fake-news-detection-revised

November 26, 2024

1 Fake News Detection using ML

[1]:
```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
import re
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
```

- **1.1 1. re (Regular Expressions)Used for text preprocessing, which is a crucial step in natural language processing (NLP).Why?To remove unnecessary characters (e.g., punctuation, numbers, special characters).To clean text data (e.g., remove URLs, hashtags, mentions, or excessive whitespace).**
- **1.2 nltk.corpus.stopwordsThe stopwords are common words (like the, and, is) that usually don't contribute much meaning in NLP tasks.Why?Removing stopwords reduces noise in the dataset and improves the model's performance by focusing on meaningful word**
- **1.3 nltk.stem.PorterStemmer Stemming reduces words to their root forms (e.g., running → run).Why?It helps in reducing the vocabulary size without losing significant meaning.Improves generalization for machine learning models by treating words with the same root as equivalent.**
- sklearn.feature_extraction.text.TfidfVectorizerThe TF-IDF Vectorizer transforms textual data into numerical features by calculating the Term Frequency-Inverse Document Frequency (TF-IDF).Why?Machine learning models require numerical inputs; raw text can't be used directly.TF-IDF assigns weights to words based on their importance, reducing the impact of common but unimportant words.

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

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

[2]:
```
True
```

[3]:
```
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"]
```

[4]:
```
# import the dataset:
df=pd.read_csv("train.csv")
df
```

[4]:

	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  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
…                                                    …    …
20795  Rapper T. I. unloaded on black celebrities who…      0
20796  When the Green Bay Packers lost to the Washing…      0
20797  The Macy’s of today grew from the union of sev…      0
20798  NATO, Russia To Hold Parallel Exercises In Bal…      1
20799    David Swanson is an author, activist, journa…      1

[20800 rows x 5 columns]
```

[5]:
```
# since there are two id one inbuilt and one from dataset so i want to use only␣
↪one.
```

[5]:
```
# import the dataset:
df=pd.read_csv("train.csv",index_col='id')
df
```

[5]:
```
                                                 title  \
id
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 Fired
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 Alive

                                      author  \
id
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  label
id
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
…                                                    …    …
20795  Rapper T. I. unloaded on black celebrities who…      0
20796  When the Green Bay Packers lost to the Washing…      0
20797  The Macy's of today grew from the union of sev…      0
20798  NATO, Russia To Hold Parallel Exercises In Bal…      1
20799    David Swanson is an author, activist, journa…      1

[20800 rows x 4 columns]
```

[6]:
```
df.head()
```

[6]:
```
                                              title              author  \
id
0   House Dem Aide: We Didn't Even See Comey's Let…       Darrell Lucus
1   FLYNN: Hillary Clinton, Big Woman on Campus - …     Daniel J. Flynn
2                     Why the Truth Might Get You Fired  Consortiumnews.com
3   15 Civilians Killed In Single US Airstrike Hav…     Jessica Purkiss
4   Iranian woman jailed for fictional unpublished…      Howard Portnoy

                                                  text  label
id
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
```

[7]:
```
df.tail()
```

[7]:
```
                                                  title  \
id
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 Alive

                                        author  \
id
20795                              Jerome Hudson
20796                            Benjamin Hoffman
20797  Michael J. de la Merced and Rachel Abrams
20798                                Alex Ansary
20799                              David Swanson

                                                  text  label
id
20795  Rapper T. I. unloaded on black celebrities who…      0
20796  When the Green Bay Packers lost to the Washing…      0
20797  The Macy’s of today grew from the union of sev…      0
20798  NATO, Russia To Hold Parallel Exercises In Bal…      1
20799    David Swanson is an author, activist, journa…      1
```

[8]:
```
df.sample()
```

[8]:
```
                                                 title  \
id
4846  Is ‘Brexit’ the Precursor to a Donald Trump Pr…

                                      author  \
id
4846  Jonathan Martin and Alexander Burns

                                                 text  label
id
4846  WASHINGTON  -   Britain’s vote to withdraw fro…      0
```

[9]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 20800 entries, 0 to 20799
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   title   20242 non-null  object
 1   author  18843 non-null  object
 2   text    20761 non-null  object
 3   label   20800 non-null  int64 
dtypes: int64(1), object(3)
memory usage: 812.5+ KB
```

1.4 key observations of each column:

- id : unique id for each news article.
- title: title of news article.
- author: name of person who wrote the article.
- text: full content of news article.
- label: 0– means original news // 1– means fake news

