A

Project Report

On

Detecting Fake News With Python And Machine Learning

Submitted in partial fulfilment of the requirement for the 3rd semester

BTECH

By

Yash Sharma 2017586

Under the Guidance of

Dr. Vishen Gupta Sir

Deptt. of Computer Science & Application

DEPARTMENT OF COMPUTER SCIENCE & APPLICATION

GRAPHIC ERA UNIVERSITY, DEHRADUN

CLEMENT TOWN, BELL ROAD
DISTRICT- DEHRADUN-248002
2021 – 2022

GraphicEra
(Deemed to be University)
Accredited by NAAC with Grade A

CANDIDATE’S DECLARATION

I hereby certify that the work which is being presented in the dissertation entitled “Detecting Fake News Using Python And Machine Learning” in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering, submitted in the Department of Computer Science and Engineering of the Graphic Era Deemed to be University, Dehradun is an authentic record of my own work carried out during a period from October 2021 to February 2022, under the supervision of Dr. Vishen Gupta Sir, Department of Computer Science and Engineering of the Graphic Era Deemed to be University, Dehradun (Uttarakhand).

The matter presented in this dissertation has not been submitted by me for the award of any other degree of this or any other Institute/University.

YASH SHARMA

2017586

ACKNOWLEDGEMENT

Here by I am submitting the project report on "Detecting Fake News using Python And Machine Learning", as per the scheme of Graphic Era University, Dehradun.

In this connection, I would like to express our deep sense of gratitude to our beloved institution Graphic Era University and, I like to express our sincere gratitude and indebtedness to Hon'ble Dr. Kamal Ghanshala, Chairman & President, GEU, Dehradun

I would like to express our sincere gratitude to Dr. Devesh Pratap Singh, Head of Dept. of Computer Science, for providing a congenial environment to work in and carry out our project.

I consider it our cardinal duty to express the deepest sense of gratitude to Dr. Vishen Gupta, Department of Computer Science and Application for the invaluable guidance extended at every stage and in every possible way.

Finally, I am very much thankful to all the faculty members of the Department of Computer Science and Application, friends and our parents for their constant encouragement, support and help throughout the period of project conduction.

YASH SHARMA

2017586

TABLE OF CONTENTS

1. ABSTRACT

1.1 Introduction

This advanced python project of detecting fake news deals with fake and real news. Using sklearn, we build a TfidfVectorizer on our dataset. Then, Train the Model using Logistic Regression . In the end, the accuracy score and the confusion matrix tell us how well our model fares.

1.2 Motivation

Fake News has become one of the major problem in the existing society. Fake News has high potential to change opinions, facts and can be the most dangerous weapon in influencing society

With our world producing an ever-growing huge amount of data exponentially per second by machines, there is a concern that this data can be false (or fake). Fake news (or data) can pose many dangers to our world. Imagine what happens if due to some false information you are given the wrong medicine.

Luckily, this problem can be addressed using machine learning with python. We can develop a machine learning model in python which can detect whether the news is fake or not

2. REQUIREMENT OF PROJECT

2.1 Hardware Requirement

- A working computer
- Internet Connection

2.2 Software Requirement

- Internet Browser such as Chrome, Edge, etc.
- Google Colaboratory
- Some DataSets (Module)

3. METHODOLOGY FOLLOWED

3.1 Pandas Library :- Pandas allows us to analyze big data and make conclusions based on statistical theories.Pandas can clean messy data sets, and make them readable and relevant.Relevant data is very important in data science.

Pandas In this Project :-

Read_csv(‘Path of file’) by using this function we can read the csv file in this we Have to give the path of the csv file which we upload

3.2 Scikit-Learn (Sklearn) Library

Scikit-learn is an indispensable part of the Python machine learning toolkit at JPMorgan. It is very widely used in predictive analytics, and very many other machine learning tasks. Its straightforward API, its breadth of algorithms, and the quality of its documentation combine to make scikit-learn simultaneously very approachable and very powerful.

3.3 TfidfVectorizer

TF (Term Frequency): The number of times a word appears in a document is its Term Frequency. A higher value means a term appears more often than others, and so, the document is a good match when the term is part of the search terms.

IDF (Inverse Document Frequency): Words that occur many times a document, but also occur many times in many others, may be irrelevant. IDF is a measure of how significant a term is in the entire corpus.

The TfidfVectorizer converts a collection of raw documents into a matrix of TF-IDF features.

3.4 NUMPY

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. ... A powerful N-dimensional array object

3.5 stopwords

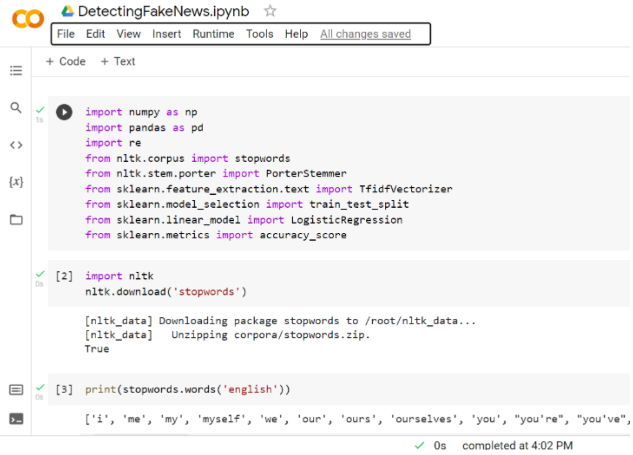
Stopwords are the English words which does not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. For example, the words like the, he, have etc. Such words are already captured this in corpus named corpus.

3.6 PorterStemmer

Porters Stemmer It is a type of stemmer which is mainly known for Data Mining and Information Retrieval. As its applications are limited to the English language only. It is based on the idea that the suffixes in the English language are made up of a combination of smaller and simpler suffixes, it is also majorly known for its simplicity and speed. The advantage is, it produces the best output from other stemmers and has less error rate.

4. Screenshots

Fig. 1



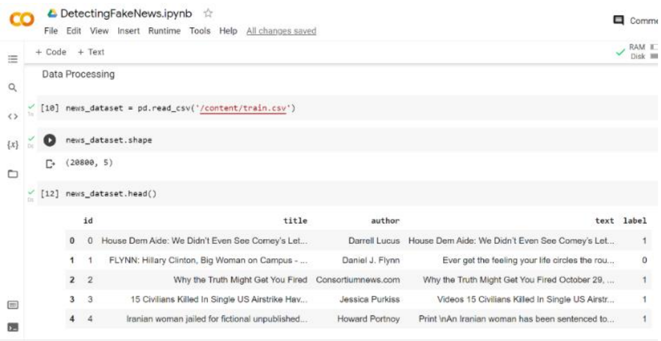
```
import numpy as np
import pandas as pd
import re
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

```
[2] import nltk
    nltk.download('stopwords')

    [nltk_data] Downloading package stopwords to /root/nltk_data...
    [nltk_data]   Unzipping corpora/stopwords.zip.
    True
```

```
[3] print(stopwords.words('english'))

    ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've",
```

Fig.1 shows importing required libraries, and use stopwords lib. And then print them

Fig.2



Data Processing

```
[10] news_dataset = pd.read_csv('/content/train.csv')
```

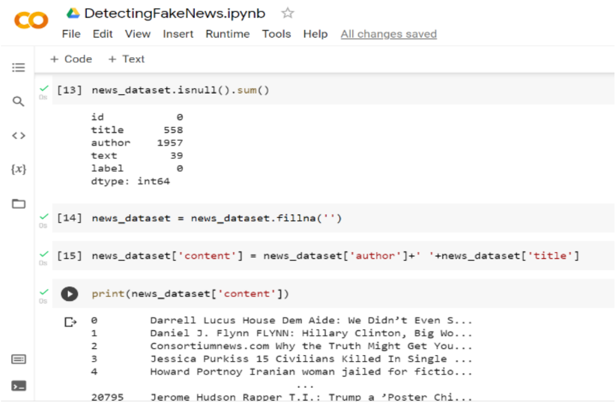
```
news_dataset.shape

(20800, 5)
```

```
[12] news_dataset.head()
```

	id	title	author	text	label
0	0	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1

Fig.2 Data Preprocessing. And print first 5 rows

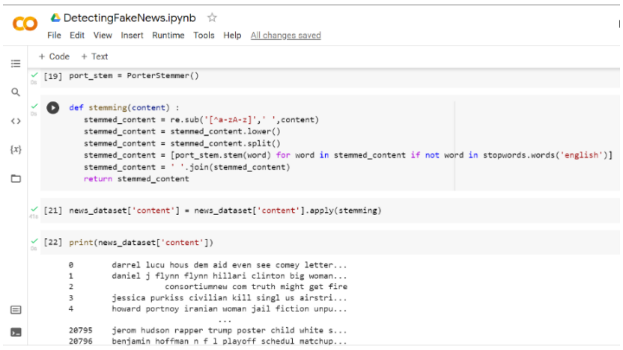
Fig. 3



```
[13] news_dataset.isnull().sum()

     id           0
     title      558
     author    1957
     text        39
     label        0
     dtype: int64

[14] news_dataset = news_dataset.fillna('')

[15] news_dataset['content'] = news_dataset['author']+' '+news_dataset['title']

     print(news_dataset['content'])

     0        Darrell Lucus House Dem Aide: We Didn’t Even S...
     1        Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
     2        Consortiumnews.com Why the Truth Might Get You...
     3        Jessica Purkiss 15 Civilians Killed In Single ...
     4        Howard Portnoy Iranian woman jailed for fictio...
                                    ...
     20795    Jerome Hudson Rapper T.I.: Trump a ’Poster Chi...
```

Fig 3. Shows Check Nulls Value and fill them with empty string and then print data

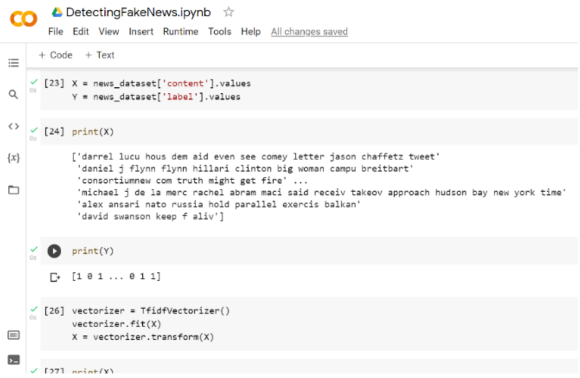
Fig 4.



```
[19] port_stem = PorterStemmer()

     def stemming(content) :
         stemmed_content = re.sub('[^a-zA-z]',' ',content)
         stemmed_content = stemmed_content.lower()
         stemmed_content = stemmed_content.split()
         stemmed_content = [port_stem.stem(word) for word in stemmed_content if not word in stopwords.words('english')]
         stemmed_content = ' '.join(stemmed_content)
         return stemmed_content

[21] news_dataset['content'] = news_dataset['content'].apply(stemming)

[22] print(news_dataset['content'])

     0        darrel lucu hous dem aid even see comey letter...
     1        daniel j flynn flynn hillari clinton big woman...
     2                 consortiumnew com truth might get fire
     3        jessica purkiss civilian kill singl us airstri...
     4        howard portnoy iranian woman jail fiction unpu...
                                  ...
     20795    jerom hudson rapper trump poster child white s...
     20796    benjamin hoffman n f l playoff schedul matchup...
```

Fig 4. Shows Performining of stemming function :- Stemming is the process of reducing a word to its Root word

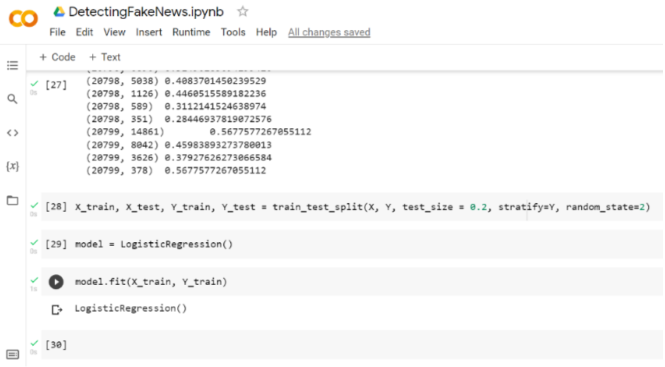
Fig. 5



