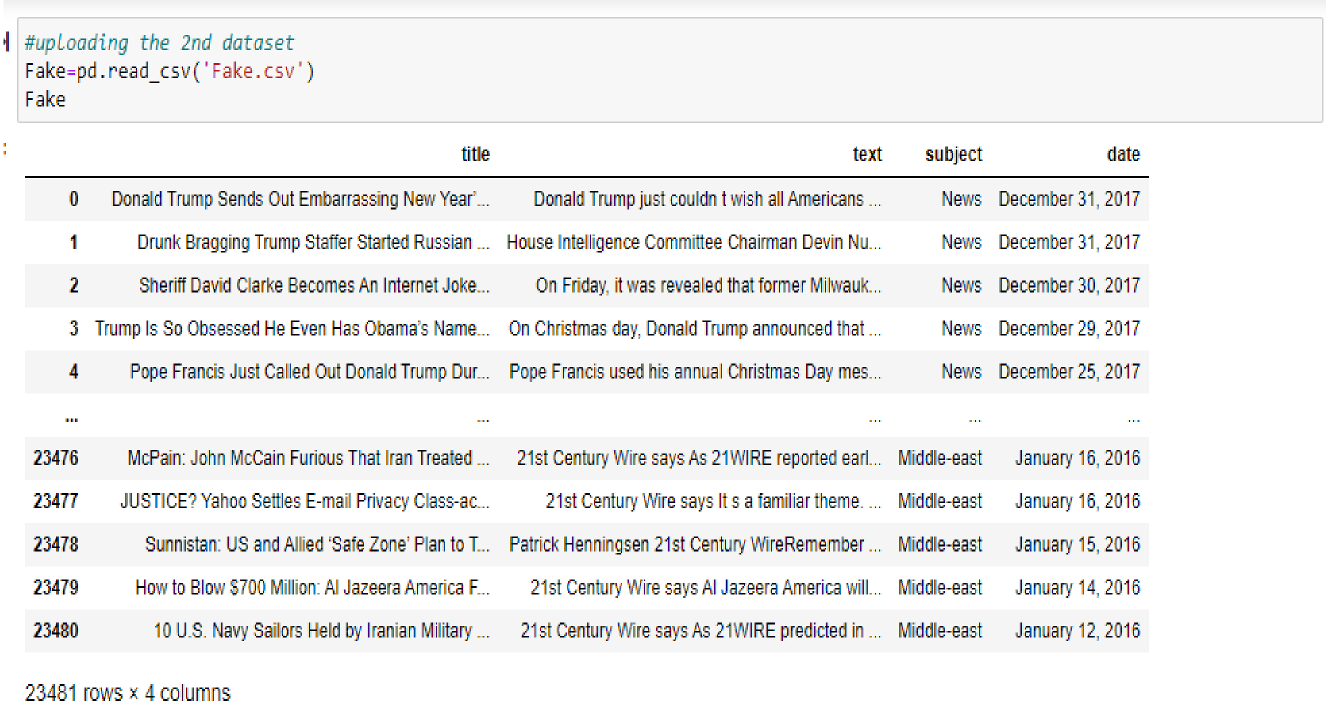
Let’s have a look on our dataset and their attributes.



```
#uploading the 1st dataset
TRUE=pd.read_csv('True.csv')
TRUE
```

	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

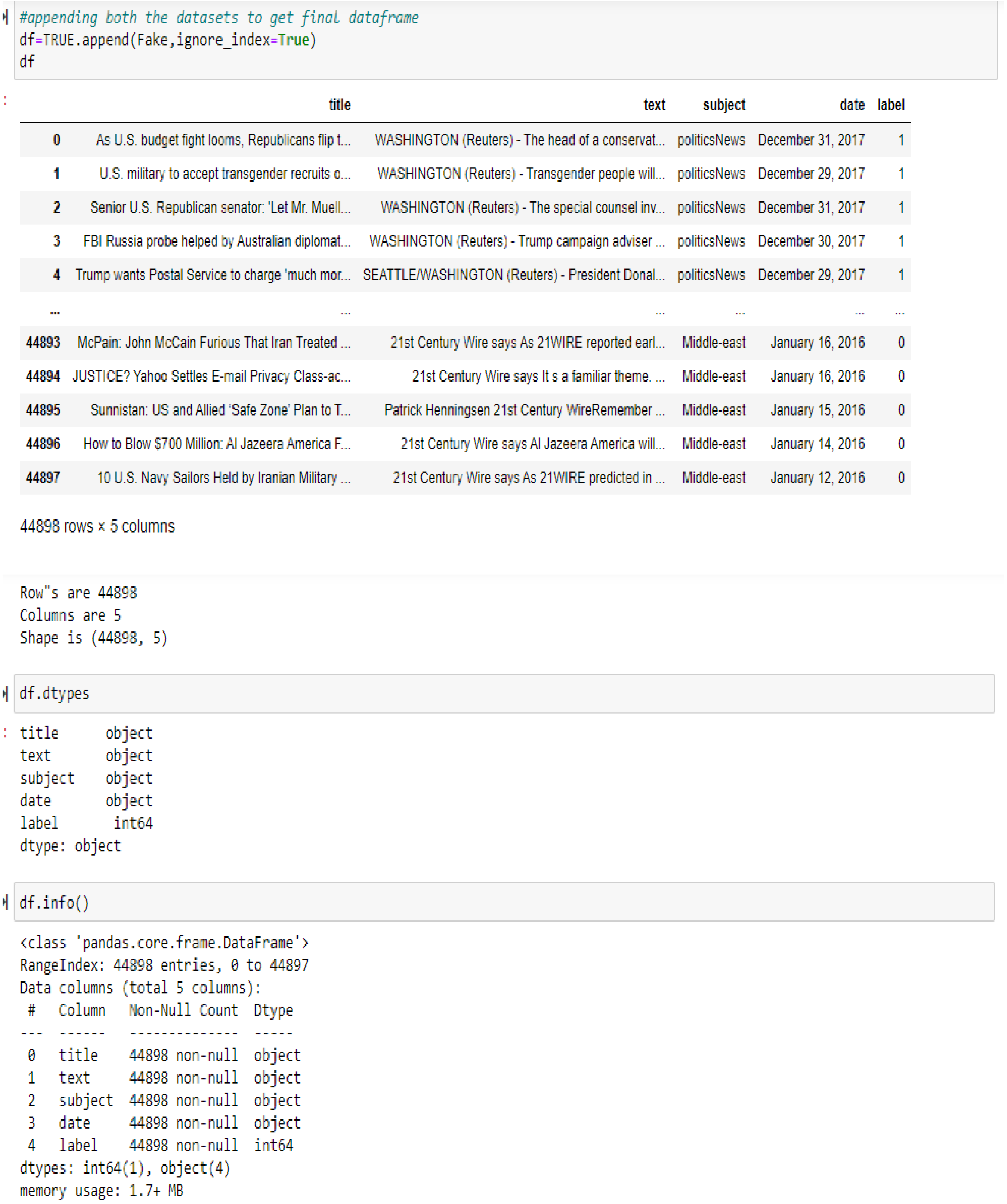
21417 rows × 4 columns



```
#uploading the 2nd dataset
Fake=pd.read_csv('Fake.csv')
Fake
```

