WELCOME

IBM NAAN MUTHALVAN PROJECT

ARTIFICIAL INTELLIGENCE(AI)

Project Title : Fake News Detection Using NLP

BATCH MEMBERS:

AKILA J -951321106001

DAYANA S -951321106007

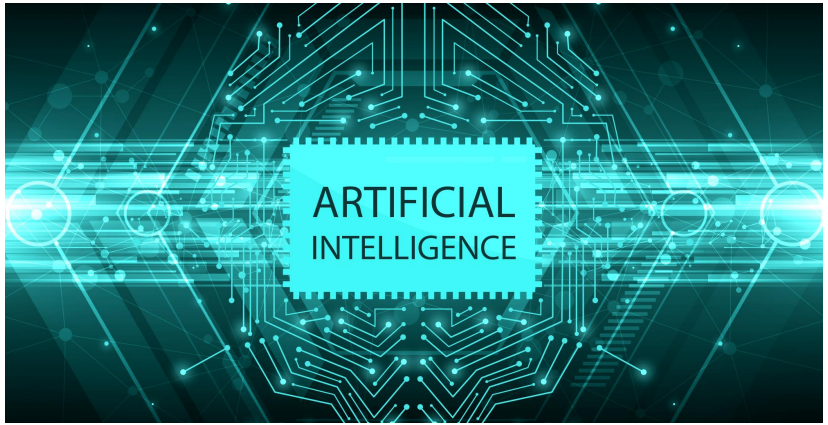
DIVYA DHARSHINI R -951321106008

JERLIN PRABHA J -951321106018

RATHINA MALA B -951321106039

MENTOR:

Dr.M.RUBAN GLADWIN,M.E,Ph.d,(AP/ECE)



INTRODUCTION OF AI:

Intelligence:"The capacity to learn and solve problems'.

Artificial intelligence is the simulation of humans intelligence by machines.

INTRODUCTION OF FAKE NEWS DETECTION :

Fake news is the intentional broadcasting of false or misleading claims as news, where the statements are purposely deceitful.

Problem Statement :

The fake news dataset is one of the classic text analytics datasets available on Kaggle. It consists of genuine and fake articles' titles and text from different authors. Our job is to create a model which predicts whether a given news is real or fake.

Problem Defnintion:

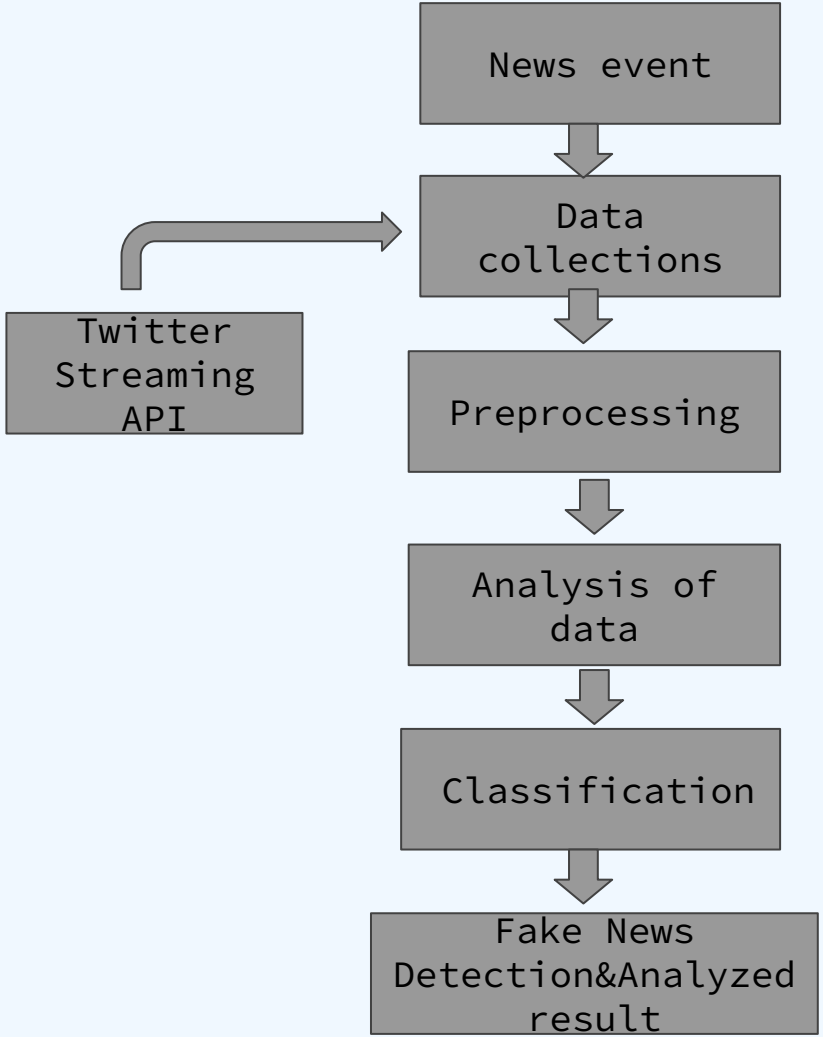
The problem is to develop a fake news detection model using a Kaggle dataset. The goal is to distinguish between genuine and fake news articles based on their titles and text. This project involves using natural language processing (NLP) techniques to preprocess the text data, building a machine learning model for classification, and evaluating the model's performance.