```
[23] X = news_dataset['content'].values
     Y = news_dataset['label'].values

[24] print(X)

     ['darrel lucu hous dem aid even see comey letter jason chaffetz tweet'
      'daniel j flynn flynn hillari clinton big woman campu breitbart'
      'consortiumnew com truth might get fire' ...
      'michael j de la merc rachel abram maci said receiv takeov approach hudson bay new york time'
      'alex ansari nato russia hold parallel exercis balkan'
      'david swanson keep f aliv']

     print(Y)

     [1 0 1 ... 0 1 1]

[26] vectorizer = TfidfVectorizer()
     vectorizer.fit(X)
     X = vectorizer.transform(X)

[27] print(X)
```

Fig.5 Printing Label and content (author , text). Column after seprating and then shows converting the texts into feature vectors so that we can find cosine similarity of the values. Shown below

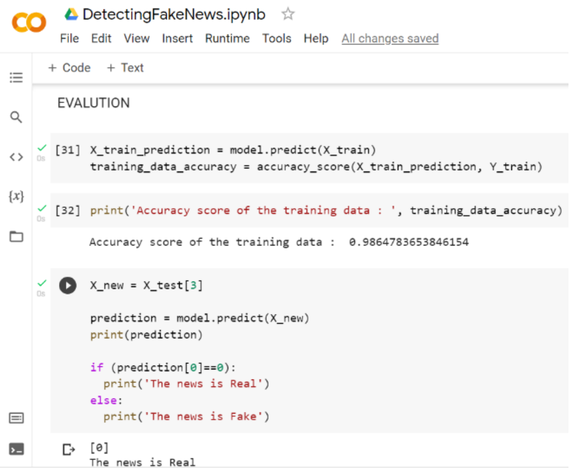
Fig. 6



```
[27]  (20798, 5038) 0.4083701450239529
      (20798, 1126) 0.4460515589182236
      (20798, 589)  0.3112141524638974
      (20798, 351)  0.28446937819072576
      (20799, 14861)        0.5677577267055112
      (20799, 8042) 0.45983893273780013
      (20799, 3626) 0.37927626273066584
      (20799, 378)  0.5677577267055112

[28] X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, stratify=Y, random_state=2)

[29] model = LogisticRegression()

     model.fit(X_train, Y_train)

     LogisticRegression()

[30]
```

Fig.6 Split the dataset into training and testing sets.

Fig. 7



EVALUTION

```
[31] X_train_prediction = model.predict(X_train)
     training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

[32] print('Accuracy score of the training data : ', training_data_accuracy)

     Accuracy score of the training data :  0.9864783653846154

     X_new = X_test[3]

     prediction = model.predict(X_new)
     print(prediction)

     if (prediction[0]==0):
       print('The news is Real')
     else:
       print('The news is Fake')

     [0]
     The news is Real
```

Fig.7 Making the predictive System. We got an accuracy of 92.64% with this model. And get result of given news is true or false

5. Conclusion

In this Project we predict the news is Fake or Not By using by Python Module Like Numpy,pandas

Etc. The completion of the project went quiet well, I learned much new things while I was building up it, and I get up to know various platforms which help us to learn all this stuff. I was able to learn the practical use of ML and Python .

Overall working on this project was great fun as I came up with great piece of knowledge and understanding of the topic

6. References

https://www.delftstack.com https://www.etutorialspoint.com

https://www.geeksforgeeks.org https://www.javatpoint.com

https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html

https://drive.google.com/file/d/1wxbX1pmdjAinILQS21YIbDJDz6I8

8qks/view?usp=sharing

https://drive.google.com/file/d/1Ef8Pmxp6KHBx2bU6eOxnhnth53T9

mr9f/view?usp=sharing

www.youtube.com