fake-and-true-news

June 28, 2024

```
[ ]: import pandas as pd
     import numpy as nm
     from sklearn.model_selection import train_test_split
     from sklearn.metrics import classification_report
     import re
     import string
     import matplotlib.pyplot as plt
```

```
[ ]: data_true=pd.read_csv("/content/drive/MyDrive/ml proj/True.csv")
     data_fake=pd.read_csv("/content/drive/MyDrive/ml proj/Fake.csv")
```

```
[ ]: from google.colab import drive
     drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
[ ]: data_true.head()
```

```
[ ]:                                              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…

                                                  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

                      date
     0  December 31, 2017
     1  December 29, 2017
     2  December 31, 2017
     3  December 30, 2017
     4  December 29, 2017
```

```
[ ]: data_true.shape , data_fake.shape
```

```
[ ]: ((21417, 4), (23481, 4))
```

```
[ ]: data_true['class']=0
     data_fake['class']=1
```

```
[ ]: data_true_manual_testing=data_true.tail(10)
     for i in range(21416,21406,-1):
       data_true.drop([i],axis=0,inplace=True)

     data_fake_manual_testing=data_fake.tail(10)
     for i in range(21416,21406,-1):
       data_fake.drop([i],axis=0,inplace=True)
```

```
[ ]: data_manual_testing = pd.concat([data_fake_manual_testing,␣
      ↪data_true_manual_testing], axis=0)
     data_manual_testing.to_csv("manual_testing.csv")
```

```
[ ]: data_merge=pd.concat([data_true,data_fake],axis=0)
     data_merge.head(10)
```

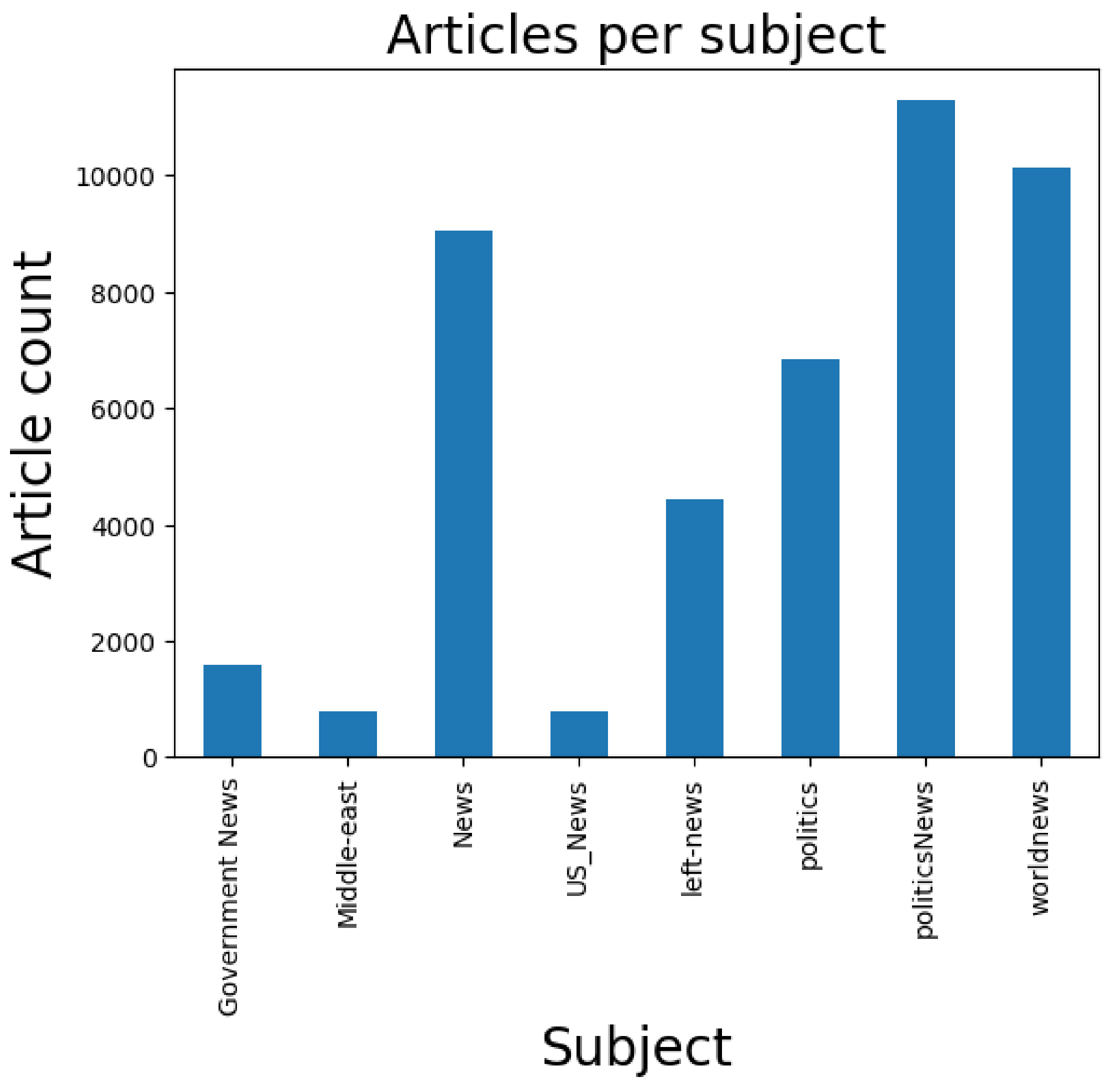
```
[ ]:                                                title  \
     0  As U.S. budget fight looms, Republicans flip t…
     1  U.S. military to accept transgender recruits o…
     2  Senior U.S. Republican senator: 'Let Mr. Muell…
     3  FBI Russia probe helped by Australian diplomat…
     4  Trump wants Postal Service to charge 'much mor…
     5  White House, Congress prepare for talks on spe…
     6  Trump says Russia probe will be fair, but time…
     7  Factbox: Trump on Twitter (Dec 29) - Approval …
     8         Trump on Twitter (Dec 28) - Global Warming
     9  Alabama official to certify Senator-elect Jone…

                                                     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
     5  WEST PALM BEACH, Fla./WASHINGTON (Reuters) - T…  politicsNews
     6  WEST PALM BEACH, Fla (Reuters) - President Don…  politicsNews
     7  The following statements were posted to the ve…  politicsNews
     8  The following statements were posted to the ve…  politicsNews
     9  WASHINGTON (Reuters) - Alabama Secretary of St…  politicsNews

                      date  class
```

```
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
5  December 29, 2017       0
6  December 29, 2017       0
7  December 29, 2017       0
8  December 29, 2017       0
9  December 28, 2017       0
```

[ ]:

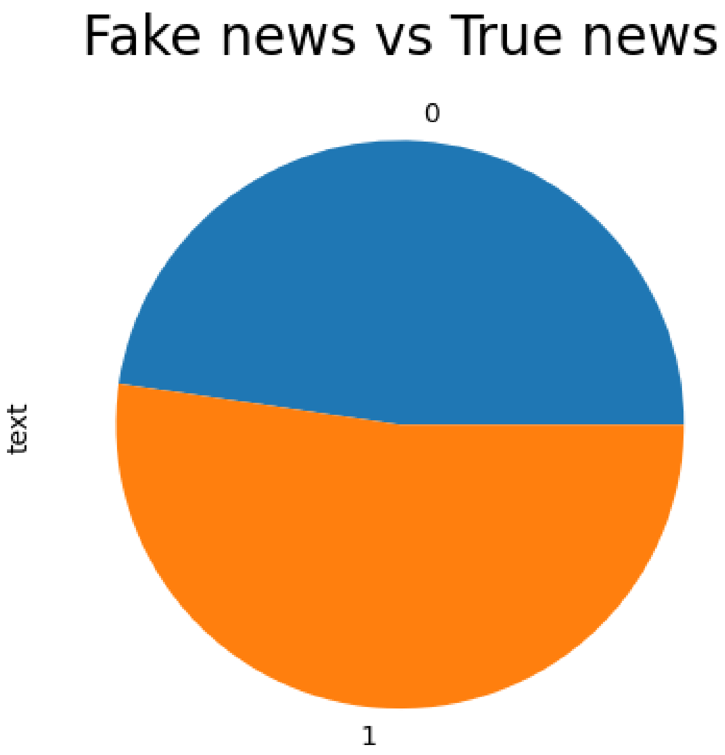
```
print(data_merge.groupby(['subject'])['text'].count())
data_merge.groupby(['subject'])['text'].count().plot(kind='bar')
plt.title("Articles per subject",size=20)
plt.xlabel("Subject",size=20)
plt.ylabel("Article count",size=20)
plt.show()
```

```
subject
Government News     1570
Middle-east          778
News                9050
US_News              783
left-news           4449
politics            6841
politicsNews       11272
worldnews          10135
Name: text, dtype: int64
```



```
[ ]: print(data_merge.groupby(['class'])['text'].count())
     print("0 = Fake news\n1 = True news")
     data_merge.groupby(['class'])['text'].count().plot(kind='pie')
     plt.title("Fake news vs True news",size=20)
     plt.show()
```

```
class
0    21407
1    23471
Name: text, dtype: int64
0 = Fake news
1 = True news
```



