spam-news-detection-1

August 3, 2024

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

[6]:
```
import pandas as pd

path1 = ""
path2 = ""

# Try reading the file with error handling and a different delimiter
try:
    true_news = pd.read_csv("true.csv", encoding='latin-1',␣
  ↪on_bad_lines='skip', sep=',')
    fake_news = pd.read_csv("fake.csv", encoding='latin-1',␣
  ↪on_bad_lines='skip', sep=',')
except pd.errors.ParserError as e:
    print(f"Error reading file: {e}")
    # Inspect the problematic line(s) for issues
    print(f"Problematic line in {path1}:")
    with open(path1, 'r', encoding='latin-1') as f:
        for i, line in enumerate(f):
            if i == 11066:  # Adjust line number based on error message
                print(line)
                break
```

[7]:
```
true_news
```

[7]:
```
                                                    title  \
0       As U.S. budget fight looms, Republicans flip t…
1       U.S. military to accept transgender recruits o…
2       Senior U.S. Republican senator: 'Let Mr. Muell…
3       FBI Russia probe helped by Australian diplomat…
4       Trump wants Postal Service to charge 'much mor…
…                                                   …
21411  'Fully committed' NATO backs new U.S. approach…
21412  LexisNexis withdrew two products from Chinese …
21413  Minsk cultural hub becomes haven from authorities
21414  Vatican upbeat on possibility of Pope Francis …
```

```
21415  Indonesia to buy $1.14 billion worth of Russia…

                                                    text       subject  \
0      WASHINGTON (Reuters) - The head of a conservat…  politicsNews
1      WASHINGTON (Reuters) - Transgender people will…  politicsNews
2      WASHINGTON (Reuters) - The special counsel inv…  politicsNews
3      WASHINGTON (Reuters) - Trump campaign adviser …  politicsNews
4      SEATTLE/WASHINGTON (Reuters) - President Donal…  politicsNews
…                                                    …             …
21411  BRUSSELS (Reuters) - NATO allies on Tuesday we…     worldnews
21412  LONDON (Reuters) - LexisNexis, a provider of l…     worldnews
21413  MINSK (Reuters) - In the shadow of disused Sov…     worldnews
21414  MOSCOW (Reuters) - Vatican Secretary of State …     worldnews
21415  JAKARTA (Reuters) - Indonesia will buy 11 Sukh…     worldnews

                     date
0      December 31, 2017
1      December 29, 2017
2      December 31, 2017
3      December 30, 2017
4      December 29, 2017
…                    …
21411    August 22, 2017
21412    August 22, 2017
21413    August 22, 2017
21414    August 22, 2017
21415    August 22, 2017

[21416 rows x 4 columns]
```

```
[8]: fake_news
```

```
[8]:                                                   title  \
0       Donald Trump Sends Out Embarrassing New Yearâ…
1       Drunk Bragging Trump Staffer Started Russian …
2       Sheriff David Clarke Becomes An Internet Joke…
3       Trump Is So Obsessed He Even Has Obamaâ s Na…
4       Pope Francis Just Called Out Donald Trump Dur…
…                                                  …
23476  McPain: John McCain Furious That Iran Treated …
23477  JUSTICE? Yahoo Settles E-mail Privacy Class-ac…
23478  Sunnistan: US and Allied â Safe Zoneâ  Plan …
23479  How to Blow $700 Million: Al Jazeera America F…
23480  10 U.S. Navy Sailors Held by Iranian Military …

                                                    text      subject  \
0      Donald Trump just couldn t wish all Americans …         News
```

```
1      House Intelligence Committee Chairman Devin Nu…         News
2      On Friday, it was revealed that former Milwauk…         News
3      On Christmas day, Donald Trump announced that …         News
4      Pope Francis used his annual Christmas Day mes…         News
…                                                   …          …
23476  21st Century Wire says As 21WIRE reported earl…  Middle-east
23477  21st Century Wire says It s a familiar theme. …  Middle-east
23478  Patrick Henningsen  21st Century WireRemember …  Middle-east
23479  21st Century Wire says Al Jazeera America will…  Middle-east
23480  21st Century Wire says As 21WIRE predicted in …  Middle-east

                    date
0      December 31, 2017
1      December 31, 2017
2      December 30, 2017
3      December 29, 2017
4      December 25, 2017
…                    …
23476   January 16, 2016
23477   January 16, 2016
23478   January 15, 2016
23479   January 14, 2016
23480   January 12, 2016

[23481 rows x 4 columns]
```

[9]:

```
true_news['label'] = 0
fake_news['label'] = 1
```

[10]:

```
true_news
```

[10]:

```
                                                   title  \
0      As U.S. budget fight looms, Republicans flip t…
1      U.S. military to accept transgender recruits o…
2      Senior U.S. Republican senator: 'Let Mr. Muell…
3      FBI Russia probe helped by Australian diplomat…
4      Trump wants Postal Service to charge 'much mor…
…                                                   …
21411  'Fully committed' NATO backs new U.S. approach…
21412  LexisNexis withdrew two products from Chinese …
21413  Minsk cultural hub becomes haven from authorities
21414  Vatican upbeat on possibility of Pope Francis …
21415  Indonesia to buy $1.14 billion worth of Russia…

                                                    text       subject  \
0      WASHINGTON (Reuters) - The head of a conservat…  politicsNews
1      WASHINGTON (Reuters) - Transgender people will…  politicsNews
```

```
2      WASHINGTON (Reuters) - The special counsel inv…  politicsNews
3      WASHINGTON (Reuters) - Trump campaign adviser …  politicsNews
4      SEATTLE/WASHINGTON (Reuters) - President Donal…  politicsNews
…                                                …            …
21411  BRUSSELS (Reuters) - NATO allies on Tuesday we…     worldnews
21412  LONDON (Reuters) - LexisNexis, a provider of l…     worldnews
21413  MINSK (Reuters) - In the shadow of disused Sov…     worldnews
21414  MOSCOW (Reuters) - Vatican Secretary of State …     worldnews
21415  JAKARTA (Reuters) - Indonesia will buy 11 Sukh…     worldnews

                     date  label
0      December 31, 2017      0
1      December 29, 2017      0
2      December 31, 2017      0
3      December 30, 2017      0
4      December 29, 2017      0
…                     …    …
21411    August 22, 2017      0
21412    August 22, 2017      0
21413    August 22, 2017      0
21414    August 22, 2017      0
21415    August 22, 2017      0

[21416 rows x 5 columns]
```

[11]:

```
fake_news
```

[11]:

```
                                                   title  \
0       Donald Trump Sends Out Embarrassing New Yearâ…
1       Drunk Bragging Trump Staffer Started Russian …
2       Sheriff David Clarke Becomes An Internet Joke…
3       Trump Is So Obsessed He Even Has Obamaâ s Na…
4       Pope Francis Just Called Out Donald Trump Dur…
…                                                …
23476  McPain: John McCain Furious That Iran Treated …
23477  JUSTICE? Yahoo Settles E-mail Privacy Class-ac…
23478  Sunnistan: US and Allied â Safe Zoneâ  Plan …
23479  How to Blow $700 Million: Al Jazeera America F…
23480  10 U.S. Navy Sailors Held by Iranian Military …

                                                    text      subject  \
0      Donald Trump just couldn t wish all Americans …         News
1      House Intelligence Committee Chairman Devin Nu…         News
2      On Friday, it was revealed that former Milwauk…         News
3      On Christmas day, Donald Trump announced that …         News
4      Pope Francis used his annual Christmas Day mes…         News
…                                                …           …
```

```
23476  21st Century Wire says As 21WIRE reported earl…  Middle-east
23477  21st Century Wire says It s a familiar theme. …  Middle-east
23478  Patrick Henningsen  21st Century WireRemember …  Middle-east
23479  21st Century Wire says Al Jazeera America will…  Middle-east
23480  21st Century Wire says As 21WIRE predicted in …  Middle-east

                    date  label
0      December 31, 2017      1
1      December 31, 2017      1
2      December 30, 2017      1
3      December 29, 2017      1
4      December 25, 2017      1
…                  …    …
23476   January 16, 2016      1
23477   January 16, 2016      1
23478   January 15, 2016      1
23479   January 14, 2016      1
23480   January 12, 2016      1

[23481 rows x 5 columns]
```

[12]:

```python
dataset1 = true_news[['text','label']]
dataset2 = fake_news[['text','label']]
```

[13]:

```python
dataset1
```

[13]:

```
                                                    text  label
0      WASHINGTON (Reuters) - The head of a conservat…      0
1      WASHINGTON (Reuters) - Transgender people will…      0
2      WASHINGTON (Reuters) - The special counsel inv…      0
3      WASHINGTON (Reuters) - Trump campaign adviser …      0
4      SEATTLE/WASHINGTON (Reuters) - President Donal…      0
…                                                  …    …
21411  BRUSSELS (Reuters) - NATO allies on Tuesday we…      0
21412  LONDON (Reuters) - LexisNexis, a provider of l…      0
21413  MINSK (Reuters) - In the shadow of disused Sov…      0
21414  MOSCOW (Reuters) - Vatican Secretary of State …      0
21415  JAKARTA (Reuters) - Indonesia will buy 11 Sukh…      0

[21416 rows x 2 columns]
```

[14]:

```python
dataset2
```

[14]:

```
                                                    text  label
0      Donald Trump just couldn t wish all Americans …      1
1      House Intelligence Committee Chairman Devin Nu…      1
2      On Friday, it was revealed that former Milwauk…      1
```

```
3      On Christmas day, Donald Trump announced that …      1
4      Pope Francis used his annual Christmas Day mes…      1
…                                                 …     …
23476  21st Century Wire says As 21WIRE reported earl…      1
23477  21st Century Wire says It s a familiar theme. …      1
23478  Patrick Henningsen  21st Century WireRemember …      1
23479  21st Century Wire says Al Jazeera America will…      1
23480  21st Century Wire says As 21WIRE predicted in …      1

[23481 rows x 2 columns]
```

[15]:
```
dataset = pd.concat([dataset1, dataset2])
```

[16]:
```
dataset
```

[16]:
```
                                                     text  label
0      WASHINGTON (Reuters) - The head of a conservat…      0
1      WASHINGTON (Reuters) - Transgender people will…      0
2      WASHINGTON (Reuters) - The special counsel inv…      0
3      WASHINGTON (Reuters) - Trump campaign adviser …      0
4      SEATTLE/WASHINGTON (Reuters) - President Donal…      0
…                                                 …     …
23476  21st Century Wire says As 21WIRE reported earl…      1
23477  21st Century Wire says It s a familiar theme. …      1
23478  Patrick Henningsen  21st Century WireRemember …      1
23479  21st Century Wire says Al Jazeera America will…      1
23480  21st Century Wire says As 21WIRE predicted in …      1

[44897 rows x 2 columns]
```

[17]:
```
dataset.shape
```

[17]:
```
(44897, 2)
```

[18]:
```
dataset.isnull().sum()
```

[18]:
```
text     0
label    0
dtype: int64
```

[19]:
```
dataset['label'].value_counts()
```

[19]:
```
label
1    23481
0    21416
Name: count, dtype: int64
```

[20]:

```
dataset = dataset.sample(frac = 1)
```

[21]:

```
dataset
```

[21]:

```
                                                    text  label
13538  If you haven t seethe viral video of a takedow…      1
9415   After reporter and model, Leeann Tweeden accus…      1
15214  DUBAI/ZURICH (Reuters) - Saudi Arabia has anno…      0
18193  GENEVA (Reuters) - Muslim Rohingyas continue t…      0
21980  Finian Cunningham SputnikGoogle is the latest …      1
…                                                    …    …
6432   An Arizona gun nut has been charged with  mult…      1
13990  Hey Glenn, suicide is no laughing matter, and …      1
20204  Maybe Hillary s Russian uranium deal included …      1
10681  President Trump has nominated Christopher Wray…      1
13477  MOSCOW (Reuters) - Russia accused the United S…      0

[44897 rows x 2 columns]
```

[23]:

```
import nltk
import re
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
```

[24]:

```
ps  = WordNetLemmatizer()
```

[25]:

```
nltk.download('wordnet')
nltk.download('omw-1.4')
nltk.download('stopwords')
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\chinu\AppData\Roaming\nltk_data…
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     C:\Users\chinu\AppData\Roaming\nltk_data…
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\chinu\AppData\Roaming\nltk_data…
[nltk_data]   Unzipping corpora\stopwords.zip.
```

