DATE:18-10-2023

PHASE-3:DEVELOPMENT

AIM:

Load the fake news dataset and preprocess the textual data.

IMPORTING LIBRARIES

[1]:
```
# import modules
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import re
import nltk
import warnings
import sklearn
%matplotlib inline

warnings.filterwarnings('ignore')
```

LOADING THE DATASET

[2]:
```
true_data = pd.read_csv('True.csv')
fake_data = pd.read_csv('Fake.csv')

true_data['Target'] = ['True'] * len(true_data)
fake_data['Target'] = ['Fake'] * len(fake_data)

# Concatenate the data frames using pd.concat
data = pd.concat([true_data, fake_data]).sample(frac=1).reset_index(drop=True)
print(data.shape)
data.head()
```

```
(44898, 5)
```

[2]:

	title	text	subject	date	Target
0	Rights groups urge EU, Japan to consider halt ...	BANGKOK (Reuters) - Rights groups on Wednesday...	worldnews	October 18, 2017	True
1	WATCH: IRRELEVANT DEM POLITICAL ANALYST James ...	On Friday s broadcast of HBO s Real Time, fo...	left-news	Oct 21, 2017	Fake
2	Trump Asks O'Reilly, 'Do you think our country...	21st Century Wire says Regardless of what one ...	US_News	February 6, 2017	Fake
3	Factbox: Trump fills top jobs for his administ...	(Reuters) - U.S. President-elect Donald Trump ...	politicsNews	December 5, 2016	True
4	ONE LAST TIME ON OUR DIME: Mooch and Barack Ar...	The hard working First Family, in need of an...	politics	Aug 6, 2016	Fake

[3]:
```
data['title'][0]
```

[3]:
```
'Rights groups urge EU, Japan to consider halt in funding for  Cambodian election'
```

```
[4]: data['text'][0]
```

```
[4]: 'BANGKOK (Reuters) - Rights groups on Wednesday urged the European Union and Japan to consider halting their funding for the election panel in Cambodia,
if the ruling party succeeds in a bid to dissolve the main opposition party ahead of next year s general election. The ruling Cambodia People s Party (C
PP) has launched a crackdown on its critics, including politicians, independent media and non-government bodies. Nearly half the opposition members of p
arliament have fled abroad since September. In a session boycotted by the opposition, Cambodia s parliament voted on Monday to change party laws to re-d
istribute seats if a party is dissolved. The measure came after the government filed a lawsuit this month seeking to dissolve the main opposition Cambod
ia National Rescue Party (CNRP).  If the government s position to dissolve the opposition Cambodia National Rescue Party succeeds, next year s election
will be a joke,  Phil Robertson, deputy director for Asia at New York-based group Human Rights Watch, told Reuters.  At that point, both the EU and Japa
n should face reality and terminate their financial and technical assistance to avoid lending credibility to what will be a charade of democracy,  he ad
ded, speaking after a news conference in Bangkok. Japan and the EU are the two biggest foreign funders of the 2018 vote. China and the United States hav
e also contributed, with the United States providing trucks and technical support, while Japan has given computers. Japan s embassy in Phnom Penh did no
t reply to a Reuters request for comment on the matter. George Edgar, head of the EU delegation to Cambodia, said the EU remains  ready to support a cre
dible electoral process  but added that the polls should only go ahead with the opposition s involvement. He urged Cambodian authorities not to go ahead
with the dissolution of the opposition party.  The EU remains ready to support a credible electoral process up to the National Assembly election in 201
8. However we do not believe that a process from which the main opposition party was arbitrarily excluded could be seen as legitimate,  Edgar told Reute
rs. Cambodia s election commission was not immediately available for comment. The CPP has used  dirty tricks  to jail opposition leaders and force other
s into exile, Robertson added. CNRP leader Kem Sokha was arrested on Sept. 3 and charged with treason after the government said he had conspired with fo
reign advisers to topple it. Cambodia is relying on a partisan judiciary to silence critics and  dismantle democracy , the Geneva-based International Co
mmission of Jurists (ICJ) said in a report on Wednesday.      Old laws are being dusted off and new laws are being created,  Kingsley Abbot, the group s
southeast Asia adviser, told reporters.  And all of these have led to a list of seemingly politicized investigations.   In recent months, the government
of Prime Minister Hun Sen, who has ruled Cambodia for more than 30 years, has revoked the licenses of about 15 independent radio stations. The Cambodia
Daily, an English-language newspaper, was also forced to shut in September following government allegations of non-payment of millions of dollars in tax
es.  Hun Sen has the backing of Beijing, which says it supports Cambodia s right to maintain its national security. '
```

```
[5]: data.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
```

PREPROCESSING:

REMOVING NULL VALUES:

(from the information above there is no null values so nothing is removed)

```
[6]: #preprocessing
     #drop null values
     data=data.dropna(axis=0)
```

```
[7]: len(data)
```

```
[7]: 44898
```

CONVERTING ALL STRINGS TO LOWERCASE:

```
[8]: #converting all strings to lowercase
     data['clean_news']=data['text'].str.lower()
     data['clean_news']
```

```
[8]: 0        bangkok (reuters) - rights groups on wednesday...
     1        on friday s broadcast of hbo s  real time,  fo...
     2        21st century wire says regardless of what one ...
     3        (reuters) - u.s. president-elect donald trump ...
     4        the  hard working  first family, in need of an...
                                ...
     44893    21st century wire says does the american ideal...
     44894    barinas, venezuela (reuters) - tirelessly trav...
     44895    phnom penh (reuters) - cambodian prime ministe...
     44896    geneva (reuters) - the united states wants to ...
     44897    beijing (reuters) - u.s. president donald trum...
     Name: clean_news, Length: 44898, dtype: object
```

REMOVING SPECIAL CHARACTERS , EXTRA SPACES AND ESCAPE CHARACTERS

[9]:
```
#removing special characters , extra spaces and escape characters
data['clean_news']=data['clean_news'].str.replace('[^A-Za-z0-9\s]','')
data['clean_news']=data['clean_news'].str.replace('[\n]','')
data['clean_news']=data['clean_news'].str.replace('[\s+]',' ')
data['clean_news']
```

[9]:
```
0        bangkok (reuters) - rights groups on wednesday...
1        on friday s broadcast of hbo s  real time,  fo...
2        21st century wire says regardless of what one ...
3        (reuters) - u.s. president-elect donald trump ...
4        the  hard working  first family, in need of an...
                              ...
44893    21st century wire says does the american ideal...
44894    barinas, venezuela (reuters) - tirelessly trav...
44895    phnom penh (reuters) - cambodian prime ministe...
44896    geneva (reuters) - the united states wants to ...
44897    beijing (reuters) - u.s. president donald trum...
Name: clean_news, Length: 44898, dtype: object
```

REMOVING STOP WORDS

[10]:
```
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\rsriv\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

[10]:
```
True
```

[11]:
```
#remove stop words
from nltk.corpus import stopwords
stop=stopwords.words('english')
data['clean_news']=data['clean_news'].apply(lambda x: " ".join([word for word in x.split() if word not in stop]))
data.head()
```

[11]:

	title	text	subject	date	Target	clean_news
0	Rights groups urge EU, Japan to consider halt ...	BANGKOK (Reuters) - Rights groups on Wednesday...	worldnews	October 18, 2017	True	bangkok (reuters) - rights groups wednesday ur...
1	WATCH: IRRELEVANT DEM POLITICAL ANALYST James ...	On Friday s broadcast of HBO s Real Time, fo...	left-news	Oct 21, 2017	Fake	friday broadcast hbo real time, former clinton...
2	Trump Asks O'Reilly, 'Do you think our country...	21st Century Wire says Regardless of what one ...	US_News	February 6, 2017	Fake	21st century wire says regardless one thinks d...
3	Factbox: Trump fills top jobs for his administ...	(Reuters) - U.S. President-elect Donald Trump ...	politicsNews	December 5, 2016	True	(reuters) - u.s. president-elect donald trump ...
4	ONE LAST TIME ON OUR DIME: Mooch and Barack Ar...	The hard working First Family, in need of an...	politics	Aug 6, 2016	Fake	hard working first family, need another taxpay...

TOKENIZATION:

[13]:
```
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\rsriv\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

[13]:
```
True
```

[14]:
```
#Tokenization
from nltk.tokenize import word_tokenize
data['tokenized_news'] = data['clean_news'].apply(lambda x: word_tokenize(x))
data.head()
```

[14]:

	title	text	subject	date	Target	clean_news	tokenized_news
0	Rights groups urge EU, Japan to consider halt ...	BANGKOK (Reuters) - Rights groups on Wednesday...	worldnews	October 18, 2017	True	bangkok (reuters) - rights groups wednesday ur...	[bangkok, (, reuters,), -, rights, groups, we...
1	WATCH: IRRELEVANT DEM POLITICAL ANALYST James ...	On Friday s broadcast of HBO s Real Time, fo...	left-news	Oct 21, 2017	Fake	friday broadcast hbo real time, former clinton...	[friday, broadcast, hbo, real, time, ,, former...
2	Trump Asks O'Reilly, 'Do you think our country...	21st Century Wire says Regardless of what one ...	US_News	February 6, 2017	Fake	21st century wire says regardless one thinks d...	[21st, century, wire, says, regardless, one, t...
3	Factbox: Trump fills top jobs for his administ...	(Reuters) - U.S. President-elect Donald Trump ...	politicsNews	December 5, 2016	True	(reuters) - u.s. president-elect donald trump ...	[(, reuters,), -, u.s., president-elect, dona...
4	ONE LAST TIME ON OUR DIME: Mooch and Barack Ar...	The hard working First Family, in need of an...	politics	Aug 6, 2016	Fake	hard working first family, need another taxpay...	[hard, working, first, family, ,, need, anothe...

