Fake news detection using NLP

Project title	**Fake news detection using NLP**
Skills taken away from this project	• **Python scripting** • **Data Preprocessing** • **Machine learning and NLP**
Domain	**Multimedia**

Introduction:

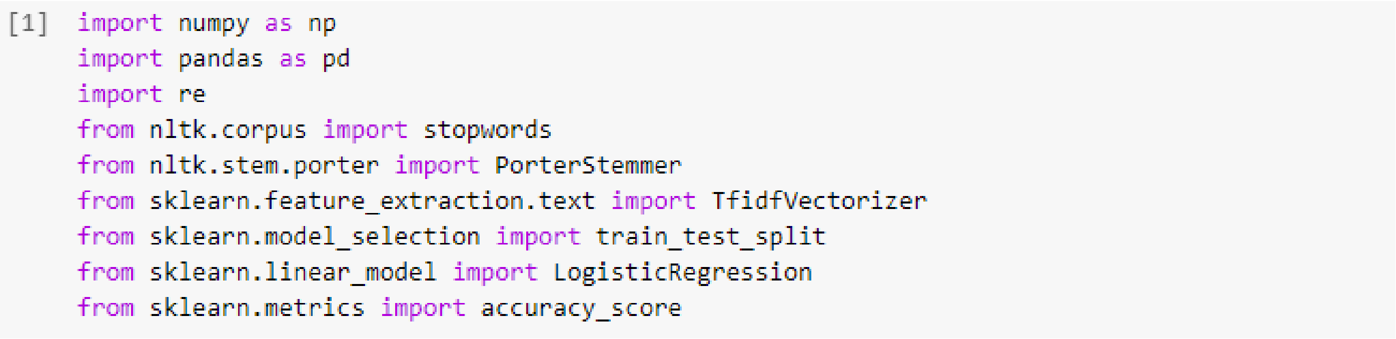
Fake news detection using Natural Language Processing (NLP) is a critical field of research and application aimed at identifying and mitigating the spread of misleading or false information in digital media. With the rapid expansion of social media and online news platforms, the dissemination of misinformation has become a pressing concern. NLP, a sub field of artificial intelligence, plays a pivotal role in addressing this issue by leveraging techniques from linguistics and machine learning to analyze and understand text data.

Objective:

- Fake news detection using machine learning is to develop a model or system that can automatically identify and classify news articles or information as either “real” or “fake”

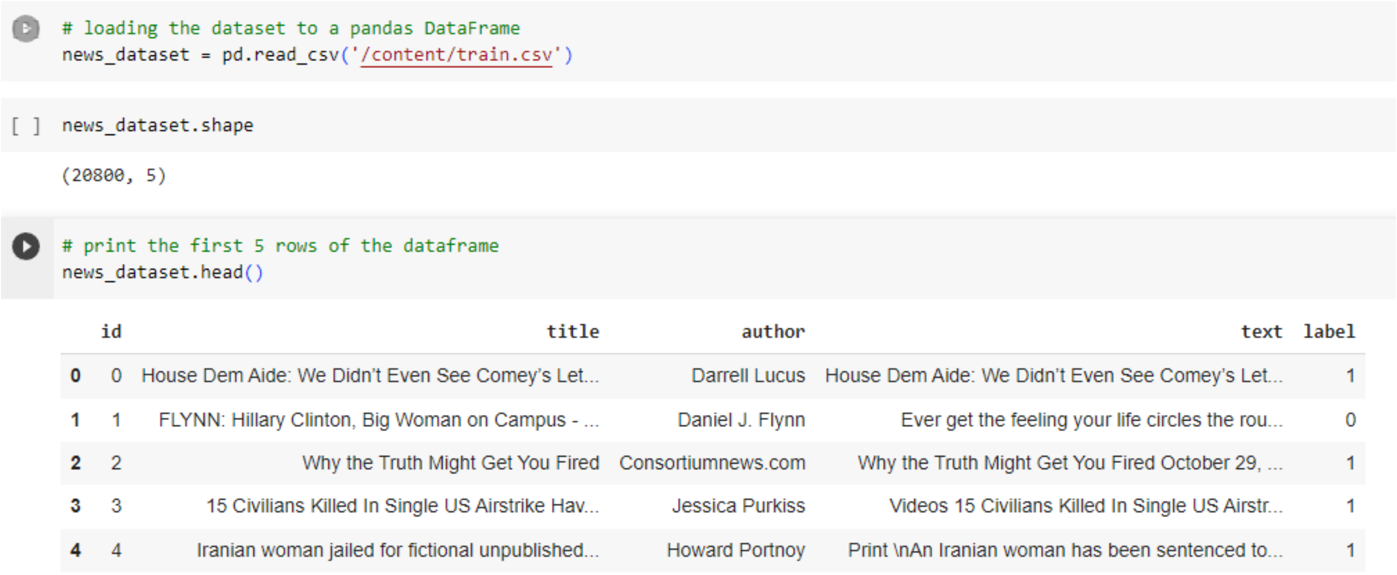
Library Installation:

Import the necessary libraries for this project



```
[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
```

Import Data:



```
# loading the dataset to a pandas DataFrame
news_dataset = pd.read_csv('/content/train.csv')
```

```
[ ] news_dataset.shape
```

```
(20800, 5)
```

```
# print the first 5 rows of the dataframe
news_dataset.head()
```

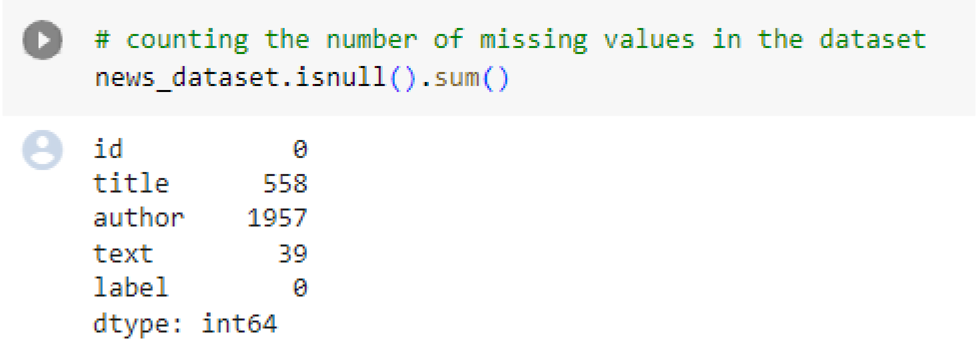
	id	title	author	text	label
0	0	House Dem Aide: We Didn’t Even See Comey’s Let...	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You 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

Import Data:

In this I used fake news data set from Kaggle

Data Preprocessing:

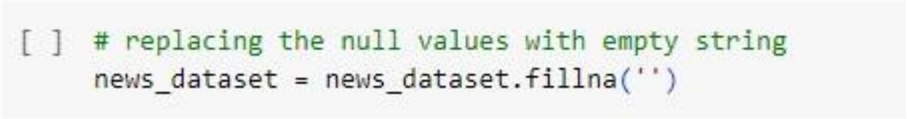
a) Missing value analysis:



```
# counting the number of missing values in the dataset
news_dataset.isnull().sum()
```

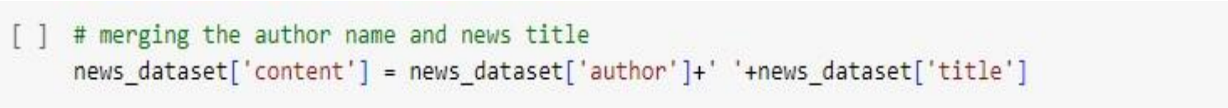
```
id           0
title      558
author    1957
text        39
label        0
dtype: int64
```

b) Fill the missing value:

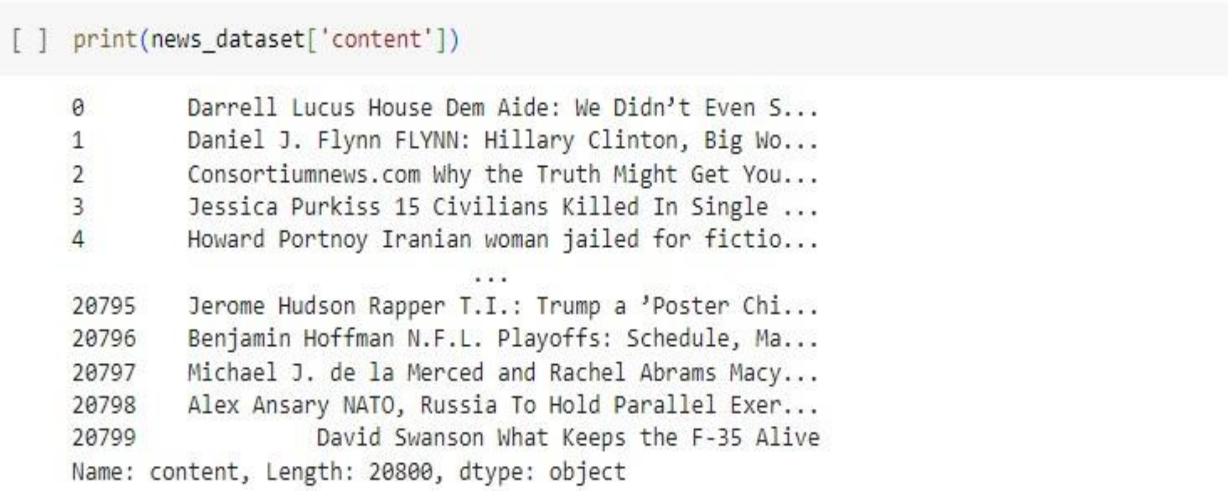


```
[ ] # replacing the null values with empty string
    news_dataset = news_dataset.fillna('')
```

c) Merging the author name & title:

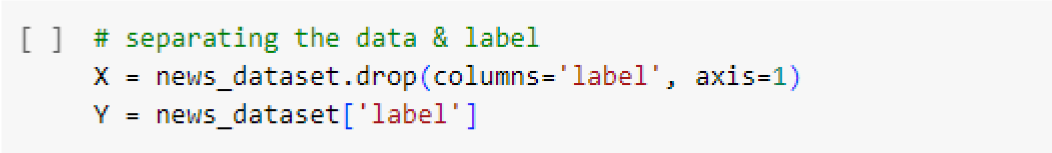


```
[ ]  # merging the author name and news title
     news_dataset['content'] = news_dataset['author']+' '+news_dataset['title']
```



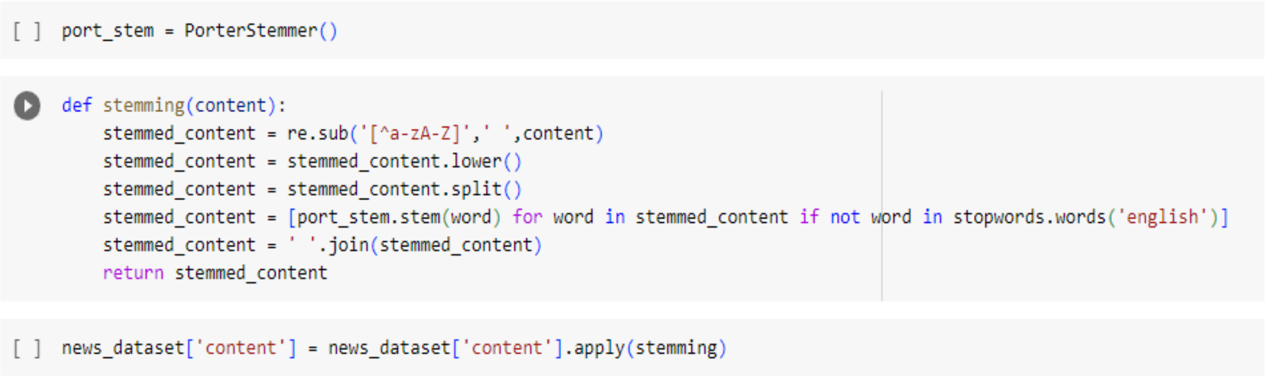
```
[ ]  print(news_dataset['content'])

     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...
     20796    Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma...
     20797    Michael J. de la Merced and Rachel Abrams Macy...
     20798    Alex Ansary NATO, Russia To Hold Parallel Exer...
     20799              David Swanson What Keeps the F-35 Alive
     Name: content, Length: 20800, dtype: object
```

d) Spreading the data & label:



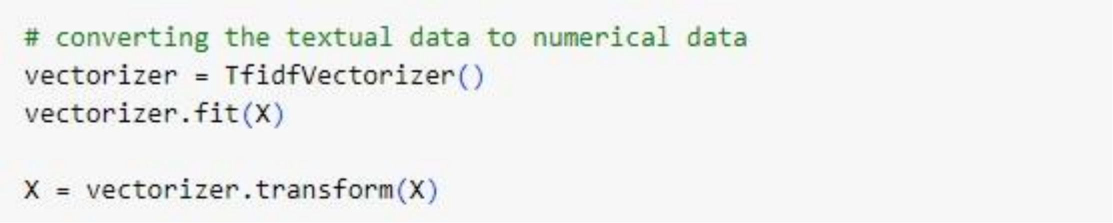
```
[ ]  # separating the data & label
     X = news_dataset.drop(columns='label', axis=1)
     Y = news_dataset['label']
```

e) Stemming process:



```
[ ]  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

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

f) Text to numerical data:



```
# converting the textual data to numerical data
vectorizer = TfidfVectorizer()
vectorizer.fit(X)

X = vectorizer.transform(X)
```

From this we can infer that after completion of the data pre-processing, The data has been cleaned by missing value analysis, fill the missing value, merging the author name& title,spreading the data and stemming process from the data set. Here, we can see that the data set has been organised, cleaned, and transformed so that it may be used for further analysis and to train a machine learning model.