	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



```
#appending both the datasets to get final dataframe
df=TRUE.append(Fake,ignore_index=True)
df
```

	title	text	subject	date	label
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1
...	...	...	...	...	...
44893	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	0
44894	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016	0
44895	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016	0
44896	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016	0
44897	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016	0

44898 rows × 5 columns

```
Row"s are 44898
Columns are 5
Shape is (44898, 5)
```

```
df.dtypes
```

```
title       object
text        object
subject     object
date        object
label        int64
dtype: object
```

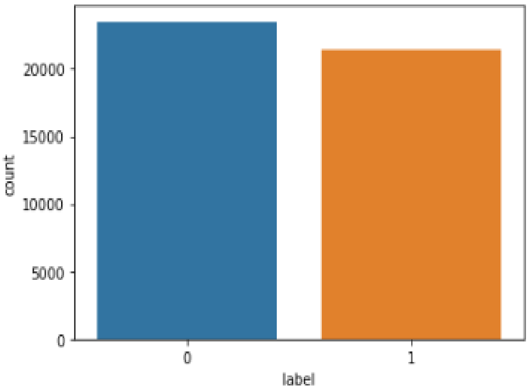
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 44898 entries, 0 to 44897
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   title    44898 non-null  object
 1   text     44898 non-null  object
 2   subject  44898 non-null  object
 3   date     44898 non-null  object
 4   label    44898 non-null  int64 
dtypes: int64(1), object(4)
memory usage: 1.7+ MB
```

This dataset doesn't have any null values in it. --------

```
#plotting our target variable 'label' to check it's count for each class
sns.countplot(df.label)
```

```
]: <AxesSubplot:xlabel='label', ylabel='count'>
```



```
df.label.value_counts()
```

```
]: 0    23481
   1    21417
   Name: label, dtype: int64
```

1 represents True news

0 represents Fake news

so in our dataset fake news are more in number as that of true news.

```
#extracting the information
df.loc[65] #This will give all the data present in 65th row
```

```
7]: title      Senator Cornyn trying to get Big Corn behind U...
    text       (Reuters) - Senator John Cornyn, the No. 2 Sen...
    subject                                         politicsNews
    date                                       December 20, 2017
    label                                                      1
    Name: 65, dtype: object
```

```
df.title.loc[49] #This will give title column data present in 49th row
```

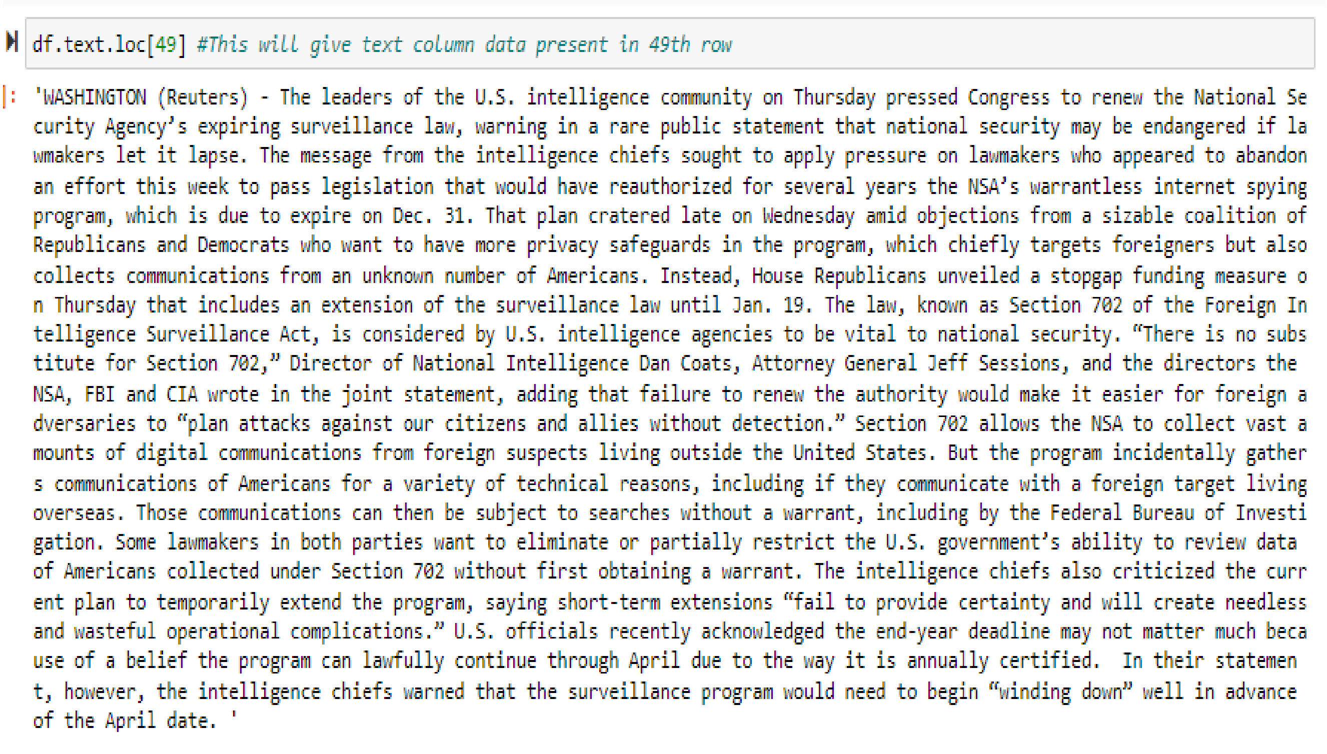
```
4]: 'Spy chiefs pressure Congress to renew expiring surveillance law'
```

```
df.subject.loc[49] #This will give subject column data present in 49th row
```

```
5]: 'politicsNews'
```

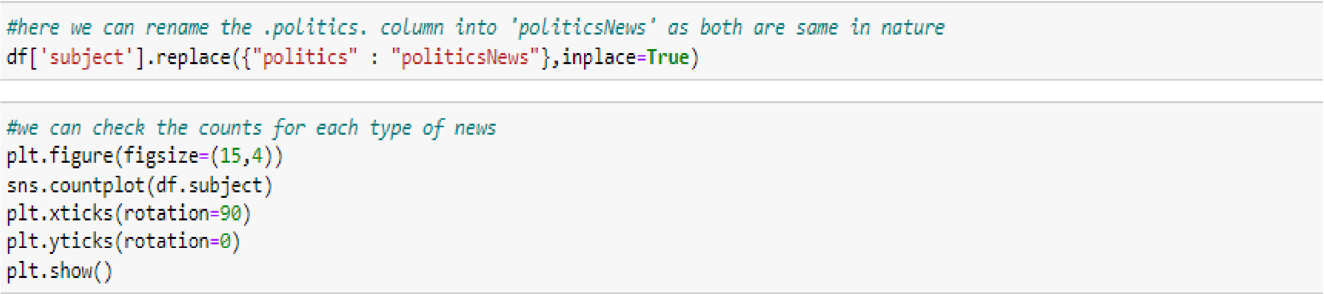
```
df.label.loc[49] #This will give label column data present in 49th row
```

```
5]: 1
```



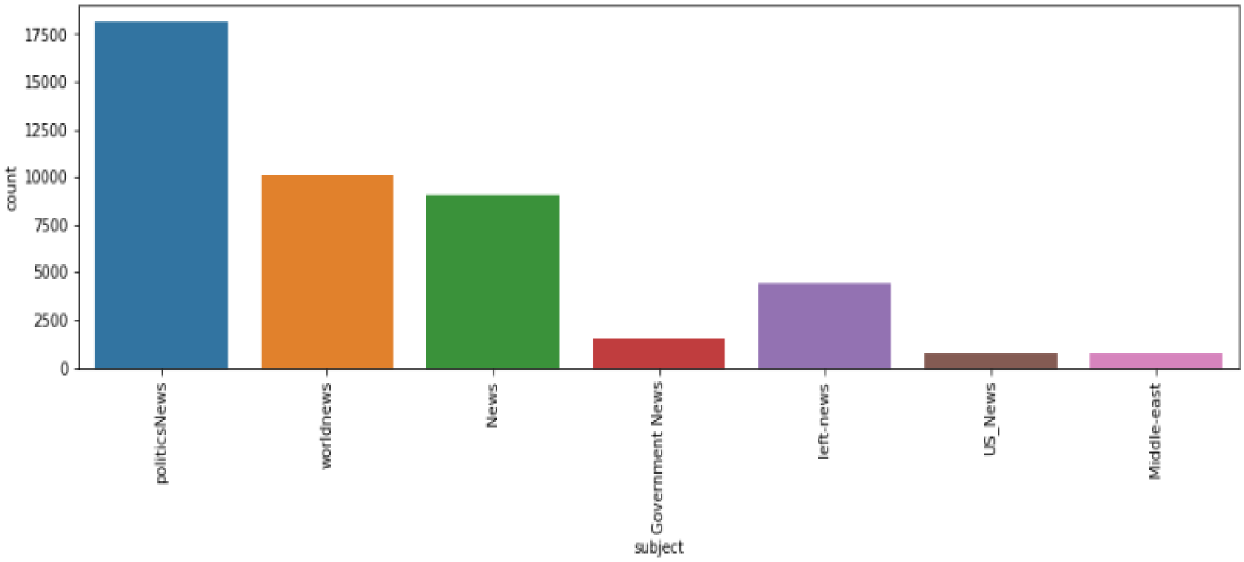
```
df.text.loc[49] #This will give text column data present in 49th row
```