---

from above info: * missing data in id,author , text ,

[11]:
```
# check the unique values in each columns:
```

[10]:
```
df['title'].unique()
```

[10]:
```
array(['House Dem Aide: We Didn’t Even See Comey’s Letter Until Jason Chaffetz
Tweeted It',
       'FLYNN: Hillary Clinton, Big Woman on Campus - Breitbart',
       'Why the Truth Might Get You Fired', …,
       'N.F.L. Playoffs: Schedule, Matchups and Odds - The New York Times',
       'Macy’s Is Said to Receive Takeover Approach by Hudson’s Bay - The New
York Times',
       'NATO, Russia To Hold Parallel Exercises In Balkans'], dtype=object)
```

[11]:
```
df['author'].unique()
```

[11]:
```
array(['Darrell Lucus', 'Daniel J. Flynn', 'Consortiumnews.com', …,
       'D. Samuelson', 'Judge Andrew Napolitano',
       'Michael J. de la Merced and Rachel Abrams'], dtype=object)
```

[17]:
```
df['text'].nunique()
```

[17]:
```
20386
```

[18]:
```
df['label'].value_counts()
```

[18]:
```
label
1    10413
0    10387
Name: count, dtype: int64
```

[19]:
```
# check the missing values;
```

[12]:
```
df['title'].isnull().sum()
```

[12]:
```
558
```

[13]:
```
df[df['title'].isnull()]
```

[13]:
```
       title             author  \
id
53       NaN        Dairy
120      NaN          Anonymous
124      NaN  SeekSearchDestory
140      NaN          Anonymous
196      NaN             Raffie
…      …                  …
20568    NaN        Cathy Milne
20627    NaN             Ramona
20636    NaN        Dave Lowery
20771    NaN         Letsbereal
20772    NaN        beersession

                                                    text  label
id
53     Sounds like he has our president pegged. What …      1
120    Same people all the time , i dont know how you…      1
124    You know, outside of any morality arguments, i…      1
140    There is a lot more than meets the eye to this…      1
196                They got the heater turned up on high.      1
…                                                    …    …
20568  Amusing comment Gary! "Those   week!" So, are …      1
20627  No she doesn’t have more money than God, every…      1
20636                                Trump all the way!      1
20771  DYN's Statement on Last Week's Botnet Attack h…      1
20772  Kinda reminds me of when Carter gave away the …      1

[558 rows x 4 columns]
```

[14]:
```
df['author'].isnull().sum()
```

[14]:
```
1957
```

[25]:
```
# here it is news so we cannot impute such values according to our prediction␣
 ↪and will.
# we cannot use mean, mediana and mode and others too.
# so we have to fill up the gaps only
```

[16]:
```
df=df.fillna('')
df.head()
```

[16]:

id	title	author	text	label
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	FLYNN: Hillary Clinton, Big Woman on Campus - …	Daniel J. Flynn	Ever get the feeling your life circles the rou…	0
2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, …	1
3	15 Civilians Killed In Single US Airstrike Hav…	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr…	1
4	Iranian woman jailed for fictional unpublished…	Howard Portnoy	Print \nAn Iranian woman has been sentenced to…	1

[17]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 20800 entries, 0 to 20799
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   title   20800 non-null  object
 1   author  20800 non-null  object
 2   text    20800 non-null  object
 3   label   20800 non-null  int64 
dtypes: int64(1), object(3)
memory usage: 812.5+ KB
```

[18]:
```
df['content']=df['author']+ ' ' + df['title']
```

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

[19]:

id	title	author	content
0	House Dem Aide: We Didn’t Even See Comey’s Let…	Darrell Lucus	Darrell Lucus House Dem Aide: We Didn’t Even See Comey’s Let…
1	FLYNN: Hillary Clinton, Big Woman on Campus - …	Daniel J. Flynn	Daniel J. Flynn FLYNN: Hillary Clinton, Big Woman on Campus - …
2	Why the Truth Might Get You Fired	Consortiumnews.com	Consortiumnews.com Why the Truth Might Get You Fired

```
3   15 Civilians Killed In Single US Airstrike Hav…     Jessica Purkiss
4   Iranian woman jailed for fictional unpublished…      Howard Portnoy

                                                  text  label  \
id
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
id
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…
```

[20]:
```
print(df['content'])
```

```
id
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
```

[21]:
```
df.drop(columns=['title','author'],axis=1,inplace=True)
```

[22]:
```
df.head()
```

[22]:
```
                                                  text  label  \
id
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
id
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…
```

[23]:
```
X=df.drop(columns='label',axis=1)
y=df['label']
```

[24]:
```
print(X)
```

```
                                                    text  \
id
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
id
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

[20800 rows x 2 columns]
```

[25]:
```
print(y)
```

```
id
0       1
```

```
1        0
2        1
3        1
4        1
        ..
20795    0
20796    0
20797    0
20798    1
20799    1
Name: label, Length: 20800, dtype: int64
```

[26]:
```
port_stream=PorterStemmer()
```

[27]:
```
def steam(content):
    stem_content = re.sub('[^a-zA-Z]', ' ', content)
    stem_content = stem_content.lower()
    stem_content = stem_content.split()
    stem_content = [port_stream.stem(word) for word in stem_content if not word␣
 ↪in stopwords.words('english')]
    stem_content = ' '.join(stem_content)
    return stem_content