- Input: News items, social contexts and associated side information
- Output: One of two labels: 'fake' or 'real'.

Design Thinking:

1. Data Source: Choose the fake news dataset available on Kaggle, containing articles titles and text, along with their labels (genuine or fake).
2. Data Preprocessing: Clean and preprocess the textual data to prepare it for analysis.
3. Feature Extraction: Utilize techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings to convert text into numerical features.
4. Model Selection: Select a suitable classification algorithm (e.g., Logistic Regression, Random Forest, or Neural Networks) for the fake news detection task.
5. Model Training: Train the selected model using the preprocessed data.
6. Evaluation: Evaluate the model's performance using metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

BLOCK DIAGRAM:



Software Tools & Program Languages:

---

1) pandas (gateway into the world of data science)

2) numby (fundamental Python package for scientific computations)

3) sklearn (the most useful and robust library for machine learning in Python and classification of the function)

4) TfidfVectorizer

Programming language:

1) Python

Innovation techniques:

IT'S SOLUTION:

It is vital to recognize and differentiate between false and accurate news. One method is to have an expert decide, and fact checks every piece of information, but this takes time and needs expertise that cannot be shared. Secondly, we can use machine learning and artificial intelligence tools to automate the identification of fake news.

Online news information includes various unstructured format data (such as documents, videos, and audio), but we will concentrate on text format news here. With the progress of machine learning and Natural language processing, we can now recognize the misleading and false character of an article or statement.

Data Exploration

In this work, we utilized the fake news dataset from Kaggle to classify untrustworthy news articles as fake news. We have a complete training dataset containing the following characteristics:

Data Cleaning for Analysis

In this section, we will clean our dataset to do some analysis:

- Drop unused rows and columns.
- Perform null value imputation.
- Remove special characters.
- Remove stop words.

Data cleaning for program:

```
1.# Constants that are used to sanitize the datasets

2.column_n = ['id', 'title', 'author', 'text', 'label']
3.remove_c = ['id','author']
4.categorical_features = []
5.target_col = ['label']

6.text_f = ['title', 'text']
```

```
# Clean Datasets

import nltk

from nltk.corpus import stopwords

import re
 ▬▬ ▬▬ ▬▬
from nltk.stem.porter import PorterStemmer

from collections import Counter


ps = PorterStemmer()

wnl = nltk.stem.WordNetLemmatizer()


stop_words = stopwords.words('english')

stopwords_dict = Counter(stop_words)


# Removed unused clumns

def remove_unused_c(df,column_n=remove_c):

    df = df.drop(column_n,axis=1)

    return df
```

```
# Cleaning text from unused characters

def clean_text(text):

    text = str(text).replace(r'http[\w:/\.]+', ' ')  # removing urls

    text = str(text).replace(r'[^\.\w\s]', ' ')  # remove everything but characters and punctuation

    text = str(text).replace('[^a-zA-Z]', ' ')

    text = str(text).replace(r'\s\s+', ' ')

    text = text.lower().strip()

    #text = ' '.join(text)

    return text

## Nltk Preprocessing include:

# Stop words, Stemming and Lemmetization

# For our project we use only Stop word removal

def nltk_preprocess(text):

    text = clean_text(text)

    wordlist = re.sub(r'[^\w\s]', '', text).split()

    #text = ' '.join([word for word in wordlist if word not in stopwords_dict])

    #text = [ps.stem(word) for word in wordlist if not word in stopwords_dict]

    text = ' '.join([wnl.lemmatize(word) for word in wordlist if word not in stopwords_dict])

    return  text
```