```
'WASHINGTON (Reuters) - The leaders of the U.S. intelligence community on Thursday pressed Congress to renew the National Se
curity Agency’s expiring surveillance law, warning in a rare public statement that national security may be endangered if la
wmakers let it lapse. The message from the intelligence chiefs sought to apply pressure on lawmakers who appeared to abandon
an effort this week to pass legislation that would have reauthorized for several years the NSA’s warrantless internet spying
program, which is due to expire on Dec. 31. That plan cratered late on Wednesday amid objections from a sizable coalition of
Republicans and Democrats who want to have more privacy safeguards in the program, which chiefly targets foreigners but also
collects communications from an unknown number of Americans. Instead, House Republicans unveiled a stopgap funding measure o
n Thursday that includes an extension of the surveillance law until Jan. 19. The law, known as Section 702 of the Foreign In
telligence Surveillance Act, is considered by U.S. intelligence agencies to be vital to national security. “There is no subs
titute for Section 702,” Director of National Intelligence Dan Coats, Attorney General Jeff Sessions, and the directors the
NSA, FBI and CIA wrote in the joint statement, adding that failure to renew the authority would make it easier for foreign a
dversaries to “plan attacks against our citizens and allies without detection.” Section 702 allows the NSA to collect vast a
mounts of digital communications from foreign suspects living outside the United States. But the program incidentally gather
s communications of Americans for a variety of technical reasons, including if they communicate with a foreign target living
overseas. Those communications can then be subject to searches without a warrant, including by the Federal Bureau of Investi
gation. Some lawmakers in both parties want to eliminate or partially restrict the U.S. government’s ability to review data
of Americans collected under Section 702 without first obtaining a warrant. The intelligence chiefs also criticized the curr
ent plan to temporarily extend the program, saying short-term extensions “fail to provide certainty and will create needless
and wasteful operational complications.” U.S. officials recently acknowledged the end-year deadline may not matter much beca
use of a belief the program can lawfully continue through April due to the way it is annually certified.  In their statemen
t, however, the intelligence chiefs warned that the surveillance program would need to begin “winding down” well in advance
of the April date. '
```

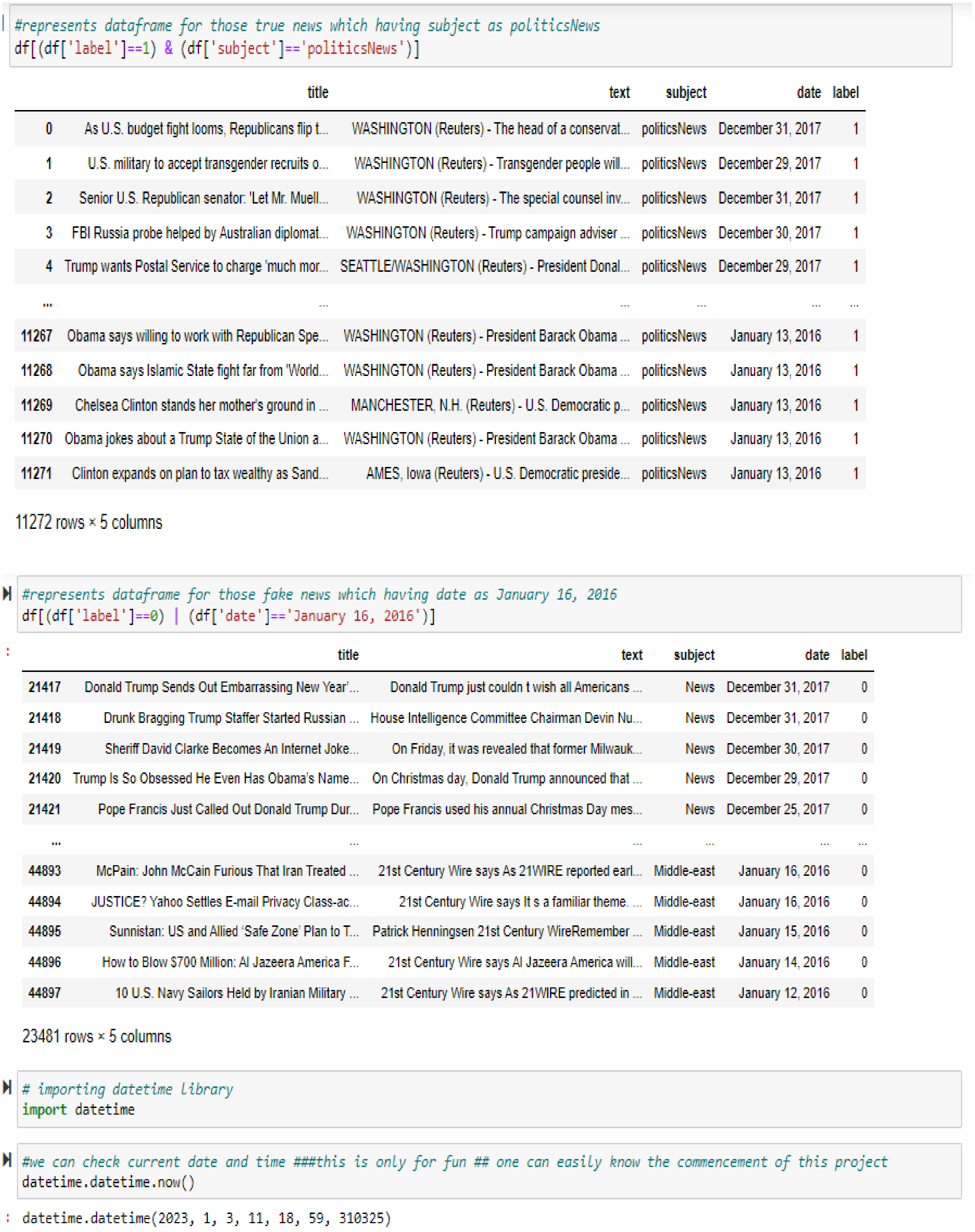


```
#here we can rename the .politics. column into 'politicsNews' as both are same in nature
df['subject'].replace({"politics" : "politicsNews"},inplace=True)
```

```
#we can check the counts for each type of news
plt.figure(figsize=(15,4))
sns.countplot(df.subject)
plt.xticks(rotation=90)
plt.yticks(rotation=0)
plt.show()
```



06-Jan-23 7:58 PM Ashutosh Mishra



```
#represents dataframe for those true news which having subject as politicsNews
df[(df['label']==1) & (df['subject']=='politicsNews')]
```

	title	text	subject	date	label
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1
...	...	...	...	...	...
11267	Obama says willing to work with Republican Spe...	WASHINGTON (Reuters) - President Barack Obama ...	politicsNews	January 13, 2016	1
11268	Obama says Islamic State fight far from 'World...	WASHINGTON (Reuters) - President Barack Obama ...	politicsNews	January 13, 2016	1
11269	Chelsea Clinton stands her mother's ground in ...	MANCHESTER, N.H. (Reuters) - U.S. Democratic p...	politicsNews	January 13, 2016	1
11270	Obama jokes about a Trump State of the Union a...	WASHINGTON (Reuters) - President Barack Obama ...	politicsNews	January 13, 2016	1
11271	Clinton expands on plan to tax wealthy as Sand...	AMES, Iowa (Reuters) - U.S. Democratic preside...	politicsNews	January 13, 2016	1

11272 rows × 5 columns