# This function cleans the input text by:

# Removing special characters,
# Converting to lowercase,
# Splitting into words,
# Removing stopwords, and
# Stemming the words to their base form.
```

[28]:
```
df['content']=df['content'].apply(steam)
```

[30]:
```
print(df['content'])
```

```
id
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…
20797    michael j de la merc rachel abram maci said re…
20798    alex ansari nato russia hold parallel exercis …
20799                            david swanson keep f aliv
Name: content, Length: 20800, dtype: object
```

[31]:
```
df['content'][1]
```

[31]:
```
'daniel j flynn flynn hillari clinton big woman campu breitbart'
```

[40]:
```
# now again
```

[32]:
```
X=df['content'].values
y=df['label'].values
# this converts the content and label column into a NumPy array
```

[33]:
```
tfv=TfidfVectorizer()
```

[34]:
```
tfv.fit(X)
```

[34]:
```
TfidfVectorizer()
```

[35]:
```
X=tfv.transform(X)
```

[36]:
```
print(X)
```

```
  (0, 15686)    0.28485063562728646
  (0, 13473)    0.2565896679337957
  (0, 8909)     0.3635963806326075
  (0, 8630)     0.29212514087043684
  (0, 7692)     0.24785219520671603
  (0, 7005)     0.21874169089359144
  (0, 4973)     0.233316966909351
  (0, 3792)     0.2705332480845492
  (0, 3600)     0.3598939188262559
  (0, 2959)     0.2468450128533713
  (0, 2483)     0.3676519686797209
  (0, 267)      0.27010124977708766
  (1, 16799)    0.30071745655510157
  (1, 6816)     0.1904660198296849
  (1, 5503)     0.7143299355715573
  (1, 3568)     0.26373768806048464
  (1, 2813)     0.19094574062359204
  (1, 2223)     0.3827320386859759
  (1, 1894)     0.15521974226349364
  (1, 1497)     0.2939891562094648
  (2, 15611)    0.41544962664721613
  (2, 9620)     0.49351492943649944
  (2, 5968)     0.3474613386728292
  (2, 5389)     0.3866530551182615
  (2, 3103)     0.46097489583229645
  :     :
  (20797, 13122)        0.2482526352197606
```

```
(20797, 12344)        0.27263457663336677
(20797, 12138)        0.24778257724396507
(20797, 10306)        0.08038079000566466
(20797, 9588) 0.174553480255222
(20797, 9518) 0.2954204003420313
(20797, 8988) 0.36160868928090795
(20797, 8364) 0.22322585870464118
(20797, 7042) 0.21799048897828688
(20797, 3643) 0.21155500613623743
(20797, 1287) 0.33538056804139865
(20797, 699)  0.30685846079762347
(20797, 43)   0.29710241860700626
(20798, 13046)        0.22363267488270608
(20798, 11052)        0.4460515589182236
(20798, 10177)        0.3192496370187028
(20798, 6889) 0.32496285694299426
(20798, 5032) 0.4083701450239529
(20798, 1125) 0.4460515589182236
(20798, 588)  0.3112141524638974
(20798, 350)  0.28446937819072576
(20799, 14852)        0.5677577267055112
(20799, 8036) 0.45983893273780013
(20799, 3623) 0.37927626273066584
(20799, 377)  0.5677577267055112
```

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

```
[1 0 1 … 0 1 1]
```

[38]:
```
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,random_state=3)
```

[39]:
```
model=LogisticRegression()
```

[40]:
```
model.fit(X_train,y_train)
```

[40]:
```
LogisticRegression()
```

[41]:
```
y_pred_train=model.predict(X_train)
print(y_pred_train)
print("accuracy score for training data␣
  ↪is",accuracy_score(y_pred_train,y_train))
```

```
[0 1 1 … 0 0 1]
accuracy score for training data is 0.9873197115384615
```

[42]:
```
y_pred_test=model.predict(X_test)
print(y_pred_test)
print("accuracy score for test data is",accuracy_score(y_pred_test,y_test))
```

```
[1 1 1 … 0 0 0]
accuracy score for test data is 0.9757211538461539
```

[43]:
```
res=X_test[1]
predict=model.predict(res)
print(predict)
if(predict[0]==0):
    print("original news")
else:
    print("fake news")
```

```
[1]
fake news
```

[44]:
```
input_data='daniel j flynn flynn hillari clinton big woman campu breitbart'
input_list = [input_data]
data = [steam(text) for text in input_list]
det=tfv.transform(data)
predic=model.predict(det)
print(predic)

if(predic[0]==0):
    print("original news")
else:
    print("fake news")
```

```
[0]
original news
```

[ ]: