FakeNewsPrediction dataset

April 19, 2025

[1]:
```
import pandas as pd
```

[2]:
```
import numpy as np
```

[3]:
```
import matplotlib.pyplot as plt
```

[4]:
```
import seaborn as sns
```

[5]:
```
from sklearn.model_selection import train_test_split
```

[7]:
```
from sklearn.linear_model import LogisticRegression
```

[8]:
```
import re
```

[9]:
```
from nltk.corpus import stopwords
```

[10]:
```
from nltk.stem.porter import PorterStemmer
```

[12]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
```

[13]:
```
from sklearn.metrics import accuracy_score,confusion_matrix
```

[14]:
```
import nltk
```

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

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\indhu\AppData\Roaming\nltk_data…
[nltk_data]   Unzipping corpora\stopwords.zip.
```

[15]:
```
True
```

Printing the stopwords in English

[16]:
```
print(stopwords.words('english'))
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're",
"you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he',
'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's",
'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what',
```

```
'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is',
'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having',
'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or',
'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about',
'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above',
'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under',
'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why',
'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some',
'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very',
's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now',
'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn',
"couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn',
"hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't",
'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn',
"shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn',
"wouldn't"]
```

Data collection and Data Preprocessing

[17]:
```
news_dataset=pd.read_csv('fakenews train.csv')
```

[19]:
```
news_dataset.head()
```

[19]:

	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

[20]:
```
news_dataset.shape
```

[20]:
```
(20800, 5)
```

[21]:
```
news_dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20800 entries, 0 to 20799
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   id      20800 non-null  int64
```

```
 1   title   20242 non-null  object
 2   author  18843 non-null  object
 3   text    20761 non-null  object
 4   label   20800 non-null  int64
dtypes: int64(2), object(3)
memory usage: 812.6+ KB
```

[22]:
```
news_dataset.isnull().sum()
```

[22]:
```
id           0
title      558
author    1957
text        39
label        0
dtype: int64
```

Hence we have large dataset we can replace missing values with empty values

[23]:
```
news_dataset=news_dataset.fillna('')
```

Merging the author name and news title

[24]:
```
news_dataset['content']=news_dataset['title']+''+news_dataset['author']
```

[25]:
```
news_dataset.head()
```

[25]:
```
   id                                              title              author  \
0   0  House Dem Aide: We Didn’t Even See Comey’s Let…       Darrell Lucus
1   1  FLYNN: Hillary Clinton, Big Woman on Campus - …     Daniel J. Flynn
2   2                  Why the Truth Might Get You Fired  Consortiumnews.com
3   3  15 Civilians Killed In Single US Airstrike Hav…     Jessica Purkiss
4   4  Iranian woman jailed for fictional unpublished…      Howard Portnoy

                                                text  label  \
0  House Dem Aide: We Didn’t Even See Comey’s Let…      1
1  Ever get the feeling your life circles the rou…      0
2  Why the Truth Might Get You Fired October 29, …      1
3  Videos 15 Civilians Killed In Single US Airstr…      1
4  Print \nAn Iranian woman has been sentenced to…      1

                                             content
0  House Dem Aide: We Didn’t Even See Comey’s Let…
1  FLYNN: Hillary Clinton, Big Woman on Campus - …
2  Why the Truth Might Get You FiredConsortiumnew…
3  15 Civilians Killed In Single US Airstrike Hav…
4  Iranian woman jailed for fictional unpublished…
```

[26]:
```
news_dataset['content']
```

[26]:
```
0        House Dem Aide: We Didn’t Even See Comey’s Let…
1        FLYNN: Hillary Clinton, Big Woman on Campus - …
2        Why the Truth Might Get You FiredConsortiumnew…
3        15 Civilians Killed In Single US Airstrike Hav…
4        Iranian woman jailed for fictional unpublished…
                             …
20795    Rapper T.I.: Trump a ’Poster Child For White S…
20796    N.F.L. Playoffs: Schedule, Matchups and Odds -…
20797    Macy’s Is Said to Receive Takeover Approach by…
20798    NATO, Russia To Hold Parallel Exercises In Bal…
20799                What Keeps the F-35 AliveDavid Swanson
Name: content, Length: 20800, dtype: object
```

we use content data and labels to make predictions

[30]:
```
X=news_dataset.drop('label',axis=1)
```

[28]:
```
Y=news_dataset['label']
```

[31]:
```
print(X)
```

```
          id                                              title  \
0          0  House Dem Aide: We Didn’t Even See Comey’s Let…
1          1  FLYNN: Hillary Clinton, Big Woman on Campus - …
2          2                  Why the Truth Might Get You Fired
3          3  15 Civilians Killed In Single US Airstrike Hav…
4          4  Iranian woman jailed for fictional unpublished…
…        …                                                 …
20795  20795  Rapper T.I.: Trump a ’Poster Child For White S…
20796  20796  N.F.L. Playoffs: Schedule, Matchups and Odds -…
20797  20797  Macy’s Is Said to Receive Takeover Approach by…
20798  20798  NATO, Russia To Hold Parallel Exercises In Bal…
20799  20799                          What Keeps the F-35 Alive

                                      author  \
0                                Darrell Lucus
1                              Daniel J. Flynn
2                           Consortiumnews.com
3                              Jessica Purkiss
4                               Howard Portnoy
…                                          …
20795                            Jerome Hudson
20796                          Benjamin Hoffman
20797  Michael J. de la Merced and Rachel Abrams
20798                              Alex Ansary
20799                            David Swanson

                                                 text  \
```

```
0      House Dem Aide: We Didn’t Even See Comey’s Let…
1      Ever get the feeling your life circles the rou…
2      Why the Truth Might Get You Fired October 29, …
3      Videos 15 Civilians Killed In Single US Airstr…
4      Print \nAn Iranian woman has been sentenced to…
…                                                    …
20795  Rapper T. I. unloaded on black celebrities who…
20796  When the Green Bay Packers lost to the Washing…
20797  The Macy’s of today grew from the union of sev…
20798  NATO, Russia To Hold Parallel Exercises In Bal…
20799    David Swanson is an author, activist, journa…

                                                  content
0      House Dem Aide: We Didn’t Even See Comey’s Let…
1      FLYNN: Hillary Clinton, Big Woman on Campus - …
2      Why the Truth Might Get You FiredConsortiumnew…
3      15 Civilians Killed In Single US Airstrike Hav…
4      Iranian woman jailed for fictional unpublished…
…                                                    …
20795  Rapper T.I.: Trump a ’Poster Child For White S…
20796  N.F.L. Playoffs: Schedule, Matchups and Odds -…
20797  Macy’s Is Said to Receive Takeover Approach by…
20798  NATO, Russia To Hold Parallel Exercises In Bal…
20799             What Keeps the F-35 AliveDavid Swanson

[20800 rows x 5 columns]
```

Stemming Procedure : stemming is the process of reducing a word to its root word

[32]:
```
port_stem=PorterStemmer()
```

[33]:
```
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
```

[34]:
```
news_dataset['content']=news_dataset['content'].apply(stemming)
```

[35]:
```
print(news_dataset['content'])
```

```
0        hous dem aid even see comey letter jason chaff…
1        flynn hillari clinton big woman campu breitbar…
2                   truth might get firedconsortiumnew com
3        civilian kill singl us airstrik identifiedjess…
4        iranian woman jail fiction unpublish stori wom…
```

```
                                  …
20795    rapper trump poster child white supremaci jero…
20796    n f l playoff schedul matchup odd new york tim…
20797    maci said receiv takeov approach hudson bay ne…
20798    nato russia hold parallel exercis balkansalex …
20799                              keep f alivedavid swanson
Name: content, Length: 20800, dtype: object
```

Seperating the data and label

[36]:
```
X=news_dataset['content'].values
```

[37]:
```
print(X)
```

```
['hous dem aid even see comey letter jason chaffetz tweet itdarrel lucu'
 'flynn hillari clinton big woman campu breitbartdaniel j flynn'
 'truth might get firedconsortiumnew com' …
 'maci said receiv takeov approach hudson bay new york timesmichael j de la merc
rachel abram'
 'nato russia hold parallel exercis balkansalex ansari'
 'keep f alivedavid swanson']
```

[38]:
```
Y=news_dataset['label'].values
```

[39]:
```
print(Y)
```

```
[1 0 1 … 0 1 1]
```

[40]:
```
Y.shape
```

[40]:
```
(20800,)
```

Converting the textual data into numerical data

[41]:
```
vectorizer=TfidfVectorizer()
vectorizer.fit(X)
X=vectorizer.transform(X)
```

[42]:
```
print(X)
```

```
  (0, 21557)    0.2736369479869461
  (0, 18009)    0.2438301027041085
  (0, 11974)    0.34466883664274506
  (0, 11617)    0.2783091851108118
  (0, 10495)    0.311553446057155
  (0, 10381)    0.41343221816522613
  (0, 9475)     0.20871803491508256
  (0, 6798)     0.22134331972572915
  (0, 5178)     0.25645024223907936
  (0, 3933)     0.23592778464338887
  (0, 3291)     0.34851330509336254
```

```
(0, 381)      0.25686395241555227
(1, 23108)    0.2952143706864955
(1, 9174)     0.18812765977413537
(1, 7534)     0.6987204016565229
(1, 3714)     0.18820851327454977
(1, 2957)     0.37436858023248293
(1, 2482)     0.3616637468521842
(1, 1956)     0.2878737833766196
(2, 21453)    0.41331452278016145
(2, 12835)    0.4627692646157023
(2, 8113)     0.3260098284202015
(2, 7391)     0.6469324358467595
(2, 3910)     0.30035267305096314
(3, 21978)    0.23565193182482072
:     :
(20797, 23430)        0.08239795721216821
(20797, 20788)        0.19858848268663878
(20797, 19961)        0.32033450422171095
(20797, 17564)        0.24595228419376997
(20797, 16511)        0.2701082984026285
(20797, 16242)        0.26138762441775215
(20797, 13686)        0.07963596797146637
(20797, 12710)        0.29268298480392924
(20797, 12062)        0.35825796183079206
(20797, 11247)        0.22115741003482425
(20797, 9530) 0.2159705521323658
(20797, 4956) 0.20977279480430666
(20797, 1719) 0.33227287481144197
(20797, 937)  0.30401505825124225
(20797, 54)   0.29434941720236085
(20798, 17456)        0.23135838404922235
(20798, 14719)        0.4545605136834298
(20798, 13532)        0.3274423107890453
(20798, 9311) 0.331161903207525
(20798, 6885) 0.4161602827826835
(20798, 1517) 0.4991866119925705
(20798, 805)  0.31715092612304463
(20799, 19781)        0.5424505449862735
(20799, 10881)        0.4393421136105601
(20799, 529)  0.7160488205788069
```

Spliting the data into training and test data

[43]:
```
train_x,test_x,train_y,test_y=train_test_split(X,Y,test_size=0.
↪2,stratify=Y,random_state=2)
```

Training the Logistic Regression Model

[44]:
```
Logistic=LogisticRegression()
```

[46]:
```
Logistic.fit(train_x,train_y)
```

[46]:
```
LogisticRegression()
```

[47]:
```
train_x_prediction=Logistic.predict(train_x)
```

[48]:
```
train_x_accuracy=accuracy_score(train_x_prediction,train_y)
```

[49]:
```
print(train_x_accuracy)
```

```
0.9825721153846154
```

[50]:
```
test_x_prediction=Logistic.predict(test_x)
```

[51]:
```
test_x_accuracy=accuracy_score(test_x_prediction,test_y)
```

[52]:
```
print(test_x_accuracy)
```

```
0.9668269230769231
```

Predictive Modeling

[54]:
```
X_new= test_x[0]
prediction=Logistic.predict(X_new)
```

[55]:
```
print(prediction)
```

```
[1]
```

[56]:
```
print(test_y[0])
```

```
1
```

[ ]:
```

```