AI BASED SOLUTION FOR FLAGGING OF FALSE INFORMATION ON ONLINE PLATFORMS

In [1]:

```
import pandas as pd
import numpy as np
data = pd.read_csv("E:/file2/Desktop/new_newsdesk.csv")
```

In [2]:

```
data = data.dropna(how = 'any', axis = 0)
```

In [3]:

```
data.isnull().sum()
```

Out[3]:

```
label    0
text     0
dtype: int64
```

In [4]:

```
data.label.value_counts()
```

Out[4]:

```
FAKE    1871
REAL    1850
Name: label, dtype: int64
```

In [5]:

```
from nltk.stem.porter import PorterStemmer
import re
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.preprocessing import LabelEncoder
from nltk.stem import WordNetLemmatizer
```

In [9]:

```
from nltk.corpus import stopwords
import nltk
```

In [10]:

```
stemming = PorterStemmer()
lemmatizer = WordNetLemmatizer()
```

In [11]:

```
from sklearn.model_selection import train_test_split
```

In [12]:

```
X=data[['text']]
Y=data['label']
```

In [116]:

```
X
```

Out[116]:

	text
0	Payal has accused filmmaker Anurag Kashyap of ...
1	A four-minute-long video of a woman criticisin...
2	Republic Poll, a fake Twitter account imitatin...
3	Delhi teen finds place on UN green list, turns...
4	Delhi: A high-level meeting underway at reside...
...	...
3724	19:17 (IST) Sep 20\n\nThe second round of coun...
3725	19:17 (IST) Sep 20\n\nThe second round of coun...
3726	The Bengaluru City Police’s official Twitter h...
3727	Sep 20, 2020, 08:00AM IST\n\nSource: TOI.in\n\...
3728	Read Also\n\nRead Also\n\nAdvocate Ishkaran Bh...

3721 rows × 1 columns

In [115]:

```
p=data['text']
print(p)
```

```
0       Payal has accused filmmaker Anurag Kashyap of ...
1       A four-minute-long video of a woman criticisin...
2       Republic Poll, a fake Twitter account imitatin...
3       Delhi teen finds place on UN green list, turns...
4       Delhi: A high-level meeting underway at reside...
                              ...
3724    19:17 (IST) Sep 20\n\nThe second round of coun...
3725    19:17 (IST) Sep 20\n\nThe second round of coun...
3726    The Bengaluru City Police’s official Twitter h...
3727    Sep 20, 2020, 08:00AM IST\n\nSource: TOI.in\n\...
3728    Read Also\n\nRead Also\n\nAdvocate Ishkaran Bh...
Name: text, Length: 3721, dtype: object
```

In [13]:

```
x_train,x_test,y_train,y_test = train_test_split(X,Y,test_size=0.2,random_state=42)
```

In [14]:

```
print('x_train:',x_train.shape)
print('y_train:',y_train.shape)
print('x_test:',x_test.shape)
print('y_test:',y_test.shape)
```

```
x_train: (2976, 1)
y_train: (2976,)
x_test: (745, 1)
y_test: (745,)
```

In [15]:

```
X_train = x_train
```

In [16]:

```
x_train.head()
```

Out[16]:

	text
209	Several mainstream news outlets such as the Ti...
3374	NEW DELHI: The Drugs Controller General of Ind...
3540	A screenshot purporting to be a newspaper clip...
2472	A viral video of a woman with infant traveling...
2510	Read Also\n\nBe it winning hearts or winning t...

In [17]:

```
X_test = x_test
```

In [18]:

```
y_train.head()
```

Out[18]:

```
209     FAKE
3374    REAL
3540    FAKE
2472    FAKE
2510    REAL
Name: label, dtype: object
```

In [19]:

```
X_train.head()
```

Out[19]:

	text
209	Several mainstream news outlets such as the Ti...
3374	NEW DELHI: The Drugs Controller General of Ind...
3540	A screenshot purporting to be a newspaper clip...
2472	A viral video of a woman with infant traveling...
2510	Read Also\n\nBe it winning hearts or winning t...

In [20]:

```
X_test.head(10)
```

Out[20]:

	text
908	NEW DELHI: A final decision on Pakistan's stat...
3454	NEW DELHI: Seven of the top 10 most valued dom...
1790	Kareena Kapoor Khan, who is all set to ring in...
1167	A photo purporting to show a television news g...
1605	A disturbing video of a woman being flogged by...
184	A graphic photo of a human skeleton found insi...
2960	Delhi: A high-level meeting underway at reside...
1067	Social media has been rife with reports of the...
2348	A disturbing CCTV footage showing a Tamil Nadu...
3417	A disturbing video of a mentally ill woman hec...