```
#represents dataframe for those fake news which having date as January 16, 2016
df[(df['label']==0) | (df['date']=='January 16, 2016')]
```

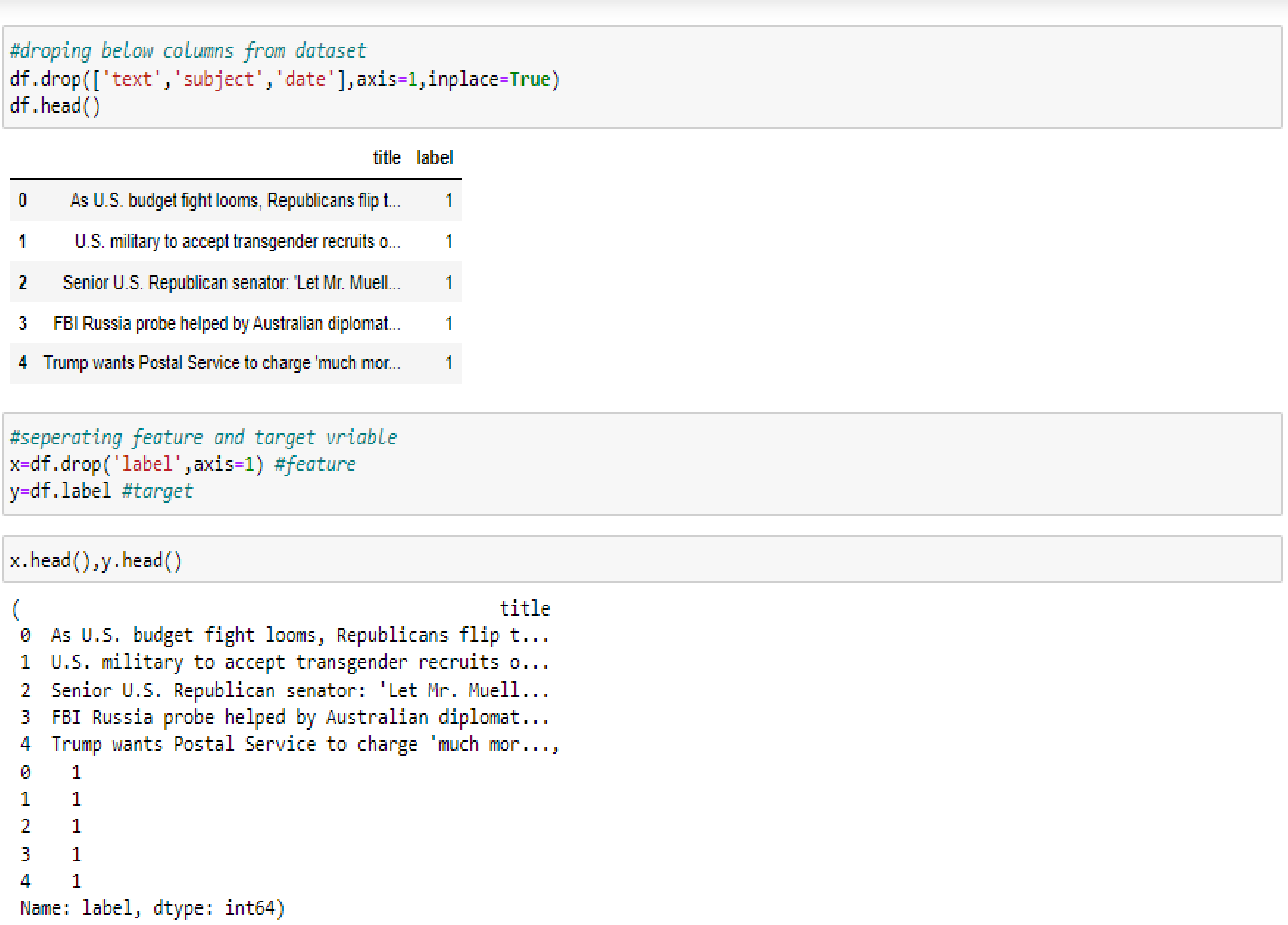
	title	text	subject	date	label
21417	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
21418	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
21419	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
21420	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
21421	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0
...	...	...	...	...	...
44893	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	0
44894	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016	0
44895	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016	0
44896	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016	0
44897	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016	0

23481 rows × 5 columns

```
# importing datetime library
import datetime
```

```
#we can check current date and time ###this is only for fun ## one can easily know the commencement of this project
datetime.datetime.now()
```

```
datetime.datetime(2023, 1, 3, 11, 18, 59, 310325)
```



```
#droping below columns from dataset
df.drop(['text','subject','date'],axis=1,inplace=True)
df.head()
```

	title	label
0	As U.S. budget fight looms, Republicans flip t...	1
1	U.S. military to accept transgender recruits o...	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	1
3	FBI Russia probe helped by Australian diplomat...	1
4	Trump wants Postal Service to charge 'much mor...	1

```
#seperating feature and target vriable
x=df.drop('label',axis=1) #feature
y=df.label #target
```

```
x.head(),y.head()
```

```
(                                               title
 0  As U.S. budget fight looms, Republicans flip t...
 1  U.S. military to accept transgender recruits o...
 2  Senior U.S. Republican senator: 'Let Mr. Muell...
 3  FBI Russia probe helped by Australian diplomat...
 4  Trump wants Postal Service to charge 'much mor...,
 0    1
 1    1
 2    1
 3    1
 4    1
 Name: label, dtype: int64)
```

Data Preprocessing Done

Eventually, I have been given four different features of object data type for making prediction for our target variable but I've consider only one feature variable and it is attribute 'title'. Here I'm not using 'text' column as our sole feature variable because the length of each documents could be very large and that too will take long time while performing text preprocessing part thereby I have dropped columns 'text', 'subject', & 'date' from our dataset. So, I've separated my independent and target variable.

In the very next step I've used **NLTK** and **RE** library for text preprocessing and cleansing. Then have replaced all the special characters of our feature **text** corpus, other than numeric data (0-9), upper and lower case alphabets, into **white space** and then made all the documents in **lower case** alphabets so that we can remove ambiguous error. At the end I've removed all unimportant words from our corpus with the help of **stopwords** and then performed **stemming** on our corpus, so that we can get root words for each similar words.

Then I've used Deep NLP techniques like tensorflow and keras so that we can easily work upon our text based corpus with the help of neural networks . Tensorflow and Keras provide high-level APIs used for easily building and training models, but Keras is more user-friendly because it's built-in Python.

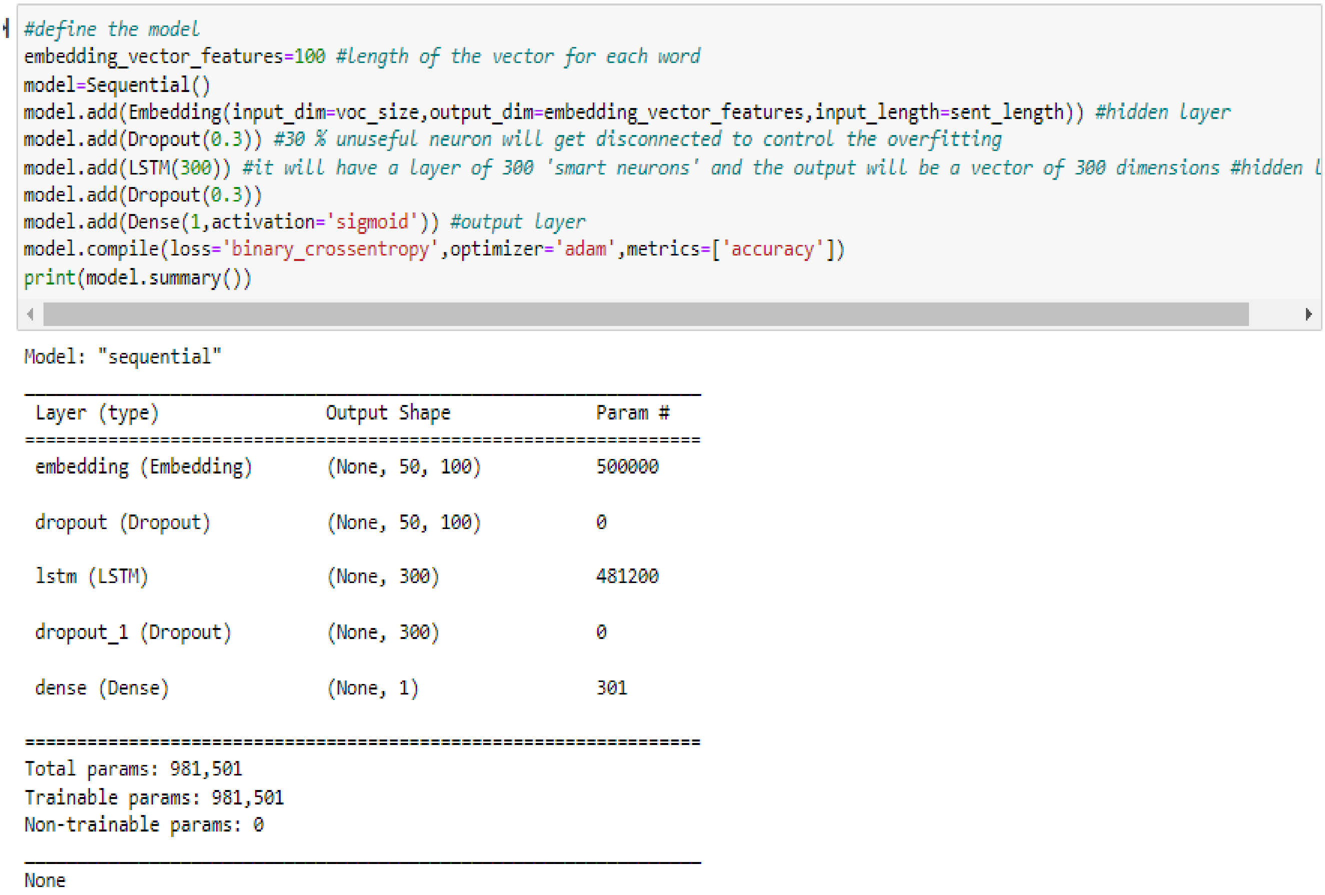
Once I get the resultant corpus I've decided the vocabulary size as 5000 because later I'll be using one_hot() function that creates an index for each word of a document as an efficient integer encoding and then will apply embedding layer to define the model.

The sequences for each documents have different lengths and keras prefers inputs to be vectorized and all inputs to have the same length. Thereby we will **pad** all the input sequences to have the length of 50 words (let's assume). I have used **pre-padding** technique to fill zero's (0) in each documents, wherever the words are missing, if the words are less than 50 in each documents.

We are now ready to define our Embedding layer as part of our neural network model.

The Embedding has a vocabulary of 5000 and an input length of 50. We will choose an embedding space i.e. our output dimension as of 100 dimensions. The model is a simple binary classification model. The output from the embedding layer will be 50 vectors of 100 dimension each, one for each word.

Let's have a look on our model summary for the same-