[25]:

```
True
```

[26]:

```
stopwords = stopwords.words('english')
```

[27]:

```
def clean_row(row):
    row = row.lower()
    row = re.sub('[^a-zA-Z]', ' ', row)
    token = row.split()
    news = [ps.lemmatize(word) for word in token if not word in stopwords]
    cleaned_news = ' '.join(news)
    return cleaned_news
```

```
[28]: dataset['text']
```

```
[28]: 13538    If you haven t seethe viral video of a takedow…
      9415     After reporter and model, Leeann Tweeden accus…
      15214    DUBAI/ZURICH (Reuters) - Saudi Arabia has anno…
      18193    GENEVA (Reuters) - Muslim Rohingyas continue t…
      21980    Finian Cunningham SputnikGoogle is the latest …
                                     …
      6432     An Arizona gun nut has been charged with  mult…
      13990    Hey Glenn, suicide is no laughing matter, and …
      20204    Maybe Hillary s Russian uranium deal included …
      10681    President Trump has nominated Christopher Wray…
      13477    MOSCOW (Reuters) - Russia accused the United S…
      Name: text, Length: 44897, dtype: object
```

```
[30]: dataset['text'] = dataset['text'].apply(lambda x : clean_row(x))
```

```
[31]: dataset['text']
```

```
[31]: 13538    seethe viral video takedown russian guard one …
      9415     reporter model leeann tweeden accused franken …
      15214    dubai zurich reuters saudi arabia announced co…
      18193    geneva reuters muslim rohingyas continue flee …
      21980    finian cunningham sputnikgoogle latest u inter…
                                     …
      6432     arizona gun nut charged multiple crime decided…
      13990    hey glenn suicide laughing matter really need …
      20204    maybe hillary russian uranium deal included ca…
      10681    president trump nominated christopher wray new…
      13477    moscow reuters russia accused united state thu…
      Name: text, Length: 44897, dtype: object
```

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

```
[33]: vectorizer = TfidfVectorizer(max_features = 50000, lowercase = False,␣
        ↪ngram_range = (1, 2))
```

```
[34]: x = dataset.iloc[:40000, 0]
      y = dataset.iloc[:40000, 1]
```

```
[35]: x
```

```
[35]: 13538    seethe viral video takedown russian guard one …
      9415     reporter model leeann tweeden accused franken …
```

```
15214    dubai zurich reuters saudi arabia announced co…
18193    geneva reuters muslim rohingyas continue flee …
21980    finian cunningham sputnikgoogle latest u inter…
                                ...
21011    rampant migrant rape violence shining light pr…
12986    steven crowder knock park brilliant imitation …
22828    st century wire say nevada assemblywoman miche…
20922    another example public school using globalist …
1026     washington reuters department justice need imm…
Name: text, Length: 40000, dtype: object
```

```
[36]: y
```

```
[36]: 13538    1
      9415     1
      15214    0
      18193    0
      21980    1
              ..
      21011    1
      12986    1
      22828    1
      20922    1
      1026     0
      Name: label, Length: 40000, dtype: int64
```

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

```
[38]: train_data, test_data, train_label, test_label = train_test_split(x, y,␣
        ↪test_size = 0.2, random_state = 0)
```

```
[39]: vec_train_data = vectorizer.fit_transform(train_data)
```

```
[40]: vec_train_data = vec_train_data.toarray()
```

```
[41]: vec_test_data = vectorizer.fit_transform(test_data)
```

```
[42]: vec_test_data = vec_test_data.toarray()
```

```
[43]: vec_train_data.shape, vec_test_data.shape
```

```
[43]: ((32000, 50000), (8000, 50000))
```

```
[44]: train_data = pd.DataFrame(vec_train_data, columns = vectorizer.
        ↪get_feature_names_out())
      test_data = pd.DataFrame(vec_test_data, columns = vectorizer.
        ↪get_feature_names_out())
```

[45]:

```
train_data
```

[45]:

```
        aa  aaf  aapl  aaron  aaronson   ab  aback  abadi  abadi office  \
0      0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
1      0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
2      0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
3      0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
4      0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
…    …  …   …    …       …  …    …    …           …
31995  0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
31996  0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
31997  0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
31998  0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
31999  0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0

       abadi said  …  zone would  zoning  zor  zor province  zschaepe  \
0             0.0  …         0.0     0.0  0.0           0.0       0.0
1             0.0  …         0.0     0.0  0.0           0.0       0.0
2             0.0  …         0.0     0.0  0.0           0.0       0.0
3             0.0  …         0.0     0.0  0.0           0.0       0.0
4             0.0  …         0.0     0.0  0.0           0.0       0.0
…           …  …         …     …  …           …       …
31995         0.0  …         0.0     0.0  0.0           0.0       0.0
31996         0.0  …         0.0     0.0  0.0           0.0       0.0
31997         0.0  …         0.0     0.0  0.0           0.0       0.0
31998         0.0  …         0.0     0.0  0.0           0.0       0.0
31999         0.0  …         0.0     0.0  0.0           0.0       0.0

       zucker  zuckerberg  zuckerberg said  zuma  zurich
0         0.0         0.0              0.0   0.0     0.0
1         0.0         0.0              0.0   0.0     0.0
2         0.0         0.0              0.0   0.0     0.0
3         0.0         0.0              0.0   0.0     0.0
4         0.0         0.0              0.0   0.0     0.0
…       …         …              …   …     …
31995     0.0         0.0              0.0   0.0     0.0
31996     0.0         0.0              0.0   0.0     0.0
31997     0.0         0.0              0.0   0.0     0.0
31998     0.0         0.0              0.0   0.0     0.0
31999     0.0         0.0              0.0   0.0     0.0

[32000 rows x 50000 columns]
```