In [ ]:

```

```

In [21]:

```
y_test
```

Out[21]:

```
908     REAL
3454    REAL
1790    REAL
1167    FAKE
1605    FAKE
        ...
1239    FAKE
2409    FAKE
1958    FAKE
2680    FAKE
955     FAKE
Name: label, Length: 745, dtype: object
```

Data Preprocessing

In [22]:

```
def preprocess(pro):
    process = re.sub('[^a-zA-Z]'," ",pro)
    lowe = process.lower()
    tokens = lowe.split()

    stop = [lemmatizer.lemmatize(i) for i in tokens if i not in stopwords.words('English')]
    lemmas =pd.Series([ " ".join(stop),len(stop)])
    return lemmas
```

In [23]:

```
px_train = X_train['text'].apply(preprocess)
```

In [109]:

```
px_train.head()
```

Out[109]:

	clean_text	text_length
209	several mainstream news outlet time india hind...	396
3374	new delhi drug controller general india approv...	257
3540	screenshot purporting newspaper clipping claim...	289
2472	viral video woman infant traveling precariousl...	355
2510	read also winning heart winning trophy easy bi...	123

In [110]:

```
type(px_train)
```

Out[110]:

```
pandas.core.frame.DataFrame
```

Test data preprocessing

In [26]:

```
px_test = X_test['text'].apply(preprocess)
```

In [27]:

```
px_test.head()
```

Out[27]:

	0	1
908	new delhi final decision pakistan status finan...	150
3454	new delhi seven top valued domestic company sa...	187
1790	kareena kapoor khan set ring birthday tomorrow...	105
1167	photo purporting show television news graphic ...	170
1605	disturbing video woman flogged law husband all...	180

In [28]:

```
px_test.columns = ['clean_text','text_length']
px_test.head()
```

Out[28]:

	clean_text	text_length
908	new delhi final decision pakistan status finan...	150
3454	new delhi seven top valued domestic company sa...	187
1790	kareena kapoor khan set ring birthday tomorrow...	105
1167	photo purporting show television news graphic ...	170
1605	disturbing video woman flogged law husband all...	180

In [29]:

```
px_train.columns = ['clean_text','text_length']
px_train.head()
```

Out[29]:

	clean_text	text_length
209	several mainstream news outlet time india hind...	396
3374	new delhi drug controller general india approv...	257
3540	screenshot purporting newspaper clipping claim...	289
2472	viral video woman infant traveling precariousl...	355
2510	read also winning heart winning trophy easy bi...	123

In [30]:

```
X_train = pd.concat([X_train,px_train],axis=1)
X_train.head()
```

Out[30]:

	text	clean_text	text_length
209	Several mainstream news outlets such as the Ti...	several mainstream news outlet time india hind...	396
3374	NEW DELHI: The Drugs Controller General of Ind...	new delhi drug controller general india approv...	257
3540	A screenshot purporting to be a newspaper clip...	screenshot purporting newspaper clipping claim...	289
2472	A viral video of a woman with infant traveling...	viral video woman infant traveling precariousl...	355
2510	Read Also\n\nBe it winning hearts or winning t...	read also winning heart winning trophy easy bi...	123

In [31]:

```
X_test = pd.concat([X_test,px_test],axis=1)
```

In [32]:

```
X_test.head()
```

Out[32]:

	text	clean_text	text_length
908	NEW DELHI: A final decision on Pakistan's stat...	new delhi final decision pakistan status finan...	150
3454	NEW DELHI: Seven of the top 10 most valued dom...	new delhi seven top valued domestic company sa...	187
1790	Kareena Kapoor Khan, who is all set to ring in...	kareena kapoor khan set ring birthday tomorrow...	105
1167	A photo purporting to show a television news g...	photo purporting show television news graphic ...	170
1605	A disturbing video of a woman being flogged by...	disturbing video woman flogged law husband all...	180

In [33]:

```
from wordcloud import WordCloud
```

In [34]:

```
y_train
```

Out[34]:

```
209     FAKE
3374    REAL
3540    FAKE
2472    FAKE
2510    REAL
        ...
1133    REAL
1297    REAL
863     FAKE
3515    REAL
3182    FAKE
Name: label, Length: 2976, dtype: object
```