```
#define the model
embedding_vector_features=100 #length of the vector for each word
model=Sequential()
model.add(Embedding(input_dim=voc_size,output_dim=embedding_vector_features,input_length=sent_length)) #hidden layer
model.add(Dropout(0.3)) #30 % unuseful neuron will get disconnected to control the overfitting
model.add(LSTM(300)) #it will have a layer of 300 'smart neurons' and the output will be a vector of 300 dimensions #hidden l
model.add(Dropout(0.3))
model.add(Dense(1,activation='sigmoid')) #output layer
model.compile(loss='binary_crossentropy',optimizer='adam',metrics=['accuracy'])
print(model.summary())
```

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 50, 100)           500000

 dropout (Dropout)           (None, 50, 100)           0

 lstm (LSTM)                 (None, 300)               481200

 dropout_1 (Dropout)         (None, 300)               0

 dense (Dense)               (None, 1)                 301

=================================================================
Total params: 981,501
Trainable params: 981,501
Non-trainable params: 0
_________________________________________________________________
None
```

So we have created our neural networks having input as length of 5000 vocabularies then we have added first hidden layer i.e. embedding layer having input length 50 and output dimension as 100. Then again added a second hidden layer as LSTM (300) it means that it will have a layer of 300 'smart neurons' and the output will be a vector of 300 dimensions as well. In the output layer I've applied the activation function as 'sigmoid' since it is a binary classification problem and will be getting only one output at a time. I've applied dropout= 30% it means that it will disconnect 30% impractical neurons to reduce the overfitting problem.

06-Jan-23 7:58 PM Ashutosh Mishra

Once the model is define will separate our target and feature variable and will convert it into array, so that we can fit our model with train_test_split method. I've selected epochs=10 and batch_size=100 while training the model. Also, I have taken the validation split over the test data to compare the accuracy and loss w.r.t. that of training data.

Since, Classification metrics can't handle a mix of binary and continuous targets, so have applied the concept as - if y_pred>0.5 then convert it into 1 otherwise 0.

Data Inputs- Logic- Output Relationships

Below are some inputs which were really helpful in this model building

Embeddings

Are a great way to deal with NLP problems because of two reasons! First it helps in dimensionality reduction over one-hot encoding as we can control the number of features. Second it is capable of understanding the context of a word so that similar words have similar embeddings.

LSTM

Long Short Term Memory Networks is an advanced RNN, a sequential network that allows information to persist. It is capable of handling the vanishing gradient problem faced by RNN. A recurrent neural network is also known as RNN and is used for persistent short term memory.

Dense

In any neural network, a dense layer is a layer that is deeply connected with its preceding layer which means the neurons of the layer are connected to every neuron of its preceding layer. This layer is the most commonly used layer in artificial neural network networks. In Keras, "dense" usually refers to a single layer.

Dropout

It refers to dropping out the nodes (input and hidden layer) in a neural network to control overfitting problem.

One-hot Encoding

Converts text into vector.

Padding

To ensure that all the input sequence data is having the same length we pad or truncate the input data points.

Sequential

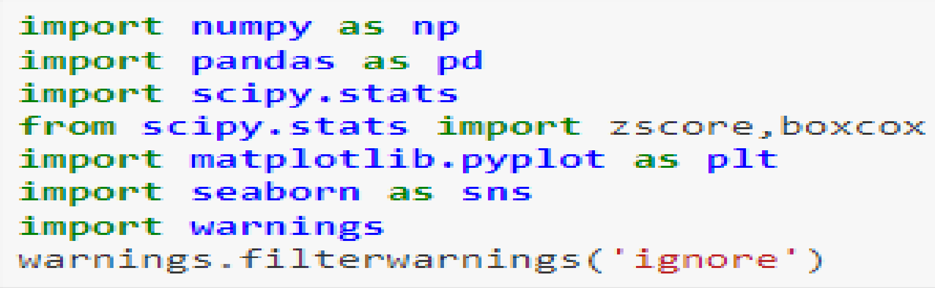
In Keras it usually refers to an entire model, not just one layer. Sequential refers to the way we build models in Keras using the sequential API. Sequential API is used to create models layer-by-layer. In sequential models, the input, hidden and output layers are stacked in the model sequentially. The information propagates from the input to the output through hidden layers.

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

No.

Hardware and Software Requirements and Tools Used

I have used **Python IDE** (Integrated Development environment) as a dedicated software throughout solving this project.



```
import numpy as np
import pandas as pd
import scipy.stats
from scipy.stats import zscore,boxcox
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

Python Libraries that I've used throughout the process are-

- Numpy- It is use for linear algebra
- Pandas- data analysis/manipulation library
- Matplotlib-Data visualization and plotting library
- Seaborn- Data visualization and statistical plotting library
- Sklearn- Machine Learning Tool
- NLTK- NLP libraries with respect to machine learning
- RE- Regular Expression
- Tensorflow- Deep NLP and Machine learning library
- Keras- Python-Inbuilt Deep NLP Library

Classes-

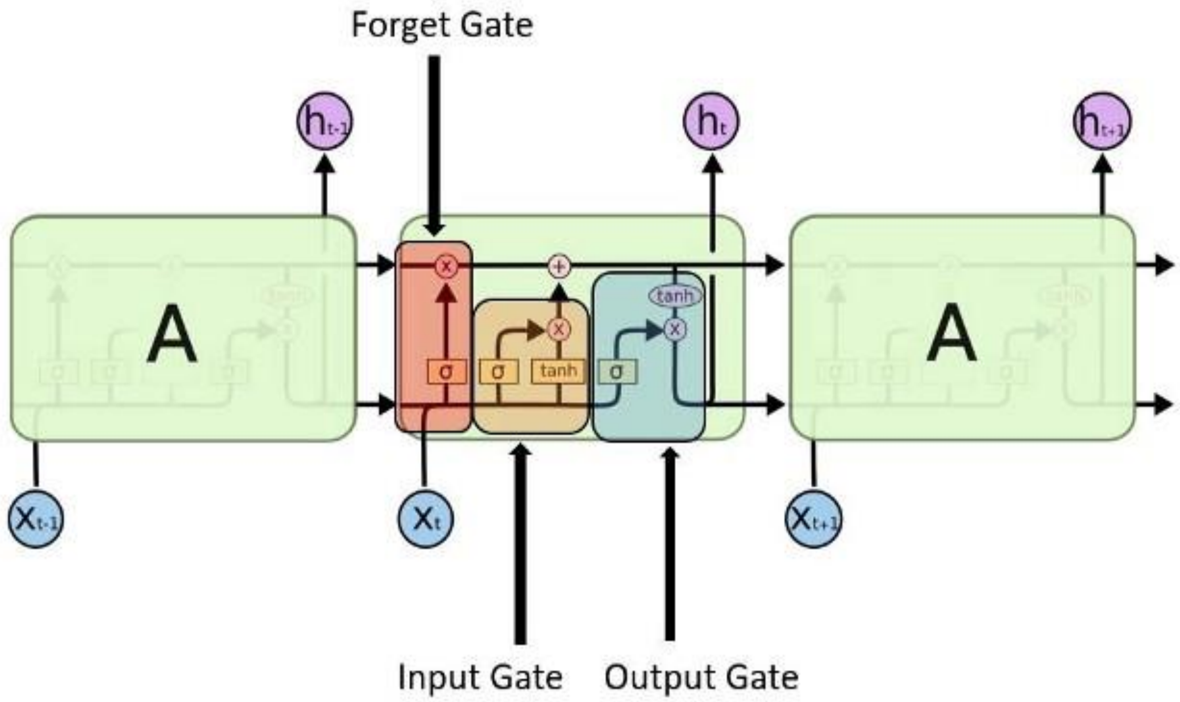
- PorterStemmer
- Stopwords
- Embedding
- LSTM
- Dense
- Dropout
- one_hot
- pad_sequences
- Sequential
- train_test_split

Model/s Development and Evaluation

Identification of possible problem-solving approaches

(LSTM-RNN)

LSTMs assign data **weights** which helps RNNs to either let new information in, forget information or give it importance enough to impact the output. LSTM networks are a modified version of recurrent neural networks, which makes it easier to remember past data in memory. The vanishing gradient problem of RNN is resolved here. LSTM is well-suited to classify, process and predict time series given time lags of unknown duration. It trains the model by using back-propagation.



06-Jan-23 7:58 PM Ashutosh Mishra

In an LSTM network, basically three gates are present-

Input gate discover which value from input should be used to modify the memory. Sigmoid function decides which values to let through **0**, **1** and tanh function gives weightage to the values which are passed deciding their level of importance ranging from -**1** to **1**.

Forget gate discover what details to be discarded from the block. It is decided by the sigmoid function. It looks at the previous state and the content input and outputs a number between **0** (omit this) and **1** (keep this) for each number in the cell state.

Output gate the input and the memory of the block is used to decide the output. Sigmoid function decides which values to let through **0**, **1**; and tanh function gives weightage to the values which are passed deciding their level of importance ranging from -**1** to **1** and multiplied with output of Sigmoid.

Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing.

- from sklearn.linear_model import train_test_split
- from nltk.corpus import stopwords
- from nltk.stem.porter import PorterStemmer
- import tensorflow as tf
- from tensorflow.keras.layers import Embedding
- from tensorflow.keras.layers import LSTM
- from tensorflow.keras.layers import Dense
- from tensorflow.keras.layers import Dropout
- from tensorflow.keras.preprocessing.text import one_hot #encoding technique converts text in to vector
- from tensorflow.keras.preprocessing.sequence import pad_sequences
- from tensorflow.keras.models import Sequential

Key Metrics for success in solving problem under consideration

Below are the some key metrics which were useful in our findings -

Confusion Matrix-

```
array([[4579,   71],
       [  68, 4262]], dtype=int64)
```

False Positive (FP) – (Type-I error)

The predicted value was falsely predicted.

The actual value was negative (0) but the model predicted a positive value (1, True News) 71 times.

False Negative (FN) – (Type-II error)

The predicted value was falsely predicted.

The actual value was positive (1) but the model predicted a negative value (0, Fake News) 68 times

Classification Report

```
precision    recall  f1-score   support

           0       0.99      0.98      0.99      4650
           1       0.98      0.98      0.98      4330

    accuracy                           0.98      8980
   macro avg       0.98      0.98      0.98      8980
weighted avg       0.98      0.98      0.98      8980
```

Precision vs. Recall vs. F-1 Score

Precision

Tells us how many of the correctly predicted cases actually turned out to be positive.

for fake news(0)- 99 of the correctly predicted cases actually turned out to be positive(Fake news).

for True news(1)- 98 of the correctly predicted cases actually turned out to be positive(True news).

Recall

Tells us how many of the actual positive cases we were able to predict correctly with our model.

for fake news(0)- 98 of the actual positive cases (Fake news) we are able to predict correctly with our model.

for True news(1)- 98 of the actual positive cases (True news) we are able to predict correctly with our model.

F-1 Score

Is basically the harmonic mean of Precision and recall

for fake news(0)- 99%

for True news(1)- 98%

CONCLUSION

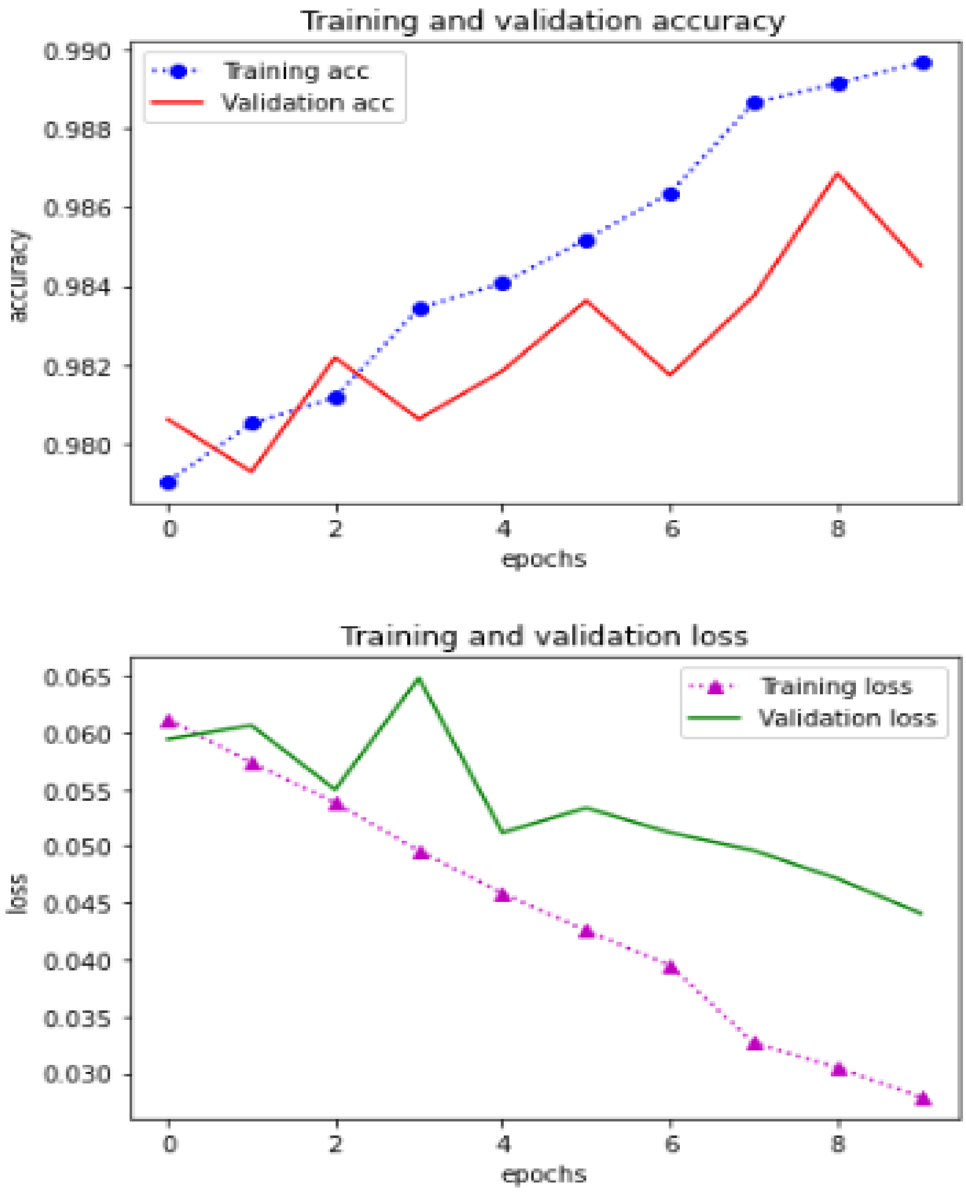
At each epoch our model training loss is getting decreasing and training accuracy is getting increasing which shows that our model is performing well with LSTM RNN neural networks and it's giving testing accuracy as 98% which is an excellent accuracy.

Also validation loss and validation accuracy is almost similar to that of Training loss and Training accuracy respectively, which indicates this model is a perfectly fit for predicting the fake news.

The training loss indicates how well the model is fitting the training data, while the validation loss indicates how well the model fits new data

Training Accuracy indicates how the model is able to classify the two output during training on the training dataset. Valid Accuracy means how the model is able to classify the outputs with the validation dataset.

Kindly see the below graph for clear understanding of how the LSTM RNN performs while predicting the fake news.


Training and validation accuracy
Training acc
Validation acc
accuracy
0.990
0.988
0.986
0.984
0.982
0.980
0
2
4
6
8
epochs
Training and validation loss
Training loss
Validation loss
loss
0.065
0.060
0.055
0.050
0.045
0.040
0.035
0.030
0
2
4
6
8
epochs

Learning Outcomes of the Study in respect of Data Science

After doing this project I got a fair idea that I can perform NLP with both Machine learning as well as Deep Learning. Also, what I've noticed that the time complexity with Deep NLP is less as compared to that of Machine Learning. This project has helped me in subdue my theoretical understanding better after practicing it practically.

Limitations of this work and Scope for Future Work

The same analysis which is done for this project can be used for any other news or spam classifier. Here I didn’t use any classification model with the help of Machine Learning so one can only use this model when they are working with Deep NLP only.