LEMMATIZATION:

[15]:
```
nltk.download('wordnet')
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\rsriv\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

[15]:
```
True
```

[16]:
```
#Lemmatization
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
def lemmatize_text(tokens, lemmatizer):
    return [lemmatizer.lemmatize(token) for token in tokens]
data['lemmatized_news'] = data['tokenized_news'].apply(lambda x: lemmatize_text(x, lemmatizer))
data.head()
```

[16]:

	title	text	subject	date	Target	clean_news	tokenized_news	lemmatized_news
0	Rights groups urge EU, Japan to consider halt ...	BANGKOK (Reuters) - Rights groups on Wednesday...	worldnews	October 18, 2017	True	bangkok (reuters) - rights groups wednesday ur...	[bangkok, (, reuters,), -, rights, groups, we...	[bangkok, (, reuters,), -, right, group, wedn...
1	WATCH: IRRELEVANT DEM POLITICAL ANALYST James ...	On Friday s broadcast of HBO s Real Time, fo...	left-news	Oct 21, 2017	Fake	friday broadcast hbo real time, former clinton...	[friday, broadcast, hbo, real, time, ,, former...	[friday, broadcast, hbo, real, time, ,, former...
2	Trump Asks O'Reilly, 'Do you think our country...	21st Century Wire says Regardless of what one ...	US_News	February 6, 2017	Fake	21st century wire says regardless one thinks d...	[21st, century, wire, says, regardless, one, t...	[21st, century, wire, say, regardless, one, th...
3	Factbox: Trump fills top jobs for his administ...	(Reuters) - U.S. President-elect Donald Trump ...	politicsNews	December 5, 2016	True	(reuters) - u.s. president-elect donald trump ...	[(, reuters,), -, u.s., president-elect, dona...	[(, reuters,), -, u.s., president-elect, dona...
4	ONE LAST TIME ON OUR DIME: Mooch and Barack Ar...	The hard working First Family, in need of an...	politics	Aug 6, 2016	Fake	hard working first family, need another taxpay...	[hard, working, first, family, ,, need, anothe...	[hard, working, first, family, ,, need, anothe...

CREATE SENTENCES TO GET CLEAN TEXT AS INPUT FOR VECTORS

[17]:
```
def return_sentences(tokenized_news):
    return " ".join([word for word in tokenized_news])
```

[18]:
```
data['clean_text'] = data['lemmatized_news'].apply(lambda x : return_sentences(x))
data.head()
```

[18]:

	title	text	subject	date	Target	clean_news	tokenized_news	lemmatized_news	clean_text
0	Rights groups urge EU, Japan to consider halt ...	BANGKOK (Reuters) - Rights groups on Wednesday...	worldnews	October 18, 2017	True	bangkok (reuters) - rights groups wednesday ur...	[bangkok, (, reuters,), -, rights, groups, we...	[bangkok, (, reuters,), -, right, group, wedn...	bangkok (reuters) - right group wednesday ur...
1	WATCH: IRRELEVANT DEM POLITICAL ANALYST James ...	On Friday s broadcast of HBO s Real Time, fo...	left-news	Oct 21, 2017	Fake	friday broadcast hbo real time, former clinton...	[friday, broadcast, hbo, real, time, ,, former...	[friday, broadcast, hbo, real, time, ,, former...	friday broadcast hbo real time , former clinto...
2	Trump Asks O'Reilly, 'Do you think our country...	21st Century Wire says Regardless of what one ...	US_News	February 6, 2017	Fake	21st century wire says regardless one thinks d...	[21st, century, wire, says, regardless, one, t...	[21st, century, wire, say, regardless, one, th...	21st century wire say regardless one think don...
3	Factbox: Trump fills top jobs for his administ...	(Reuters) - U.S. President-elect Donald Trump ...	politicsNews	December 5, 2016	True	(reuters) - u.s. president-elect donald trump ...	[(, reuters,), -, u.s., president-elect, dona...	[(, reuters,), -, u.s., president-elect, dona...	(reuters) - u.s. president-elect donald trum...
4	ONE LAST TIME ON OUR DIME: Mooch and Barack Ar...	The hard working First Family, in need of an...	politics	Aug 6, 2016	Fake	hard working first family, need another taxpay...	[hard, working, first, family, ,, need, anothe...	[hard, working, first, family, ,, need, anothe...	hard working first family , need another taxpa...

PREPARE DATA FOR THE MODEL. CONVERT LABEL IN TO BINARY

[20]:
```
data['Target'] = [1 if x == 'Fake' else 0 for x in data['Target']]
data.head()
```

[20]:

	title	text	subject	date	Target	clean_news	tokenized_news	lemmatized_news	clean_text
0	Rights groups urge EU, Japan to consider halt ...	BANGKOK (Reuters) - Rights groups on Wednesday...	worldnews	October 18, 2017	0	bangkok (reuters) - rights groups wednesday ur...	[bangkok, (, reuters,), -, rights, groups, we...	[bangkok, (, reuters,), -, right, group, wedn...	bangkok (reuters) - right group wednesday ur...
1	WATCH: IRRELEVANT DEM POLITICAL ANALYST James ...	On Friday s broadcast of HBO s Real Time, fo...	left-news	Oct 21, 2017	1	friday broadcast hbo real time, former clinton...	[friday, broadcast, hbo, real, time, ,, former...	[friday, broadcast, hbo, real, time, ,, former...	friday broadcast hbo real time , former clinto...
2	Trump Asks O'Reilly, 'Do you think our country...	21st Century Wire says Regardless of what one ...	US_News	February 6, 2017	1	21st century wire says regardless one thinks d...	[21st, century, wire, says, regardless, one, t...	[21st, century, wire, say, regardless, one, th...	21st century wire say regardless one think don...
3	Factbox: Trump fills top jobs for his administ...	(Reuters) - U.S. President-elect Donald Trump ...	politicsNews	December 5, 2016	0	(reuters) - u.s. president-elect donald trump ...	[(, reuters,), -, u.s., president-elect, dona...	[(, reuters,), -, u.s., president-elect, dona...	(reuters) - u.s. president-elect donald trum...
4	ONE LAST TIME ON OUR DIME: Mooch and Barack Ar...	The hard working First Family, in need of an...	politics	Aug 6, 2016	1	hard working first family, need another taxpay...	[hard, working, first, family, ,, need, anothe...	[hard, working, first, family, ,, need, anothe...	hard working first family , need another taxpa...

SPLIT THE DATASET

[21]:
```
from sklearn.model_selection import train_test_split
```

[22]:
```
X_train, X_test, y_train, y_test = train_test_split(data['clean_text'], data['Target'], test_size=0.2, random_state=5)

print(X_train.shape)
print(X_test.shape)
```

```
(35918,)
(8980,)
```

FEATURE EXTRACTION

(words to vectors)

Count vectorizer which considers the frequency of occurrence of a word across the corpus.

[23]:
```
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer

# Assuming 'Lemmatized_news' is a column in your DataFrame 'data'
count_vectorizer = CountVectorizer()
X = count_vectorizer.fit_transform(data['clean_text'])

# Get feature names
feature_names_count = count_vectorizer.get_feature_names_out()

print("CountVectorizer feature names:", feature_names_count)
```

```
CountVectorizer feature names: ['00' '000' '0000' ... 'zzzzzzzz' 'zzzzzzzzzzzz' 'émigré']
```

TF-IDF : Term Frequency - Inverse Document Frequency

The term frequency is the number of times a term occurs in a document. Inverse document frequency is an inverse function of the number of documents in which that a given word occurs.

The product of these two terms gives tf-idf weight for a word in the corpus.

[26]:
```
tfidf_vectorizer = TfidfVectorizer()
X_tfidf = tfidf_vectorizer.fit_transform(data['clean_text'])
feature_names_tfidf = tfidf_vectorizer.get_feature_names_out()
print("TfidfVectorizer feature names:", feature_names_tfidf)
```

```
TfidfVectorizer feature names: ['00' '000' '0000' ... 'zzzzzzzz' 'zzzzzzzzzzzz' 'émigré']
```

[27]:
```
tfidf = TfidfVectorizer()
tfidf_train = tfidf.fit_transform(X_train)
tfidf_test = tfidf.transform(X_test)

print(tfidf_train.shape)
print(tfidf_test.shape)
```

```
(35918, 106465)
(8980, 106465)
```

CONCLUSION:

In the preprocessing stage of the fake news detection dataset, a series of crucial steps were performed using NLP techniques. Punctuation and stopwords were removed to clean the text data, and tokenization was applied, ensuring that each document was divided into individual words. Further refinement was achieved through lemmatization, standardizing words to their base form. Feature vectors were then generated, capturing the essence of the text data. Additionally, it is important to highlight that using sparse representations for text data, like TF-IDF, can significantly reduce memory usage and improve model performance.