In [35]:

```
y_test
```

Out[35]:

```
908     REAL
3454    REAL
1790    REAL
1167    FAKE
1605    FAKE
        ...
1239    FAKE
2409    FAKE
1958    FAKE
2680    FAKE
955     FAKE
Name: label, Length: 745, dtype: object
```

In [36]:

```
real_n = X_train.loc[y_train=='REAL', :]
real_n.head()
```

Out[36]:

	text	clean_text	text_length
3374	NEW DELHI: The Drugs Controller General of Ind...	new delhi drug controller general india approv...	257
2510	Read Also\n\nBe it winning hearts or winning t...	read also winning heart winning trophy easy bi...	123
599	WASHINGTON: Enter Journey's Crossing Church in...	washington enter journey crossing church washi...	215
1707	NEW DELHI: The finance ministry on Saturday in...	new delhi finance ministry saturday informed l...	266
3676	PANAJI: The second phase of reviving the cocon...	panaji second phase reviving coconut tree line...	98

In [37]:

```
words = ' '.join(real_n['clean_text'])
clean_word = " ".join([word for word in words.split()])
```

In [38]:

```
real_word = WordCloud(stopwords=stopwords.words("english"),
                      background_color='black',
                      width=1600,
                      height=800).generate(clean_word)
```

In [39]:

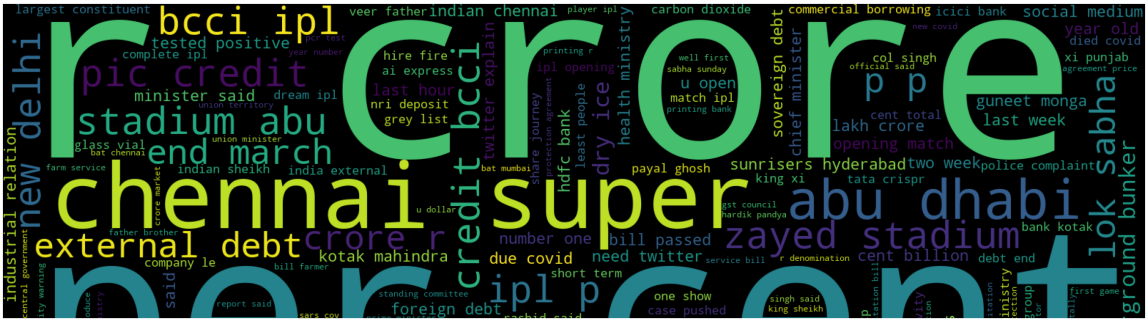
```
plt.figure(1,figsize=(30,20))
plt.imshow(real_word)
plt.axis('off')
plt.show()
```

```
<IPython.core.display.Javascript object>

<IPython.core.display.Javascript object>

<IPython.core.display.Javascript object>

<IPython.core.display.Javascript object>
```



In [40]:

```
fake_n = X_train.loc[y_train=='FAKE', :]
fake_n.head()
```

Out[40]:

	text	clean_text	text_length
209	Several mainstream news outlets such as the Ti...	several mainstream news outlet time india hind...	396
3540	A screenshot purporting to be a newspaper clip...	screenshot purporting newspaper clipping claim...	289
2472	A viral video of a woman with infant traveling...	viral video woman infant traveling precariousl...	355
2704	A press release detailing restrictions imposed...	press release detailing restriction imposed mo...	233
1224	A video of a customer losing his cool at a bak...	video customer losing cool bakery manager kara...	238

In [41]:

```
words_f = ' '.join(fake_n['clean_text'])
clean_word_f = " ".join([word for word in words_f.split()])
```

In [42]:

```
real_word_f = WordCloud(stopwords=stopwords.words("english"),
                        background_color='black',
                        width=1600,
                        height=800).generate(clean_word_f)
```

In [43]:

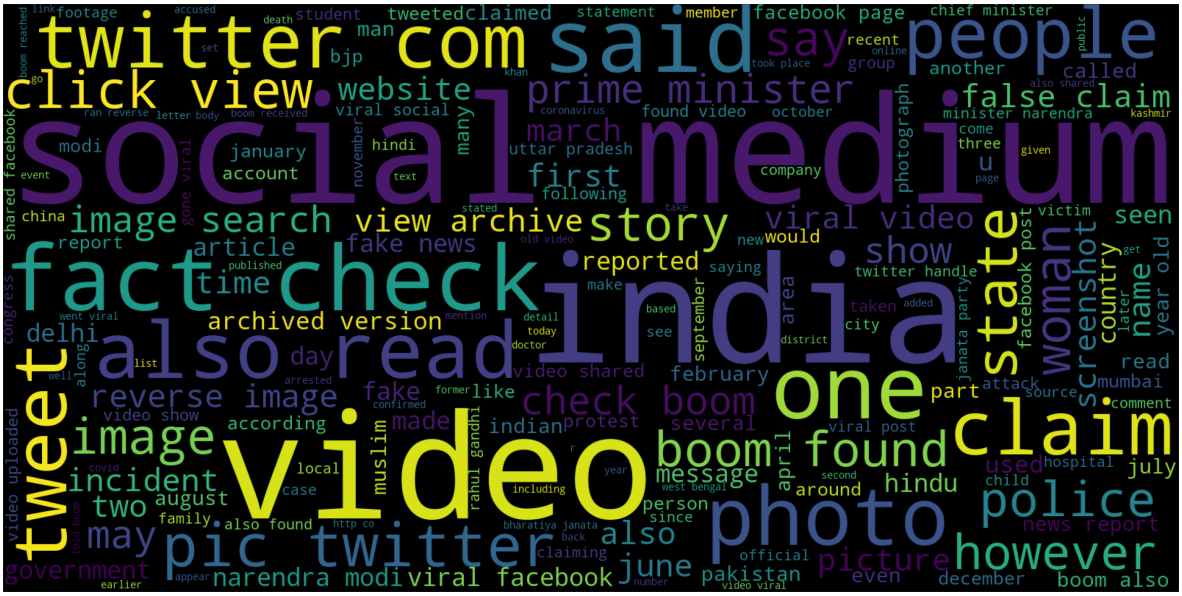
```
plt.figure(1,figsize=(30,20))
plt.imshow(real_word_f)
plt.axis('off')
plt.show()
```

```
<IPython.core.display.Javascript object>


<IPython.core.display.Javascript object>


<IPython.core.display.Javascript object>


<IPython.core.display.Javascript object>
```



Tfidf Vectorizer

In [44]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [45]:

```
tf_vector = TfidfVectorizer()
```

In [46]:

```
X_train_t = tf_vector.fit_transform(X_train['clean_text'])
```

In [47]:

```
(X_train_t)
```

Out[47]:

```
<2976x28314 sparse matrix of type '<class 'numpy.float64'>'
        with 442232 stored elements in Compressed Sparse Row format>
```

In [48]:

```
print('unique words:',len(tf_vector.vocabulary_))
print('Shape of input data:',X_train_t.shape)
```

```
unique words: 28314
Shape of input data: (2976, 28314)
```

Test data

In [49]:

```
X_test_tf = tf_vector.transform(X_test['clean_text'])
```

In [50]:

```
X_test_tf
```

Out[50]:

```
<745x28314 sparse matrix of type '<class 'numpy.float64'>'
        with 107305 stored elements in Compressed Sparse Row format>
```

Label Encoding

In [51]:

```
label = LabelEncoder()
```

In [52]:

```
y_train = label.fit_transform(y_train)
```

In [53]:

```
y_train
```

Out[53]:

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

In [54]:

```
Y_test = label.transform(y_test)
```

In [55]:

```
Y_test
```

Out[55]:

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

Logistic Regression Model

In [56]:

```
from sklearn.linear_model import LogisticRegression
```

In [57]:

```
models = LogisticRegression()
```

In [58]:

```
models.fit(X_train_t,y_train)
```

Out[58]:

```
LogisticRegression()
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [59]:

```
from sklearn.metrics import accuracy_score
```

In [103]:

```
l_train_score = models.predict(X_train_t)
l_train_accuracy = accuracy_score(l_train_score,y_train)
```

In [111]:

```
print('train_accuracy:',l_train_accuracy)
```

```
train_accuracy: 0.998991935483871
```

In [105]:

```
l_test_score = models.predict(X_test_tf)
```

In [106]:

```
l_test_accuracy = accuracy_score(test_score,Y_test)
```

In [107]:

```
print('test_acccuracy:',l_test_accuracy)
```

```
test_acccuracy: 0.9919463087248322
```

In [108]:

```
cmx_1=confusion_matrix(Y_test,l_test_score)
print("\nNo. of test samples : ",len(X_test))
print("\n Confustion Matrix : \n",cmx_2)
print("\nPerfomance measures are: \n",classification_report(Y_test, l_test_score))
```

```
No. of test samples :  745

 Confustion Matrix :
 [[352   9]
 [ 35 349]]

Perfomance measures are:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       361
           1       1.00      0.99      0.99       384

    accuracy                           0.99       745
   macro avg       0.99      0.99      0.99       745
weighted avg       0.99      0.99      0.99       745
```

In [ ]:

```

```

In [65]:

```
news=X_train_t[1]
```

In [66]:

```
prediction = models.predict(news)
print(prediction)

if (prediction[0]==0):
    print('The news is fake')
else:
    print('The news is real')
```

```
[1]
The news is real
```

In [67]:

```
from sklearn import metrics
```

In [68]:

```
confusion = metrics.confusion_matrix(Y_test, test_score)
```

In [69]:

```
confusion
```

Out[69]:

```
array([[360,   1],
       [  5, 379]], dtype=int64)
```

SVM

In [70]:

```
from sklearn.svm import SVC
```

In [71]:

```
support = svm.SVC()
```

```
<IPython.core.display.Javascript object>
```

In [72]:

```
support
```

Out[72]:

```
SVC()
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [73]:

```
support.fit(X_train_t,y_train)
```

Out[73]:

```
SVC()
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [74]:

```
train_score_1 = support.predict(X_train_t)
train_accuracy_1 = accuracy_score(train_score_1,y_train)
```

In [75]:

```
print('train_accuracy:',train_accuracy_1)
```

```
train_accuracy: 1.0
```

In [76]:

```
test_score_1 = support.predict(X_test_tf)
```

In [77]:

```
test_accuracy_1 = accuracy_score(test_score_1,Y_test)
```

In [78]:

```
print('test_acccuracy:',test_accuracy_1)
```

```
test_acccuracy: 0.9892617449664429
```

In [79]:

```
news_1=X_train_t[1]
```

In [80]:

```
prediction_1 = support.predict(news_1)
print(prediction_1)

if (prediction_1[0]==0):
    print('The news is fake')
else:
    print('The news is real')
```

```
[1]
The news is real
```

In [81]:

```
from sklearn.metrics import classification_report, confusion_matrix
```

In [82]:

```
confusion = metrics.confusion_matrix(Y_test, test_score_1)
```

In [83]:

```
cmx=confusion_matrix(Y_test,test_score)
print("\nNo. of test samples : ",len(X_test))
print("\n Confustion Matrix : \n",cmx)
print("\nPerfomance measures are: \n",classification_report(Y_test, test_score))
```

```
No. of test samples :  745

 Confustion Matrix :
 [[360   1]
 [  5 379]]

Perfomance measures are:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       361
           1       1.00      0.99      0.99       384

    accuracy                           0.99       745
   macro avg       0.99      0.99      0.99       745
weighted avg       0.99      0.99      0.99       745
```

KNN

In [85]:

```
from sklearn.neighbors import KNeighborsClassifier
```

In [94]:

```
knn_model = KNeighborsClassifier(n_neighbors=5)
```

In [95]:

```
knn_model.fit(X_train_t,y_train)
```

Out[95]:

```
KNeighborsClassifier()
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [96]:

```
knn_1_train_score = knn_model.predict(X_train_t)
knn_train_accuracy = accuracy_score(knn_1_train_score,y_train)
```

In [112]:

```
print('train_accuracy:',knn_train_accuracy)
```

```
train_accuracy: 0.9684139784946236
```

In [98]:

```
knn_test_score = knn_model.predict(X_test_tf)
```

In [99]:

```
knn_test_accuracy = accuracy_score(knn_test_score,Y_test)
```

In [100]:

```
print('test_acccuracy:',knn_test_accuracy)
```

```
test_acccuracy: 0.9409395973154362
```

In [102]:

```
cmx_2=confusion_matrix(Y_test,knn_test_score)
print("\nNo. of test samples : ",len(X_test))
print("\n Confustion Matrix : \n",cmx_2)
print("\nPerfomance measures are: \n",classification_report(Y_test, knn_test_score))
```

```
No. of test samples :  745

 Confustion Matrix :
 [[352   9]
 [ 35 349]]

Perfomance measures are:
              precision    recall  f1-score   support

           0       0.91      0.98      0.94       361
           1       0.97      0.91      0.94       384

    accuracy                           0.94       745
   macro avg       0.94      0.94      0.94       745
weighted avg       0.94      0.94      0.94       745
```

In [ ]:

```

```

In [ ]:

```

```