Project Report

On

Fake News Detection Using Machine Learning

IN PARTIAL FULFILMENT FOR THE REQUIREMENT OF SECOND YEAR, PROJECT,

Masters of Computer Network and Information Security

Submi&ed by

M.tech-S.Y-[CNIS]

Under the guidance of

(Assistant Professor)

Department of Computer Science & Engineering

SHRI GURU GOBIND SINGHJI INSTITUTE
OF ENGINEERING AND TECHNOLOGY
VISHNUPURI NANDED - 431606 (M.S)
INDIA ACADEMIC YEAR: 2021-2022

Abstract

In our modern world where the internet is necessary, everyone relies on various online resources for news. Along with the increase in the use of social media platforms like Facebook, Twitter, etc. news spread rapidly among millions of users within a very short span of time. The spread of fake news has far-reaching consequences like the creation of biased opinions to swaying election outcomes for the benefit of certain candidates. Moreover, spammers use appealing news headlines to generate revenue using advertisements via click-baits. In this paper, we aim to perform different classification algorithms of various news articles available online with the help of concepts pertaining to Machine Learning. We aim to provide the user with the ability to classify the news as fake or real and also check the authenticity of the website publishing the news.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 BASIC INFORMATION

Fake news's simple meaning is to include information that leads people to the wrong path. Nowadays fake news spreading like water and people share this information without verifying it. And this type of news are spread for personal reasons or political agendas and so on. So it is necessary to detect fake news.

Many scientists believe that fake news issue may be solved by means of machine learning and artificial intelligence. Basically, machine learning is a technique in which computer program learn from experiences or past experiences. Then specifically related to that it performs the task. So in this project, Our goal is to develop a best model that classifies a given news article as either fake or true. It can be detected using various machine learning techniques.

1.2 MACHINE LEARNING

Machine Learning is a subset of Artificial Intelligence. It is a set of algorithms that train on a data set to make predictions or take actions in order to optimize some systems.

It is a technique in which computer program learn from experiences or past experiences. Then specifically related to that it performs the task.

Machine Learning algorithm create model using sample data or training data. Which takes the decision or prediction to perform the task. Then model checks the accuracy of that prediction, if accuracy is acceptable and prediction is correct, then machine learning algorithm gets deploy. But if accuracy is not acceptable then model again get trained until accuracy doesn’t get accepted.

In machine learning whatever the processes are involved are similar to data mining and predictive modeling.

- Data mining is the process of finding the patterns from the large dataset in which machine learning, statistics, and database systems are used. It is also known as KDD(Knowledge discovery in databases).
- Predictive Modeling is the process of predicting future outcomes by using data mining and probability.

It uses mathematical module to find out hidden pattern from the data. This solves the real life problem or business problem. Machine Learning do complex task easily. For example face detection, self driving car, Alexa, etc.

Machine learning algorithms has three parts:-

- Input
- Output
- Objective Function or Performance matrix

The steps involved in machine learning are:-

- Problem Identification
- Data Collection and validation
- Model Building
- Feedback

1.2.1 MACHINE LEARNING MODELS

In Machine Learning, techniques and data are more important, but to identify the type of problem (supervised or unsupervised) is equally important.

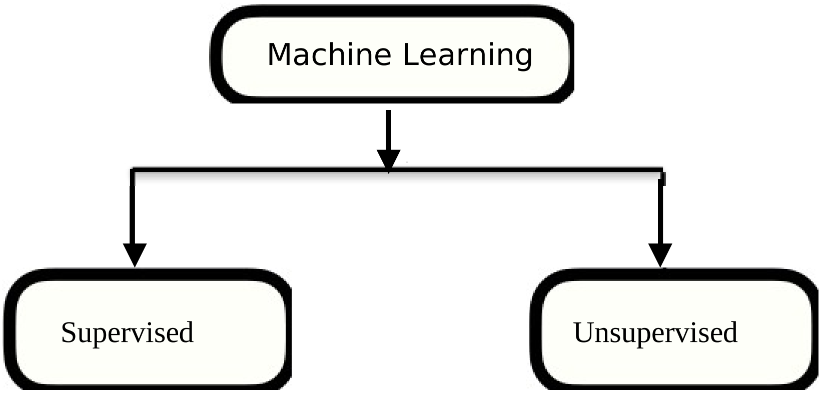


Fig 1.1 Types of ML

Supervised Lear:ning

In this input and output are already available. On the basis of this input and output or labeled data model is created and on that model new input are given and checked whether valid output is coming or not. In this model machine knows the features.

Supervised Learning are,

· Already Tagged Data

· Features and labels present

There are two types of supervised Learning:-

• Classification:-

- Classify which label a given set of features belongs to.
- This algorithm is used when there are limited number of answers Like, yes or no, 0 or 1, true or false, etc.

  For example, Is it cold? -> Yes or No

- In these algorithm we have only two outputs. So, when there are only two outputs (only two choices) then that type is called as 2 class classification algorithm. But if there are more than two choices then this type is called as multi class classification.
- To solve the classification problem the algorithms required are, - Logistic regression
  - K-Nearest Neighbors (KNN)
  - Support Vector machine (SVM)
  - Naive Bayes
  - Decision Tree Classification
  - Random Forest Classification

• Regression :-

- Find out the value of the label using previous data.

  For example, What will be the temperature tomorrow?

  So the output will be suppose, 28.
- To solve the regression problem the algorithms required are,
  - Linear Regression
  - Multiple Linear Regression
  - Polynomial Regression
  - Support Vector Regression
  - Decision Tree Regression
  - Random Forest Regression

Unsupervised Learning

All a machine knows is the data in front of it. No Features, No Labels. Machine seeing the data for the first time.

In this model,

- Data are not already tagged
- Features and labels not present

- Training not done
- Clustering:- Discover the inherit groupings in the data, such as growing customers by purchasing behaviour. There are two types of algorithm used for clustering and they are,
  - K-Means Clustering
  - Hierarchical Clustering

- Association:- Association rule learning problem, such as people that buy X also tend to buy Y.
  Algorithm used for association are, - Apriori
  - Eclat

To know how a problem cane solved using machine learning models, some of the examples are:-

1. Is this person is diabetic? (Yes or no)

Algorithm used will be Classification Algorithm.

2. Is this A or B?

Algorithm used will be Classification Algorithm

3. Is this weird?

Algorithm used will be Anomaly Detection Algorithm.

4. How many?

Algorithm used will be Regression Algorithms.

5. How is this organised?

Algorithm used will be Clustering Algorithm.

CHAPTER 3

PROPOSED METHODOLOGY

3.1 PROBLEM STATEMENT

- The goal of the project is to :
  - Prepare the data-set using several methods to train the model.
  - Build a model which can give high accuracy of predicting the fake news.
- Requirements:
  - SOFTWARE : Anaconda navigator's Jupiter notebook
  - WEBSITE : Here Dataset is taken from Kaggle site.
  - LIBRARIES : Pandas, Numpy, Matplotlib, Seaborn, Scikit Learn, etc.
- So, by observing this problem statement, we can say that it comes under Classification technique which is a type of Supervised Machine Learning, here the model is built using different classification algorithms.

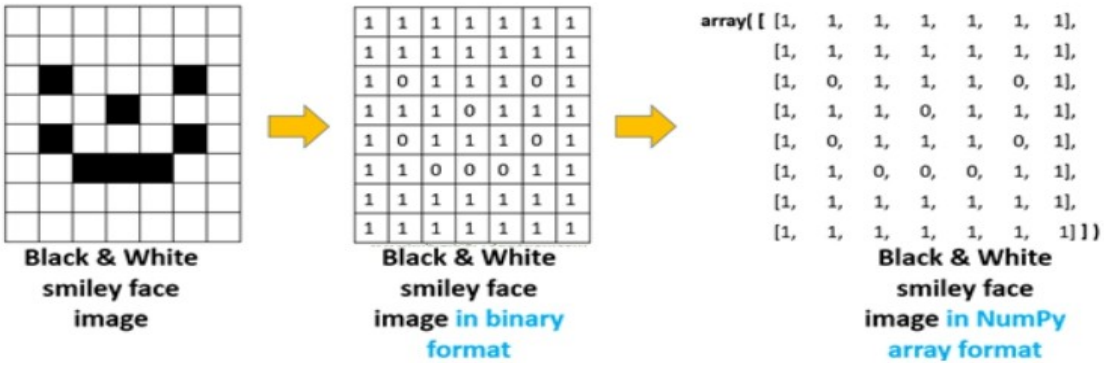
3.2 PROJECT DETAILS

It is a Machine Learning project in which we need to predict whether a given data is fake or real. Since it comes under the Classification technique which is a type of Supervised Machine Learning, here the model is built using different classification algorithms like Logistic regression, K-Nearest Neighbors, Naive Bayes , and Support Vector Machine. The model, which will give higher accuracy out of all in predicting whether the given data is fake or real will be chosen. That model or algorithm will be use further for predicting the false or real news.

Libraries used in this project are:-

- **NumPy :**

- Here, Num means numeric and Py means python.
- It is a scientific computing library for python.
- It supports multi-dimensional array. It is used to represent the large number of data in the form of array.
- NumPy Library is used for numeric calculation.
- To store data it uses less memory. It is very convenient and process fast.
- NumPy for Machine Learning:- Any machine learning model can't directly perform operation on images so for that images are converted into the numpy array format then machine learning can perform operations.



1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	0	1	1	1	0	1
1	1	1	0	1	1	1
1	0	1	1	1	0	1
1	1	0	0	0	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

```
array( [ [1,   1,   1,   1,   1,   1,   1],
         [1,   1,   1,   1,   1,   1,   1],
         [1,   0,   1,   1,   1,   0,   1],
         [1,   1,   1,   0,   1,   1,   1],
         [1,   0,   1,   1,   1,   0,   1],
         [1,   1,   0,   0,   0,   1,   1],
         [1,   1,   1,   1,   1,   1,   1],
         [1,   1,   1,   1,   1,   1,    1] ] )
```

Fig 3.2.1 NumPy for Machine Learning

Here in fig., image is converted to binary format(numeric format) then in NumPy array format. Now machine learning model can perform operation on this format.

- **Pandas:**

- Pandas is a powerful python data analysis toolkit.
- Pandas is used for data manipulation, analysis and cleaning.
- Open Source
- For data manipulation it is a fast and efficient data frame.
- It can read different types of formats like, csv, JSON, etc.

- Importance of Pandas :- Suppose in data science projects, to find useful insights or patterns from raw data firstly raw data should be in proper format. This is known as data preprocessing. In data preprocessing panda libraries are used. To prepare and process the data this, library is very fast and efficient. This library have many features like, it can use for handling missing values, reshaping the datasets, etc.

•SKLearn:

Using SKLearn for machine learning:-

- Simple and efficient tools for data mining and data analysis.
- Accessible to everybody and reusable in various contexts.
- Built on Numpy, Scipy, and Matplotlib.
- Open source, commercially usable-BSD license.

•Matplotlib:

- Matplotlib is a 2D and 3D plotting python library.
- This library is used when there is a bulk of data at that time we can transform that data in a graphical representation and can analyse the data easily.
- It is use to create high quality graph.
- Matplotlib graphs are Histogram, Bar charts, Scatterplots, etc.

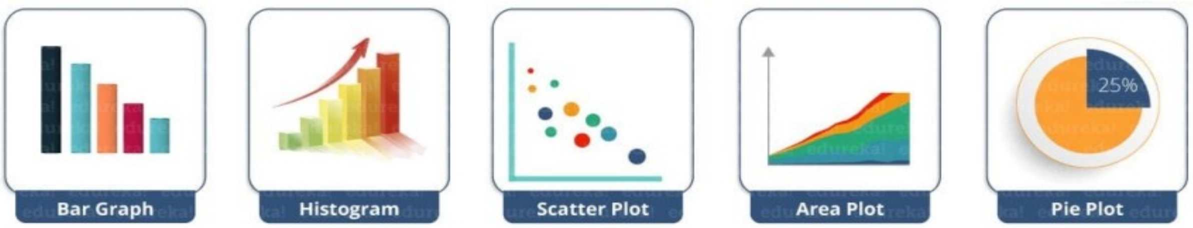


Fig 3.2.2 Matplotlib Plots

• Seaborn :-

- It is a python library use for data visualization. This library is build using matplotlib library.

- It is use for statistical library.
- Seaborn graphics are Heatmap, Pair plot, Facet grid, etc.
- Dependency of Seaborn:- Seaborn library is dependent upon Python, NumPy, Pandas, Scipy, Matplotlib.

3.3 PROCESS FLOW

This is the system architecture,

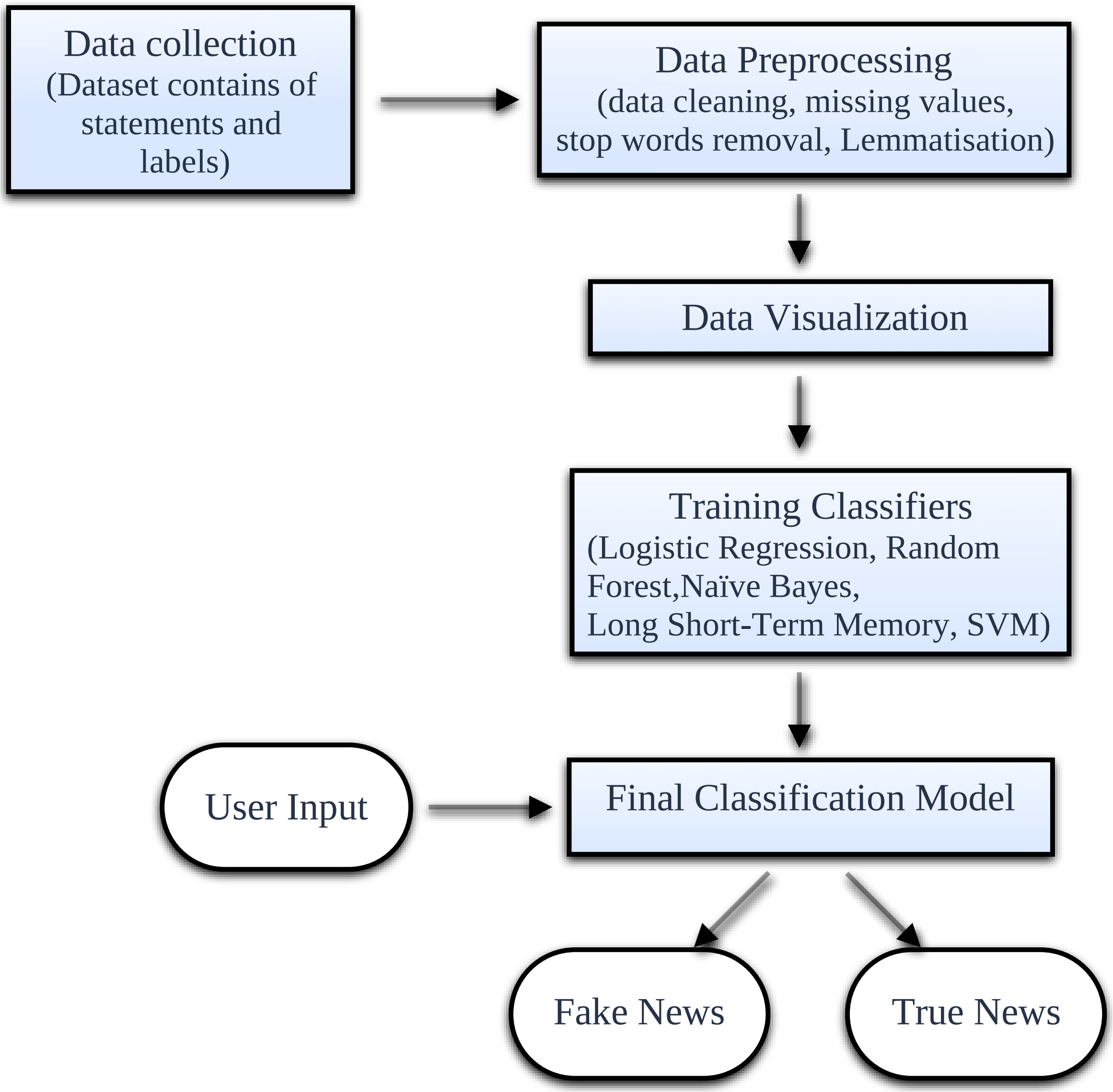
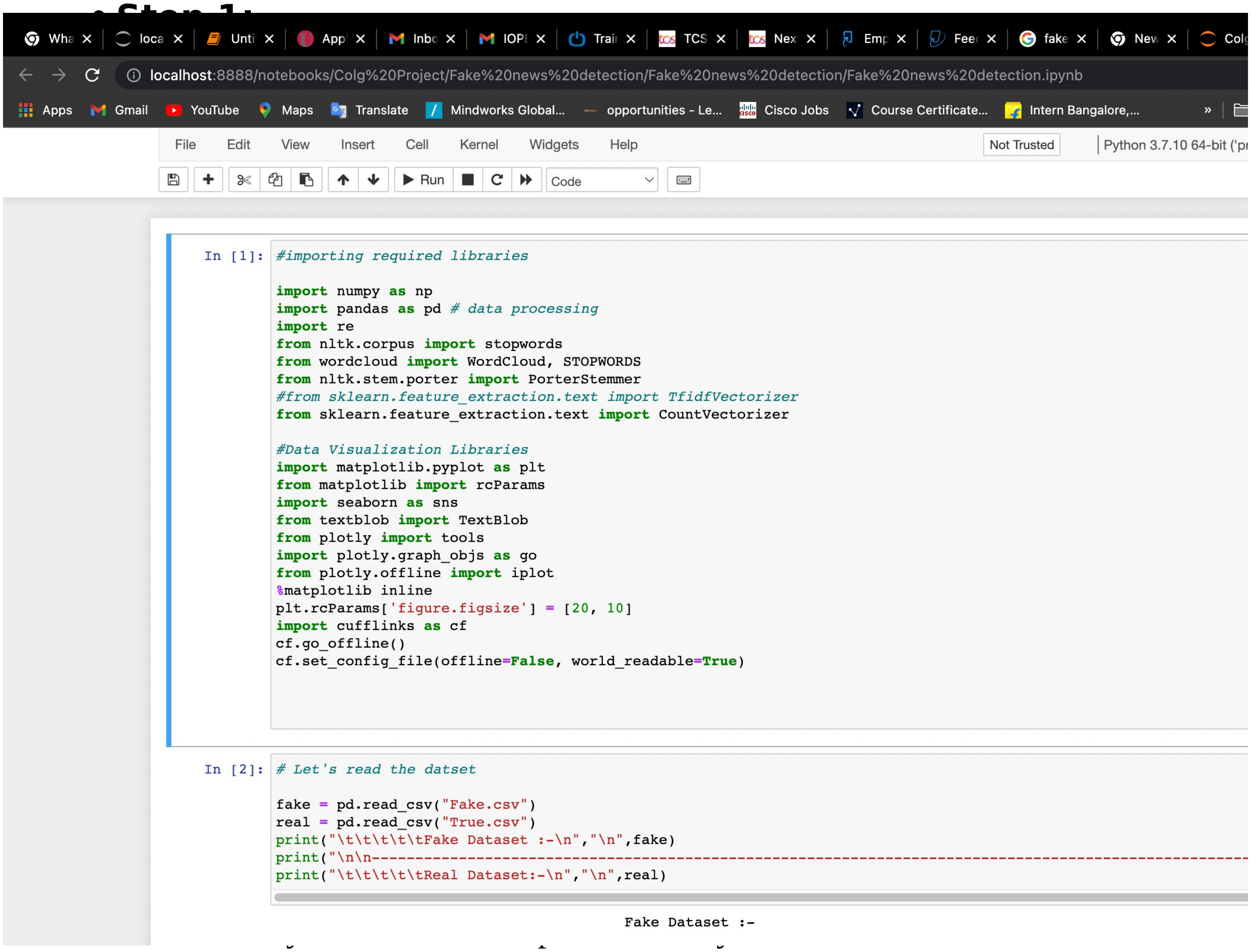

Data collection
(Dataset contains of statements and labels)
Data Preprocessing
(data cleaning, missing values, stop words removal, Lemmatisation)
Data Visualization
Training Classifiers
(Logistic Regression, Random Forest,Naïve Bayes,
Long Short-Term Memory, SVM)
User Input
Final Classification Model
Fake News
True News

Fig 3.3.1 System Architecture

Implementation STEPS:

Step by Step implementation of the project:-

- **Step 1:-**



localhost:8888/notebooks/Colg%20Project/Fake%20news%20detection/Fake%20news%20detection/Fake%20news%20detection.ipynb
Apps Gmail YouTube Maps Translate Mindworks Global... opportunities - Le... Cisco Jobs Course Certificate... Intern Bangalore,...
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3.7.10 64-bit ('pr
Run
Code

```
In [1]: #importing required libraries

        import numpy as np
        import pandas as pd # data processing
        import re
        from nltk.corpus import stopwords
        from wordcloud import WordCloud, STOPWORDS
        from nltk.stem.porter import PorterStemmer
        #from sklearn.feature_extraction.text import TfidfVectorizer
        from sklearn.feature_extraction.text import CountVectorizer

        #Data Visualization Libraries
        import matplotlib.pyplot as plt
        from matplotlib import rcParams
        import seaborn as sns
        from textblob import TextBlob
        from plotly import tools
        import plotly.graph_objs as go
        from plotly.offline import iplot
        %matplotlib inline
        plt.rcParams['figure.figsize'] = [20, 10]
        import cufflinks as cf
        cf.go_offline()
        cf.set_config_file(offline=False, world_readable=True)
```

```
In [2]: # Let's read the dataset

        fake = pd.read_csv("Fake.csv")
        real = pd.read_csv("True.csv")
        print("\t\t\t\t\tFake Dataset :-\n","\n",fake)
        print("\n\n---------------------------------------------------------------------------------------------------
        print("\t\t\t\t\tReal Dataset:-\n","\n",real)
```

```
                          Fake Dataset :-
```

- Numpy can be used to perform a wide variety of mathematical operations on arrays.
- we will use functionalities of the ‘nltk‘ library named Removing Stopwords, Tokenization, and Lemmatization.

  This is the syntax one should import libraries like this. But this libraries should be installed first then import the libraries.

Step 2:-

Second step is to read the dataset and stored it into new variable.

Here, Dataset is taken from kaggle site (https://www.kaggle.com/

clmentbisaillon/fake-and-real-news-dataset).

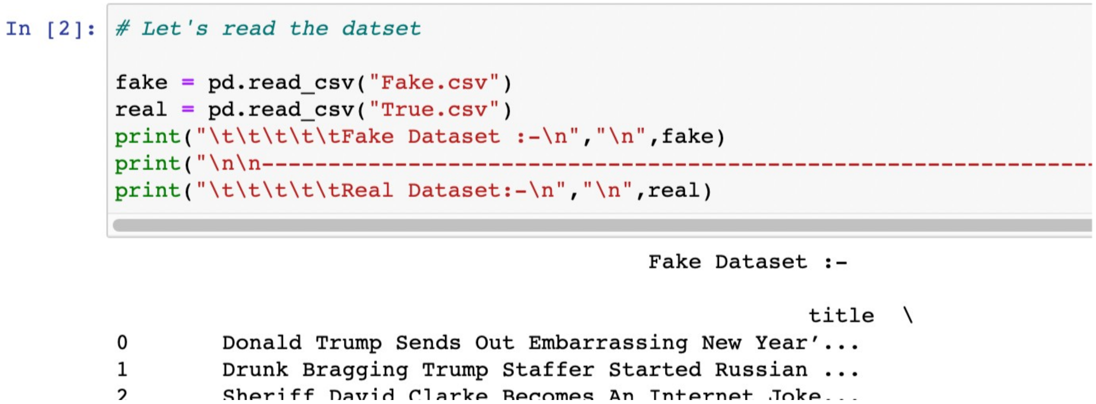
In this, data there are two datasets,

1. Fake.csv : This dataset contains a list of fake news.
2. True.csv : This dataset contains a list of Real news.

To read any dataset we need pandas library. In this project, datasets are in the csv format. Here csv means comma separated value and to read the data of csv format we use,

read_csv()

For example, In this Project

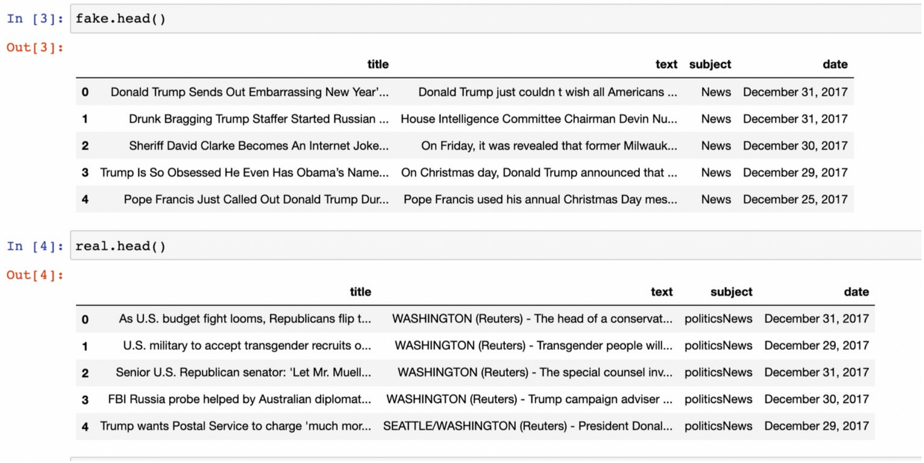


```
In [2]: # Let's read the datset

        fake = pd.read_csv("Fake.csv")
        real = pd.read_csv("True.csv")
        print("\t\t\t\t\tFake Dataset :-\n","\n",fake)
        print("\n\n--------------------------------------------------------------------
        print("\t\t\t\t\tReal Dataset:-\n","\n",real)
```

```
                                         Fake Dataset :-

                                                     title  \
0        Donald Trump Sends Out Embarrassing New Year’...
1        Drunk Bragging Trump Staffer Started Russian ...
2        Sheriff David Clarke Becomes An Internet Joke...
```

• By using head() function we can see the first n rows of the datasets.

In this Project there are two datasets so, Below screenshot shows the first 5 rows of the fake.csv and true.csv datasets.



```
In [3]: fake.head()
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

```
In [4]: real.head()
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

• Step 3:-

Next step is “Data Preprocessing step”

In this process, raw data or original data are purified. For data preprocessing the python libraries like NumPy, Panda, Sklearn, etc are used.

The steps in this process are,

1. Data Cleaning:

In this step, missing data are filled using nan or 0. Any row who’s heading or label is not given are ignored or removed. Any type of irrelevant data are removed. After this, noisy data (meaningless data or corrupted data) are deleted or removed.

2. DataTransformation:

In this step, feature scaling is done. In this scikit-learn library is used. Normally for solving classification problem data transformation is used. 3. Dimensionality Reduction:

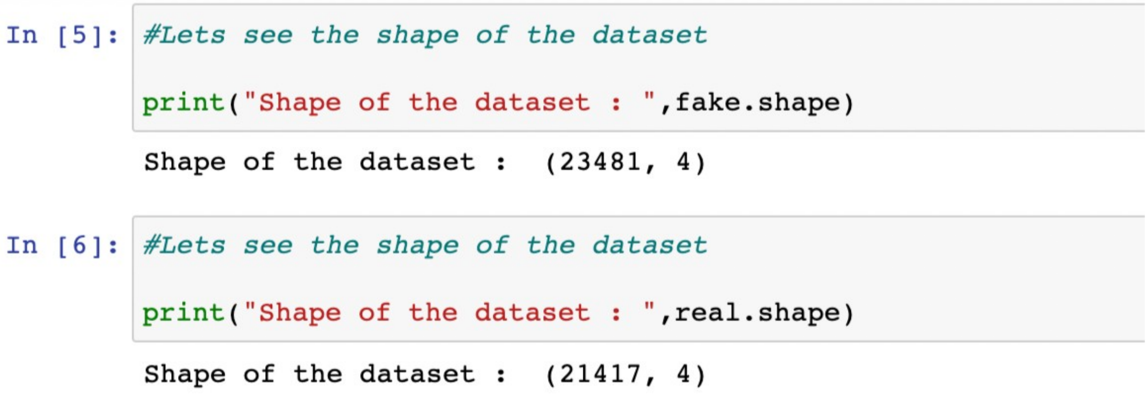
In this step, dimensions are reduced like 3D data is converted to 2D data. So in this way data are analysed.

• In this step,

we will start from checking the shape of the datasets.

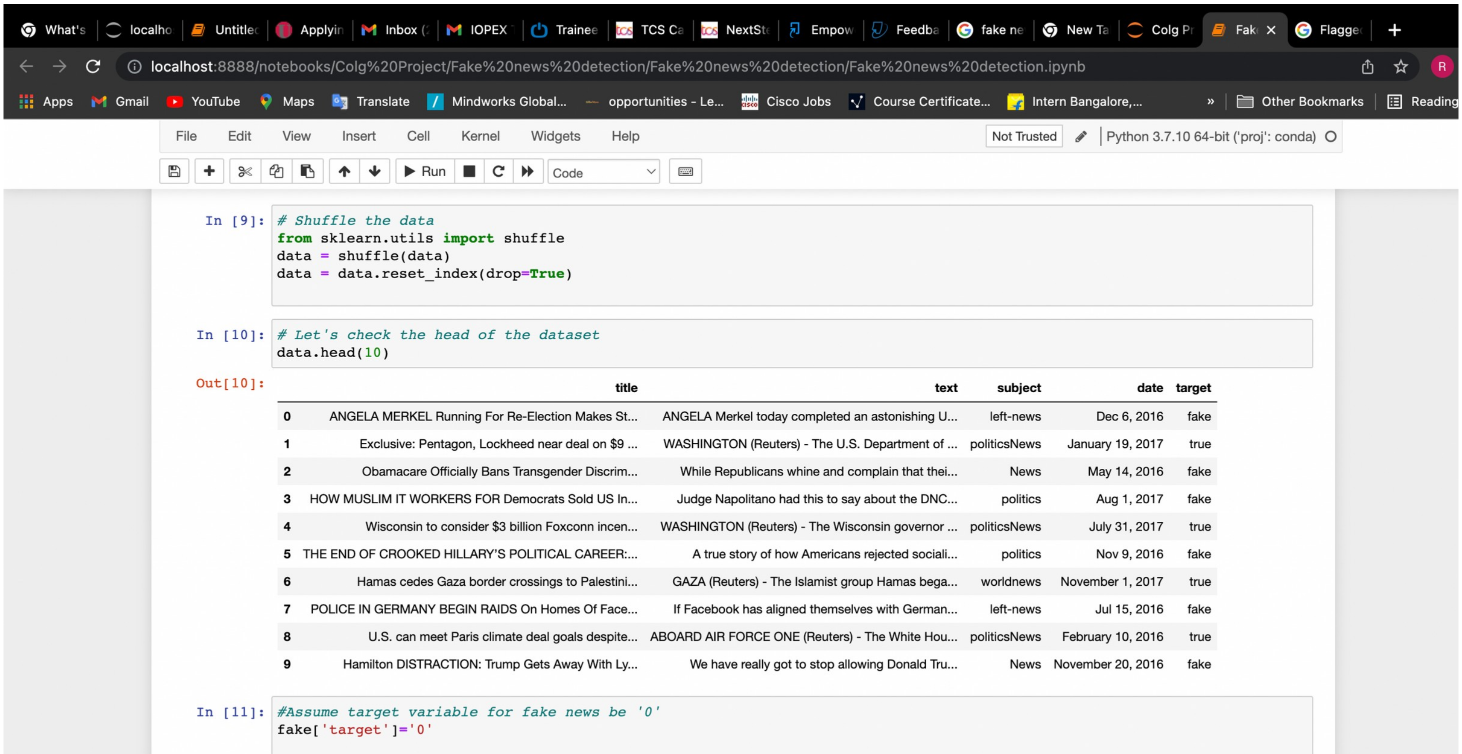
This means, In “fake.csv”dataset there are 23481 rows and 4 columns.

