fake-news

November 6, 2024

About the Dataset:

1. id: unique id for a new article
2. title: the title of a news article
3. author: author of the news article
4. text: the text of the article; could be incomplete
5. label: a label that makes whether the news article is real or fake

```
0. Fake news
1. Real news
```

0.1 Importing libraries

[1]:
```python
import numpy as np
import pandas as pd
import re
from nltk.corpus import stopwords #words that are of no importance
from nltk.stem import PorterStemmer #root word for a particular word
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
```

[2]:
```python
#loading the dataset to a pandas dataframe
news_dataset = pd.read_csv('train.csv')
```

0.1.1 About dataset

[3]:
```python
#first 5 rows of the dataframe
news_dataset.head()
```

[3]:
```
   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4   4  Iranian woman jailed for fictional unpublished…      Howard Portnoy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
```

[4]:
```
news_dataset.shape
```

[4]:
```
(20800, 5)
```

[5]:
```
# Dataset Overview
news_dataset.info()
```

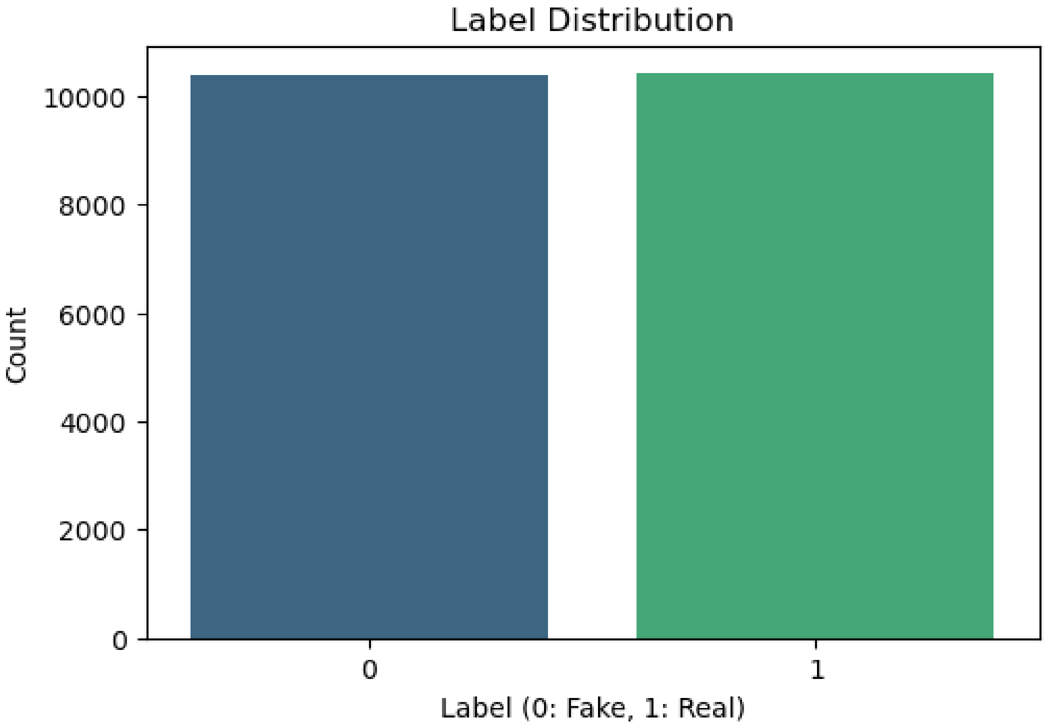
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20800 entries, 0 to 20799
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   id      20800 non-null  int64
 1   title   20242 non-null  object
 2   author  18843 non-null  object
 3   text    20761 non-null  object
 4   label   20800 non-null  int64
dtypes: int64(2), object(3)
memory usage: 812.6+ KB
```

[6]:
```
#counting the number of missing values in each column
news_dataset.isnull().sum()
```

[6]:
```
id           0
title      558
author    1957
text        39
label        0
dtype: int64
```

[7]:
```
import matplotlib.pyplot as plt
import seaborn as sns

# Label distribution plot
plt.figure(figsize=(6, 4))
sns.countplot(data=news_dataset, x='label', palette='viridis')
plt.title("Label Distribution")
plt.xlabel("Label (0: Fake, 1: Real)")
plt.ylabel("Count")
plt.show()
```



0.1.2 Data Preprocessing

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

```
[8]:           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]
```

```
[9]: # merging the author name and news title
     news_dataset['content'] = news_dataset['author'] + ' ' + news_dataset['title']
     news_dataset['content'].head()
```

```
[9]: 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…
     Name: content, dtype: object
```

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

```
[11]: X.head()
```

```
[11]:    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  \
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…

                                             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…
```

[12]:

```python
Y.head()
```

[12]:

```
0    1
1    0
2    1
3    1
4    1
Name: label, dtype: int64
```

0.1.3

StopWords

[13]:

```python
import nltk
nltk.download('stopwords')
```

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

[13]:

```
True
```

[31]:

```python
#printing the stopwords in English
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"]
```

0.1.4

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

example:

- actor, actress, acting –> act

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

```
[16]: 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
```

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

```
[17]: 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…
      Name: content, dtype: object
```

```
[18]: # separating the data and label
      X = news_dataset['content'].values
      Y = news_dataset['label'].values
```

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

```
['darrel lucu hous dem aid even see comey letter jason chaffetz tweet'
```

```
 'daniel j flynn flynn hillari clinton big woman campu breitbart'
 'consortiumnew com truth might get fire' …
 'michael j de la merc rachel abram maci said receiv takeov approach hudson bay
new york time'
 'alex ansari nato russia hold parallel exercis balkan'
 'david swanson keep f aliv']
```

[20]:

```python
print(Y)
```

```
[1 0 1 … 0 1 1]
```

0.1.5 ———

feature engineering

[21]:

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

X = vectorizer.transform(X)
```

[22]:

```python
X
```

[22]:

```
<20800x17128 sparse matrix of type '<class 'numpy.float64'>'
        with 210687 stored elements in Compressed Sparse Row format>
```

0.1.6 Splitting the dataset to training and test data

[23]:

```python
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2,␣
 ↪stratify = Y, random_state = 2)
```

0.1.7 Training the model: Logistic Regression

[24]:

```python
model = LogisticRegression()
```

[25]:

```python
model.fit(X_train, Y_train)
```

[25]:

```
LogisticRegression()
```

0.1.8 Making a Predictive system

[26]:

```python
X_new = X_test[1]

prediction = model.predict(X_new)
print(prediction)

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

```
[0]
The news is real
```

[27]:

```
print(Y_test[1])
```

```
0
```

0.1.9 Evaluation

[28]:

```
#accuracy score on the traing data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

[29]:

```
#accuracy score on the test data
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

[30]:

```
print('Accuracy score of the training data : ', training_data_accuracy)
print('Accuracy score of the test data : ', test_data_accuracy)
```

```
Accuracy score of the training data :  0.9865985576923076
Accuracy score of the test data :  0.9790865384615385
```

[ ]:

```

```