Preprocessing:

```
# Perform data cleaning on train and test dataset by calling clean_dataset function

df = clean_dataset(news_d)

# apply preprocessing on text through apply method by calling the function nltk_preprocess

df["text"] = df.text.apply(nltk_preprocess)

# apply preprocessing on title through apply method by calling the function nltk_preprocess

    df["title"] = df.title.apply(nltk_preprocess)
```

Copy

```
# Dataset after cleaning and preprocessing step

    df.head()
```

Copy

Output:

	title	text	label
0	house dem aide didnt even see comeys letter ja..	house dem aide didnt even see comeys letter ja..	1
1	flynn hillary clinton big woman campus breitbart	ever get feeling life circle roundabout rather..	0
2	truth might get fired	truth might get fired october 29 2016 tension ...	1
3	15 civilian killed single u airstrike identified	video 15 civilian killed single u airstrike id...	1
4	iranian woman jailed fictional unpublished sto..	print iranian woman sentenced six year pri	

Let's make a function that accepts the article text as an argument and return whether it's fake or not:

```
def get_prediction(text, convert_to_label=False):
    # prepare our text into tokenized sequence
    inputs = tokenizer(text, padding=True, truncation=True, max_length=max_length, return_tensors="pt").to("cuda")
    # perform inference to our model
    outputs = model(**inputs)
    # get output probabilities by doing softmax
    probs = outputs[0].softmax(1)
    # executing argmax function to get the candidate label
    d = {
        0: "reliable",
        1: "fake"
    }
    if convert_to_label:
      return d[int(probs.argmax())]
    else:
        return int(probs.argmax())
```

```
real_news = """

Tim Tebow Will Attempt Another Comeback, This Time in Baseball - The New York Times",Daniel Victor,"If at first you don’t succeed, try a different sport. Tim Tebow, who was a Heisman quarterback at the University of Florida but was unable to hold an N. F. L. job, is pursuing a career in Major League Baseball. <SNIPPED>

"""

get_prediction(real_news, convert_to_label=True)
```

OUTPUT:

reliable

Steps For Detecting Fake News with Python:

– – –

Follow the below steps for detecting fake news and complete your first advanced Python Project –

Step 1:Make necessary imports:

```
1. import numpy as np

2.import pandas as pd

3.import itertools
4.from sklearn.model_selection import train_test_split
5.from sklearn.feature_extraction.text import TfidfVectorizer
6.from sklearn.linear_model import PassiveAggressiveClassifier
7.from sklearn.metrics import accuracy_score, confusion_matrix
```

Output screenshot:

— — —

step2:. Now, let's read the data into a DataFrame, and get the shape of the data and the first 5 records

```
1.#Read the data
 2.df=pd.read_csv('D:\\DataFlair\\news.csv')
 3.#Get shape and head
 4.df.shape
 5.df.head()
```

output:

```
[2]: #Read the data
     df=pd.read_csv('D:\\DataFlair\\news.csv')

     #Get shape and head
     df.shape
     df.head()
```

[2]:

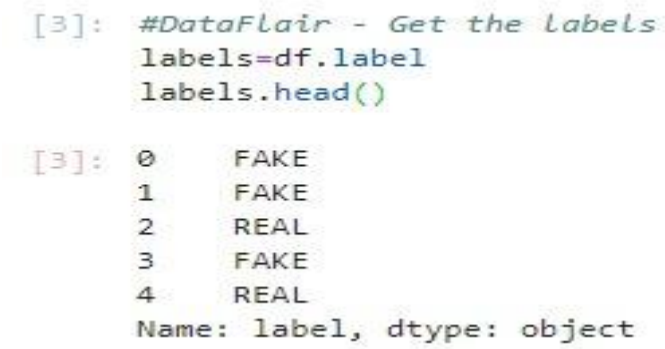
	Unnamed: 0	title	text	label
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...	FAKE
4	875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...	REAL

Output for data Frame:

step3: And get the labels from the DataFrame.

```
1.#DataFlair - Get the labels
2.labels=df.label
3.labels.head()
```

output:



Split the dataset:

– – –

step4: Split the dataset into training and testing sets.

```
1.#DataFlair - Split the dataset

2.x_train,x_test,y_train,y_test=train_test_split(df['text'], labels, test_size=0.2, random_state=7)
```

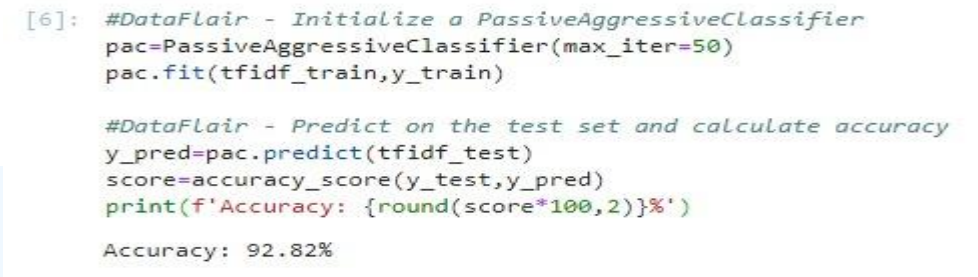
step5: Let's initialize a TfidfVectorizer with stop words from the English language and a maximum document frequency of 0.7 (terms with a higher document frequency will be discarded). Stop words are the most common words in a language that are to be filtered out before processing the natural language data. And a TfidfVectorizer turns a collection of raw documents into a matrix of TF-IDF features.

Classified the dataset:

```
1.#DataFlair - Initialize a TfidfVectorizer
2.tfidf_vectorizer=TfidfVectorizer(stop_words='english', max_df=0.7)
3.#DataFlair - Fit and transform train set, transform test set
4.tfidf_train=tfidf_vectorizer.fit_transform(x_train)
5.tfidf_test=tfidf_vectorizer.transform(x_test)
```

step6:. Next, we'll initialize a PassiveAggressiveClassifier. This is. We'll fit this on tfidf_train and y_train.

Then, we'll predict on the test set from the TfidfVectorizer and calculate the accuracy with accuracy_score() from sklearn.metrics. **output:**

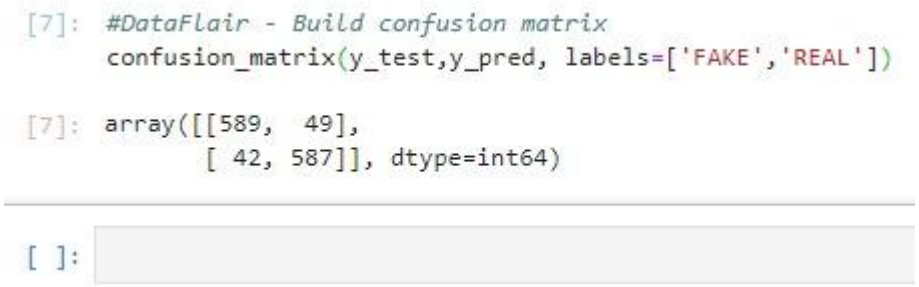


Fake News Detection & Analysed result:

Step7: We got an accuracy of 92.82% with this model. Finally, let's print out a confusion matrix to gain insight into the number of false and true negatives and positives.

```
1.#DataFlair - Build confusion matrix
2.confusion_matrix(y_test,y_pred, labels=['FAKE','REAL'])
```

Output screenshot:



Source code:

```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load
import warnings
warnings.filterwarnings('ignore')
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

import nltk
from nltk.sentiment import SentimentIntensityAnalyzer
import warnings
warnings.filterwarnings("ignore")
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, classification_report

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save & Run All"
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session
```

```
/kaggle/input/fake-and-real-news-dataset/True.csv
/kaggle/input/fake-and-real-news-dataset/Fake.csv
```

Loading Data:

```
true = pd.read_csv('/kaggle/input/fake-and-real-news-dataset/True.csv')
fake = pd.read_csv('/kaggle/input/fake-and-real-news-dataset/Fake.csv')In [3]:
```

linkcode

```
fake['Category'] = 'fake'
```

```
fake
```

output:

s.n	title	text	subject	date	category
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	fake
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	fake
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	fake
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	fake
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	fake

23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	fake
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016	fake
23478	Sunnistan: US and Allied ‘Safe Zone’ Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016	fake
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016	fake
23480	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016	fake

```
true['Category'] = 'true'
True
```

Outout:

s.no	Title	text	subject	date	Category
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	true
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	true
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	true
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	true
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	true
...	...	...	...	...	...

21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	true
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	true
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	true
21415	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	true
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	true

```
#Now let's combine the whole dataset into one
data = pd.concat([fake, true], ignore_index = True)
data
```

s.no	title	text	subject	date	Category
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	fake
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	fake
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	fake
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	fake
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	fake

44893	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	true
44894	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	true
44895	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	true
44896	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	true
44897	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	true

```
Data.shape
```

Out[6]:

(44898, 5)

Preprocessing:

```
data['Category'].value_counts()
```

Out[7]:

```
Category
fake    23481
true    21417

    Name: count, dtype: int64
```

In [8]:

```
#Transforming category values to numerical
from sklearn.preprocessing import LabelEncoder
encoder = LabelEncoder()
data['Category'] = encoder.fit_transform(data['Category'])
```

In [9]:

```
data['Category']
```

Out[9]:

```
0        0
1        0
2        0
3        0
4        0
       ..
44893    1
44894    1
44895    1
44896    1
44897    1

    Name: Category, Length: 44898, dtype: int64
```

In [10]:

```
title = vectorizer.fit_transform(data['title'])
title
```

Out[10]:

```
<44898x20896 sparse matrix of type '<class 'numpy.float64'>'
	with 546512 stored elements in Compressed Sparse Row format>
```

Modeling:

```
from sklearn.model_selection import train_test_split
X = title
y = data['Category']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2,
random_state = 42)
```

In [12]:

```
model = SVC()
model.fit(X_train, y_train)
```

Out[12]:

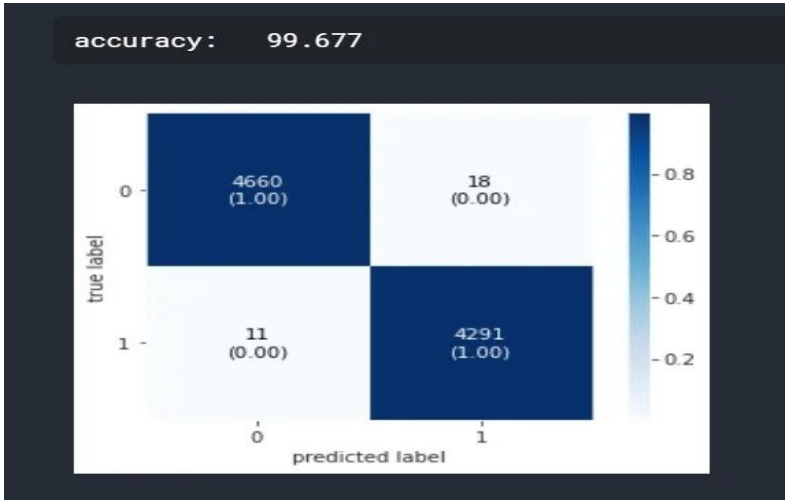
```
SVC
SVC()
```

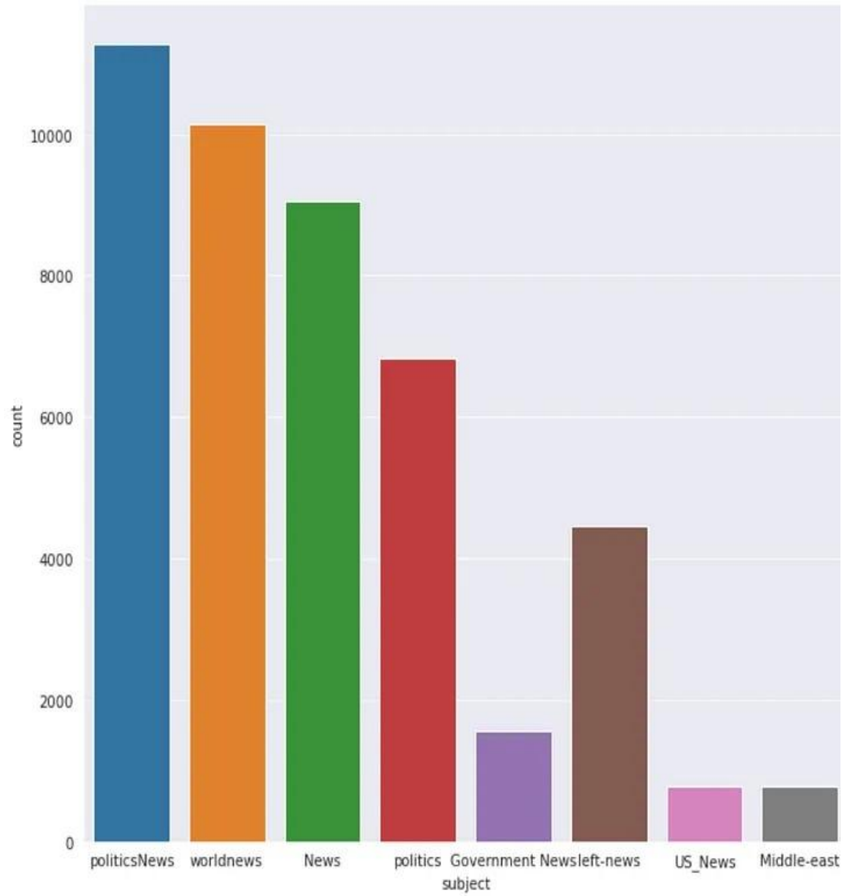
```
y_pred = model.predict(X_test)
print('Classification Report: ')
print(classification_report(y_test, y_pred))                    In[13]:
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.96      0.96      4733
           1       0.95      0.97      0.96      4247

    accuracy                           0.96      8980
   macro avg       0.96      0.96      0.96      8980
weighted avg       0.96      0.96      0.96      8980
```





count
subject
0
2000
4000
6000
8000
10000
politicsNews
worldnews
News
politics
Government News
left-news
US_News
Middle-east

Donald Trump
one of
wa
Hillary Clinton
according to
going to
United State
Featured image
the United
White House
thi
hi
would be
Trump is
the White
President Obama
the people
the new
Fox New
the Republican
twitter com
This is
pic twitter
the president
image via
President Trump
to get
to make
want to
Getty Image
New York
the world
the first
trying to
said the
at least
need to
Trump and
to take
he said
as well
that Trump
In the
the election
the media
the American
the FBI
the time
the right
part of
the Obama
the fact
Trump has
Trump supporter
the last
would have
the GOP
doe
said he
the Clinton
the country
ha
people who
fact that
the US
able to
including
the Democrat
of State
year old
of course
the state

Reference:

1)Artificial Intelligence for Humans by Jeff Heaton https://youtu.be/f7ckMaEJvh4?si=0EYjBfRCUjoKahgf

2)https://youtu.be/2o-_bMKGi_o?si=y_L8U6EVfIU_tc93

3)https://youtu.be/DdgymahD7Zg?si=MTWrP0W1s7QvSQid

4) Combating Fake News with Computational Intelligence Techniques

books.googl

Mohamed Lahby, Al-Sakib Khan Pathan, Yassine Maleh · 2021

5)https://books.google.co.in/books?id=GfFUEAAAQBAJ&printsec=frontcover&dq=fake+news+detection+using+python+and+nlp+reference&hl=en&newbks=1&newbks_redir=1&sa=X&ved=2ahUKEwj7hOiU3uOBAxWsbWwGHdD7BFYQ6AF6BAgFEAI

Fake news Detection using NLP

THANKING YOU