In “True.csv”dataset there are 21417 rows and 4 columns.



```
In [5]: #Lets see the shape of the dataset

        print("Shape of the dataset : ",fake.shape)

        Shape of the dataset :  (23481, 4)
```

```
In [6]: #Lets see the shape of the dataset

        print("Shape of the dataset : ",real.shape)

        Shape of the dataset :  (21417, 4)
```


localhost:8888/notebooks/Colg%20Project/Fake%20news%20detection/Fake%20news%20detection/Fake%20news%20detection.ipynb
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3.7.10 64-bit ('proj': conda)
Run
Code

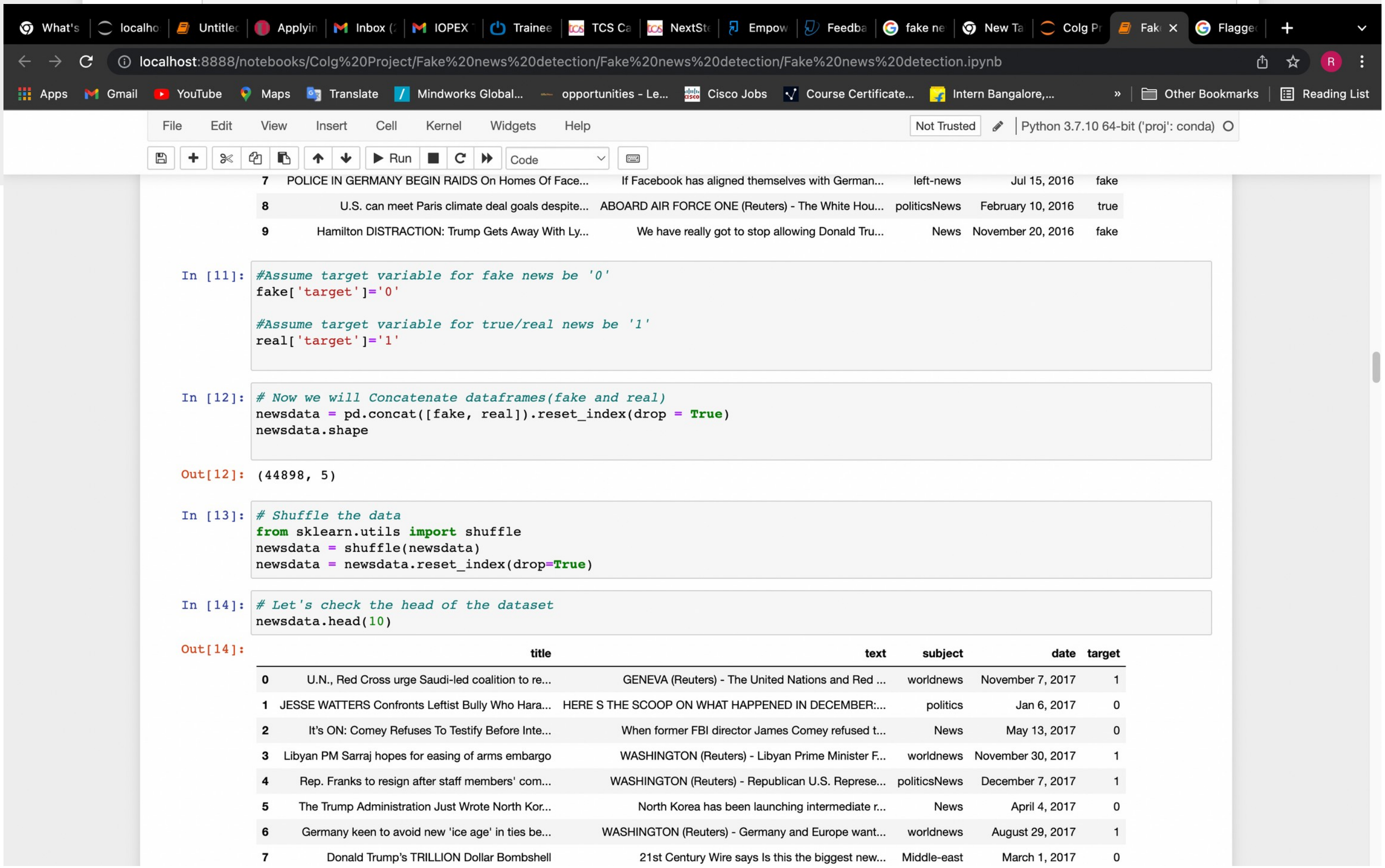
```
In [9]:  # Shuffle the data
         from sklearn.utils import shuffle
         data = shuffle(data)
         data = data.reset_index(drop=True)
```

```
In [10]: # Let's check the head of the dataset
         data.head(10)
```

Out[10]:

	title	text	subject	date	target
0	ANGELA MERKEL Running For Re-Election Makes St...	ANGELA Merkel today completed an astonishing U...	left-news	Dec 6, 2016	fake
1	Exclusive: Pentagon, Lockheed near deal on $9 ...	WASHINGTON (Reuters) - The U.S. Department of ...	politicsNews	January 19, 2017	true
2	Obamacare Officially Bans Transgender Discrim...	While Republicans whine and complain that thei...	News	May 14, 2016	fake
3	HOW MUSLIM IT WORKERS FOR Democrats Sold US In...	Judge Napolitano had this to say about the DNC...	politics	Aug 1, 2017	fake
4	Wisconsin to consider $3 billion Foxconn incen...	WASHINGTON (Reuters) - The Wisconsin governor ...	politicsNews	July 31, 2017	true
5	THE END OF CROOKED HILLARY'S POLITICAL CAREER:...	A true story of how Americans rejected sociali...	politics	Nov 9, 2016	fake
6	Hamas cedes Gaza border crossings to Palestini...	GAZA (Reuters) - The Islamist group Hamas bega...	worldnews	November 1, 2017	true
7	POLICE IN GERMANY BEGIN RAIDS On Homes Of Face...	If Facebook has aligned themselves with German...	left-news	Jul 15, 2016	fake
8	U.S. can meet Paris climate deal goals despite...	ABOARD AIR FORCE ONE (Reuters) - The White Hou...	politicsNews	February 10, 2016	true
9	Hamilton DISTRACTION: Trump Gets Away With Ly...	We have really got to stop allowing Donald Tru...	News	November 20, 2016	fake

```
In [11]: #Assume target variable for fake news be '0'
         fake['target']='0'
```


localhost:8888/notebooks/Colg%20Project/Fake%20news%20detection/Fake%20news%20detection/Fake%20news%20detection.ipynb
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3.7.10 64-bit ('proj': conda)
Run
Code

7	POLICE IN GERMANY BEGIN RAIDS On Homes Of Face...	If Facebook has aligned themselves with German...	left-news	Jul 15, 2016	fake
8	U.S. can meet Paris climate deal goals despite...	ABOARD AIR FORCE ONE (Reuters) - The White Hou...	politicsNews	February 10, 2016	true
9	Hamilton DISTRACTION: Trump Gets Away With Ly...	We have really got to stop allowing Donald Tru...	News	November 20, 2016	fake

```
In [11]: #Assume target variable for fake news be '0'
         fake['target']='0'

         #Assume target variable for true/real news be '1'
         real['target']='1'
```

```
In [12]: # Now we will Concatenate dataframes(fake and real)
         newsdata = pd.concat([fake, real]).reset_index(drop = True)
         newsdata.shape
```

```
Out[12]: (44898, 5)
```

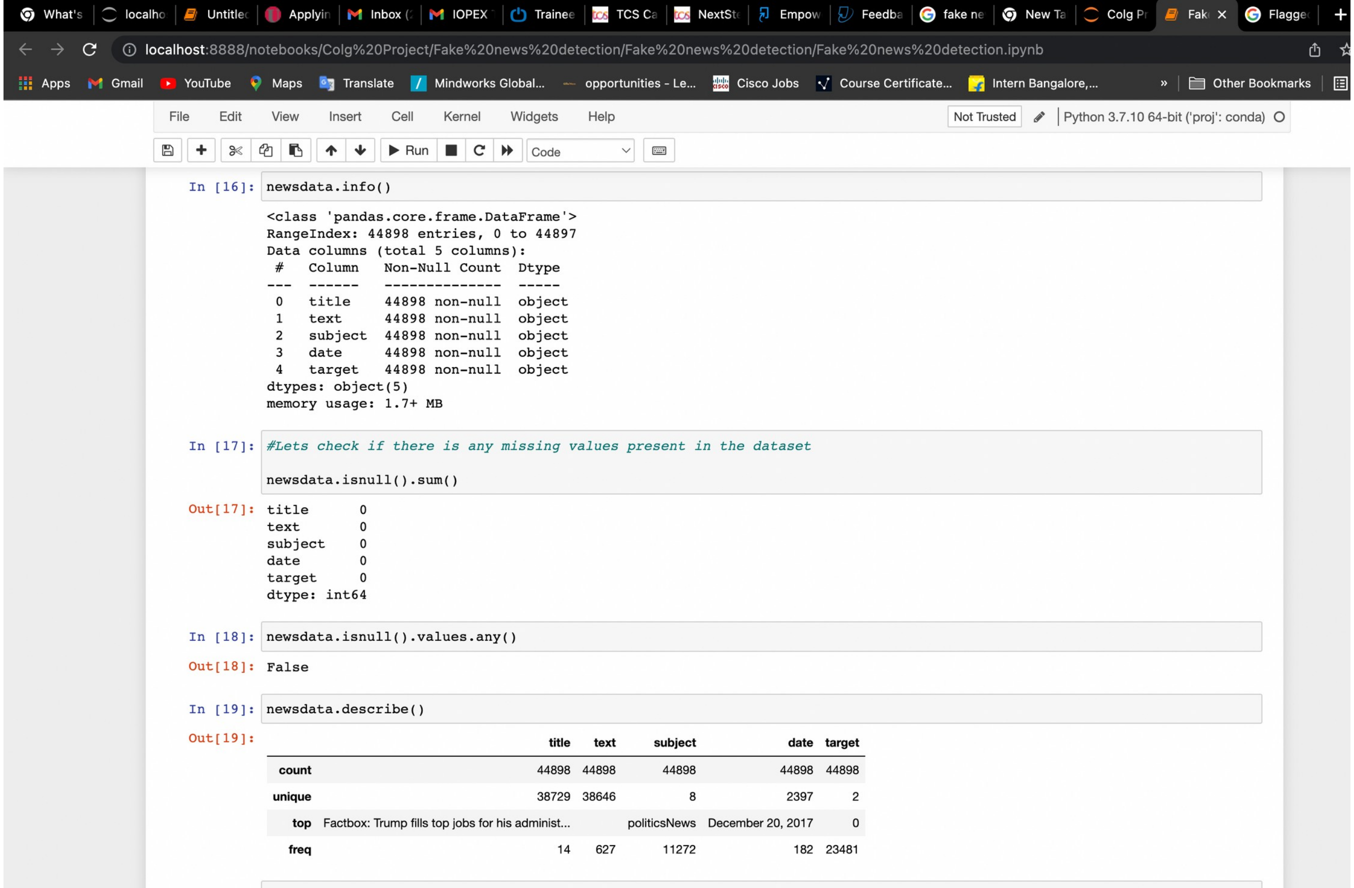
```
In [13]: # Shuffle the data
         from sklearn.utils import shuffle
         newsdata = shuffle(newsdata)
         newsdata = newsdata.reset_index(drop=True)
```

```
In [14]: # Let's check the head of the dataset
         newsdata.head(10)
```

Out[14]:

	title	text	subject	date	target
0	U.N., Red Cross urge Saudi-led coalition to re...	GENEVA (Reuters) - The United Nations and Red ...	worldnews	November 7, 2017	1
1	JESSE WATTERS Confronts Leftist Bully Who Hara...	HERE S THE SCOOP ON WHAT HAPPENED IN DECEMBER:...	politics	Jan 6, 2017	0
2	It's ON: Comey Refuses To Testify Before Inte...	When former FBI director James Comey refused t...	News	May 13, 2017	0
3	Libyan PM Sarraj hopes for easing of arms embargo	WASHINGTON (Reuters) - Libyan Prime Minister F...	worldnews	November 30, 2017	1
4	Rep. Franks to resign after staff members' com...	WASHINGTON (Reuters) - Republican U.S. Represe...	politicsNews	December 7, 2017	1
5	The Trump Administration Just Wrote North Kor...	North Korea has been launching intermediate r...	News	April 4, 2017	0
6	Germany keen to avoid new 'ice age' in ties be...	WASHINGTON (Reuters) - Germany and Europe want...	worldnews	August 29, 2017	1
7	Donald Trump's TRILLION Dollar Bombshell	21st Century Wire says Is this the biggest new...	Middle-east	March 1, 2017	0

- Now, we will check is there any null values or missing values,


localhost:8888/notebooks/Colg%20Project/Fake%20news%20detection/Fake%20news%20detection/Fake%20news%20detection.ipynb
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3.7.10 64-bit ('proj': conda)
Run
Code

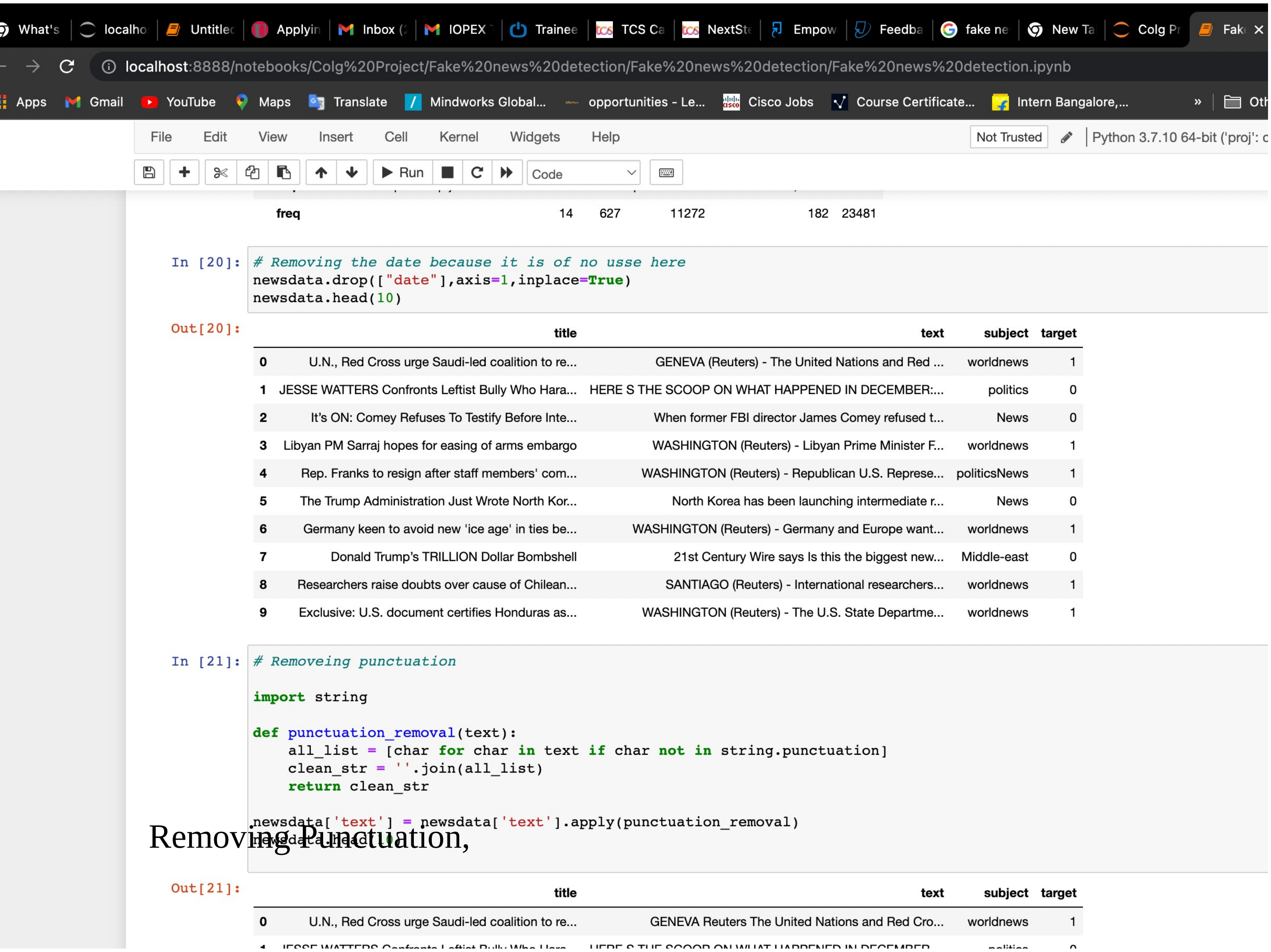
```
In [16]: newsdata.info()

         <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   target   44898 non-null  object
         dtypes: object(5)
         memory usage: 1.7+ MB
```

```
In [17]: #Lets check if there is any missing values present in the dataset

         newsdata.isnull().sum()

Out[17]: title      0
         text       0
         subject    0
         date       0
         target     0
         dtype: int64
```

```
In [18]: newsdata.isnull().values.any()

Out[18]: False
```

```
In [19]: newsdata.describe()

Out[19]:
```

	title	text	subject	date	target
count	44898	44898	44898	44898	44898
unique	38729	38646	8	2397	2
top	Factbox: Trump fills top jobs for his administ...		politicsNews	December 20, 2017	0
freq	14	627	11272	182	23481


localhost:8888/notebooks/Colg%20Project/Fake%20news%20detection/Fake%20news%20detection/Fake%20news%20detection.ipynb
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3.7.10 64-bit ('proj': c
Run
Code

freq	14	627	11272	182	23481

```
In [20]: # Removing the date because it is of no usse here
         newsdata.drop(["date"],axis=1,inplace=True)
         newsdata.head(10)

Out[20]:
```

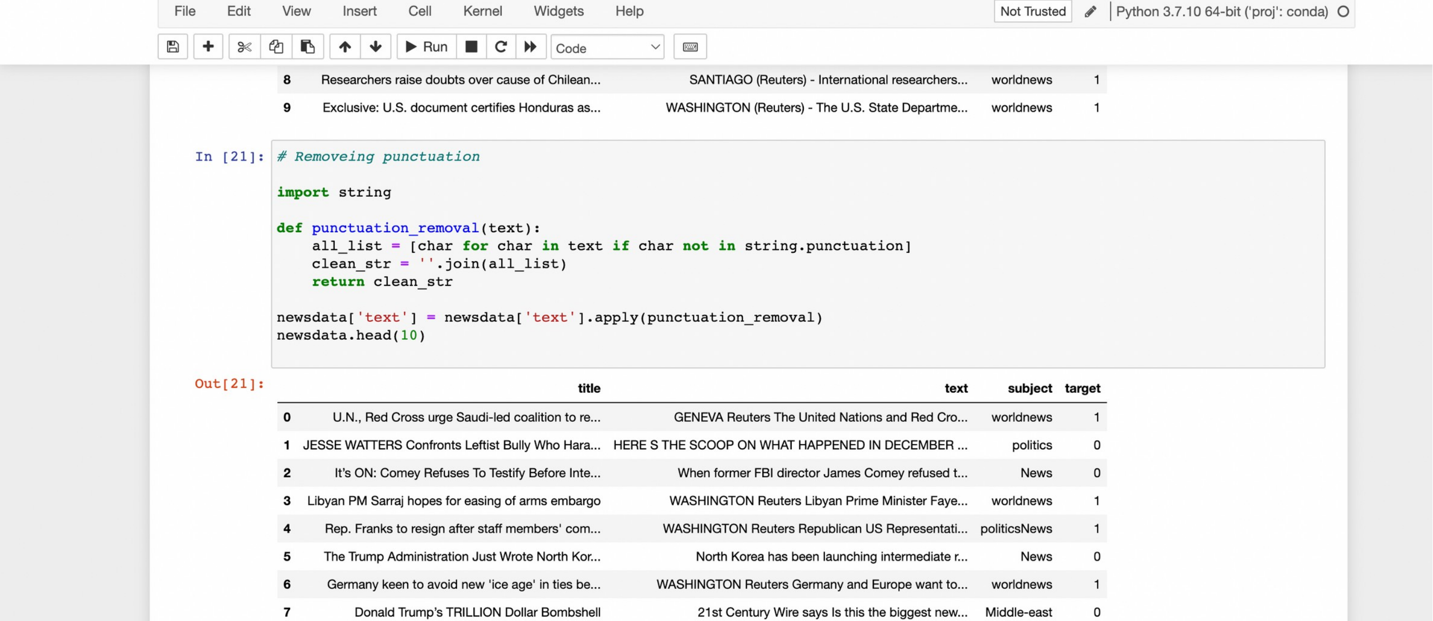
	title	text	subject	target
0	U.N., Red Cross urge Saudi-led coalition to re...	GENEVA (Reuters) - The United Nations and Red ...	worldnews	1
1	JESSE WATTERS Confronts Leftist Bully Who Hara...	HERE S THE SCOOP ON WHAT HAPPENED IN DECEMBER:...	politics	0
2	It's ON: Comey Refuses To Testify Before Inte...	When former FBI director James Comey refused t...	News	0
3	Libyan PM Sarraj hopes for easing of arms embargo	WASHINGTON (Reuters) - Libyan Prime Minister F...	worldnews	1
4	Rep. Franks to resign after staff members' com...	WASHINGTON (Reuters) - Republican U.S. Represe...	politicsNews	1
5	The Trump Administration Just Wrote North Kor...	North Korea has been launching intermediate r...	News	0
6	Germany keen to avoid new 'ice age' in ties be...	WASHINGTON (Reuters) - Germany and Europe want...	worldnews	1
7	Donald Trump's TRILLION Dollar Bombshell	21st Century Wire says Is this the biggest new...	Middle-east	0
8	Researchers raise doubts over cause of Chilean...	SANTIAGO (Reuters) - International researchers...	worldnews	1
9	Exclusive: U.S. document certifies Honduras as...	WASHINGTON (Reuters) - The U.S. State Departme...	worldnews	1

```
In [21]: # Removeing punctuation

         import string

         def punctuation_removal(text):
             all_list = [char for char in text if char not in string.punctuation]
             clean_str = ''.join(all_list)
             return clean_str

         newsdata['text'] = newsdata['text'].apply(punctuation_removal)
         newsdata.head(10)
```