[46]:

```
test_data
```

[46]:

```
       aa  aaf  aapl  aaron  aaronson   ab  aback  abadi  abadi office  \
0     0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
```

```
1     0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
2     0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
3     0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
4     0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
…   …  …   …    …       …  …    …    …           …
7995  0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
7996  0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
7997  0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
7998  0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0
7999  0.0  0.0   0.0    0.0       0.0  0.0    0.0    0.0           0.0

      abadi said  …  zone would  zoning  zor  zor province  zschaepe  \
0            0.0  …         0.0     0.0  0.0           0.0       0.0
1            0.0  …         0.0     0.0  0.0           0.0       0.0
2            0.0  …         0.0     0.0  0.0           0.0       0.0
3            0.0  …         0.0     0.0  0.0           0.0       0.0
4            0.0  …         0.0     0.0  0.0           0.0       0.0
…          …  …         …     …  …           …       …
7995         0.0  …         0.0     0.0  0.0           0.0       0.0
7996         0.0  …         0.0     0.0  0.0           0.0       0.0
7997         0.0  …         0.0     0.0  0.0           0.0       0.0
7998         0.0  …         0.0     0.0  0.0           0.0       0.0
7999         0.0  …         0.0     0.0  0.0           0.0       0.0

      zucker  zuckerberg  zuckerberg said  zuma  zurich
0        0.0         0.0              0.0   0.0     0.0
1        0.0         0.0              0.0   0.0     0.0
2        0.0         0.0              0.0   0.0     0.0
3        0.0         0.0              0.0   0.0     0.0
4        0.0         0.0              0.0   0.0     0.0
…      …         …              …   …     …
7995     0.0         0.0              0.0   0.0     0.0
7996     0.0         0.0              0.0   0.0     0.0
7997     0.0         0.0              0.0   0.0     0.0
7998     0.0         0.0              0.0   0.0     0.0
7999     0.0         0.0              0.0   0.0     0.0

[8000 rows x 50000 columns]
```

[47]:

```python
from sklearn.naive_bayes import MultinomialNB
#from sklearn.linear_model import LogisticRegression
```

[48]:

```python
clf = MultinomialNB()
#clf = LogisticRegression()
```

[49]:

```python
clf.fit(train_data, train_label)
```

```
[49]: MultinomialNB()
```

```
[50]: y_pred = clf.predict(test_data)
```

```
[51]: test_label
```

```
[51]: 10600    0
      12491    0
      14991    0
      2069     0
      17311    1
              ..
      17877    1
      19238    0
      13998    0
      1315     0
      9854     1
      Name: label, Length: 8000, dtype: int64
```

```
[52]: y_pred
```

```
[52]: array([1, 0, 1, …, 0, 0, 0], dtype=int64)
```

```
[53]: from sklearn.metrics import accuracy_score
```

```
[54]: accuracy_score(test_label, y_pred)
```

```
[54]: 0.75675
```

```
[55]: y_pred_train = clf.predict(train_data)
```

```
[56]: accuracy_score(train_label, y_pred_train)
```

```
[56]: 0.9585625
```

```
[57]: # SL>NO              METHOD                  ACCURACY[TEST]
      #    1             naive_bayes                    75%
      #    2          logistic_regression               53%
```

```
[58]: '''txt = input("Enter News")
      news = clean_row(str(txt))
      pred = clf1.predict(vectorizer.transform([news]).toarray())'''
```

```
[58]: 'txt = input("Enter News")\nnews = clean_row(str(txt))\npred =
      clf1.predict(vectorizer.transform([news]).toarray())'
```

[59]:

```
'''if pred == 0:
  print('News is True')
else:
  print('News is Fake')'''
```

[59]:

```
"if pred == 0:\n  print('News is True')\nelse:\n  print('News is Fake')"
```

[ ]:

```

```