Removing Punctuation,

```
Out[21]:
```

	title	text	subject	target
0	U.N., Red Cross urge Saudi-led coalition to re...	GENEVA Reuters The United Nations and Red Cro...	worldnews	1
1	JESSE WATTERS Confronts Leftist Bully Who Hara...	HERE S THE SCOOP ON WHAT HAPPENED IN DECEMBER	politics	0



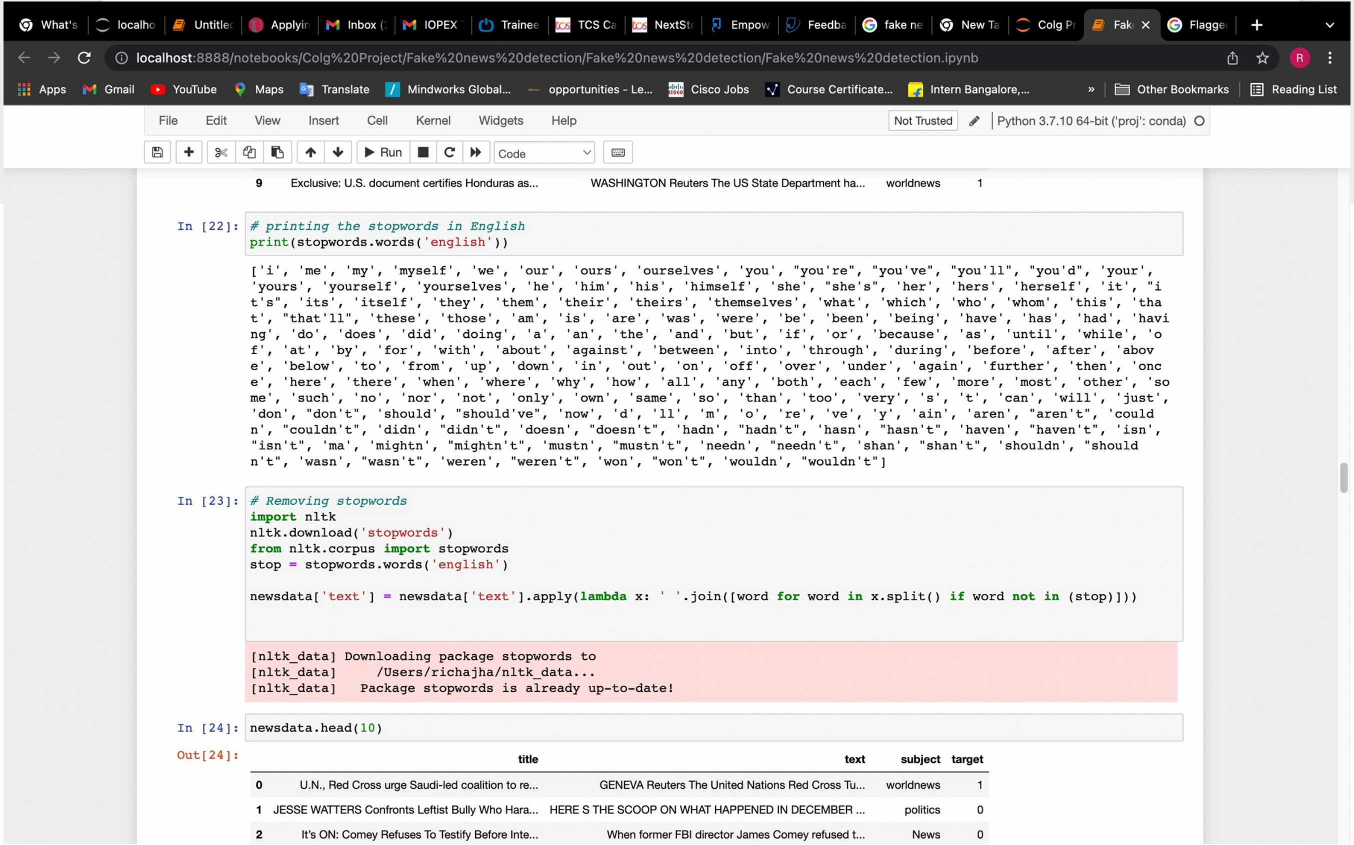
	title	text	subject	target
8	Researchers raise doubts over cause of Chilean...	SANTIAGO (Reuters) - International researchers...	worldnews	1
9	Exclusive: U.S. document certifies Honduras as...	WASHINGTON (Reuters) - The U.S. State Departme...	worldnews	1

In [21]:

```
# Removeing punctuation

import string

def punctuation_removal(text):
    all_list = [char for char in text if char not in string.punctuation]
    clean_str = ''.join(all_list)
    return clean_str

newsdata['text'] = newsdata['text'].apply(punctuation_removal)
newsdata.head(10)
```

Out[21]:

	title	text	subject	target
0	U.N., Red Cross urge Saudi-led coalition to re...	GENEVA Reuters The United Nations and Red Cro...	worldnews	1
1	JESSE WATTERS Confronts Leftist Bully Who Hara...	HERE S THE SCOOP ON WHAT HAPPENED IN DECEMBER ...	politics	0
2	It's ON: Comey Refuses To Testify Before Inte...	When former FBI director James Comey refused t...	News	0
3	Libyan PM Sarraj hopes for easing of arms embargo	WASHINGTON Reuters Libyan Prime Minister Faye...	worldnews	1
4	Rep. Franks to resign after staff members' com...	WASHINGTON Reuters Republican US Representati...	politicsNews	1
5	The Trump Administration Just Wrote North Kor...	North Korea has been launching intermediate r...	News	0
6	Germany keen to avoid new 'ice age' in ties be...	WASHINGTON Reuters Germany and Europe want to...	worldnews	1
7	Donald Trump's TRILLION Dollar Bombshell	21st Century Wire says Is this the biggest new...	Middle-east	0



	title	text	subject	target
9	Exclusive: U.S. document certifies Honduras as...	WASHINGTON Reuters The US State Department ha...	worldnews	1

In [22]:

```
# printing the stopwords in English
print(stopwords.words('english'))
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your',
'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "i
t's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'tha
t', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'havi
ng', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'o
f', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'abov
e', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'onc
e', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'so
me', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just',
'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'could
n', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn',
"isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "should
n't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

In [23]:

```
# Removing stopwords
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
stop = stopwords.words('english')

newsdata['text'] = newsdata['text'].apply(lambda x: ' '.join([word for word in x.split() if word not in (stop)]))
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/richajha/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [24]:

```
newsdata.head(10)
```

Out[24]:

	title	text	subject	target
0	U.N., Red Cross urge Saudi-led coalition to re...	GENEVA Reuters The United Nations Red Cross Tu...	worldnews	1
1	JESSE WATTERS Confronts Leftist Bully Who Hara...	HERE S THE SCOOP ON WHAT HAPPENED IN DECEMBER ...	politics	0
2	It's ON: Comey Refuses To Testify Before Inte...	When former FBI director James Comey refused t...	News	0

which occurs frequently and that are of no use. Like; the, a, on, for, etc.

- **Step 4:-**

Next Step is “Data Visualization ”.

Data visualization is a graphical representation of data. Here, data are transformed to piecharts, graphs, bar graphs, histograms, etc. For data visualization python libraries like matplotlib, seaborn, etc are used.

CHAPTER 4

REFERENCES

[1] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, “Fake News Detection on Social Media: A Data Mining Perspective” arXiv:1708.01967v3 [cs.SI], 3 Sep 2017

[2] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903.

[3] Fake news websites. (n.d.) Wikipedia. [Online]. Available: https:// en.wikipedia.org/wiki/Fake_news_website. Accessed Feb. 6, 2017

[4] Conroy, N., Rubin, V. and Chen, Y. (2015). “Automatic deception detection: Methods for finding fake news” at Proceedings of the Association for Information Science and Technology, 52(1), pp.1-4.

[5] Markines, B., Cattuto, C., & Menczer, F. (2009, April). “Social spam detection”. In Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web (pp. 41-48)