[ ]:
```
data = data_merge.drop(['title','subject','date'],axis=1)
data.head(10)
```

[ ]:
```
                                                text  class
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
5  WEST PALM BEACH, Fla./WASHINGTON (Reuters) - T…      0
6  WEST PALM BEACH, Fla (Reuters) - President Don…      0
7  The following statements were posted to the ve…      0
8  The following statements were posted to the ve…      0
9  WASHINGTON (Reuters) - Alabama Secretary of St…      0
```

[ ]:
```
data = data.sample(frac=1)
data.head(10)
```

[ ]:
```
                                                   text  class
1322   WASHINGTON (Reuters) - NFL team owners will co…      0
6982   Children are proof that hate is taught and lea…      1
8114   FAIRFAX, Va. (Reuters) - Democratic presidenti…      0
```

```
19339  Native Americans continue to battle poverty, j…      1
3677   President Obama couldn t be more different fro…      1
19696  Just look away. The Democrats don t have any p…      1
3550    (Corrects Comey firing to May 9 in fifth para…      0
11088  WASHINGTON (Reuters) - The leader of an influe…      0
15067  KRAKOW, Poland (Reuters) - Demanding reparatio…      0
11690  U.S. immigration authorities arrested hundreds…      1
```

```
[ ]: data.isnull().sum()
```

```
[ ]: text     0
     class    0
     dtype: int64
```

```
[ ]: def filter_text(data):
         text=data.lower()
         text=re.sub('\[.*?\]', '', text)
         text=re.sub("\\W"," ",text)
         text=re.sub('https?://\S+|www\.\S+', '', text)
         text=re.sub('<.*?>+', '', text)
         text=re.sub('[%s]' % re.escape(string.punctuation), '', text)
         text=re.sub('\n', '', text)
         text=re.sub('\w*\d\w*', '', text)
         return text
```

```
[ ]: data['text']=data['text'].apply(filter_text)
     data.head(10)
```

```
[ ]:                                                    text  class
     1322   washington  reuters    nfl team owners will co…      0
     6982   children are proof that hate is taught and lea…      1
     8114   fairfax  va   reuters    democratic presidenti…      0
     19339  native americans continue to battle poverty  j…      1
     3677   president obama couldn t be more different fro…      1
     19696  just look away  the democrats don t have any p…      1
     3550     corrects comey firing to may  in fifth parag…      0
     11088  washington  reuters    the leader of an influe…      0
     15067  krakow  poland  reuters    demanding reparatio…      0
     11690  u s  immigration authorities arrested hundreds…      1
```

```
[ ]: x=data['text'] #ind
     y=data['class'] #dep
```

```
[ ]: import pandas as pd
     from sklearn.model_selection import train_test_split
     from sklearn.feature_extraction.text import TfidfVectorizer
     from sklearn.linear_model import LogisticRegression
```

```python
from sklearn.metrics import accuracy_score, classification_report

# Sample data (replace with your actual data)
# Increased the size of the dataset to include more samples and ensure both␣
 ↪classes are present in the training set.
data = {'text': ['This is a positive sentence.', 'This is a negative sentence.
 ↪', 'Another positive one.', 'And a negative one.'],
        'class': [1, 0, 1, 0]}
data = pd.DataFrame(data) # Convert the dictionary to a DataFrame

# Now 'x' and 'y' can be defined
x=data['text'] #ind
y=data['class'] #dep

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2,␣
 ↪random_state=42) # Split the data into training and testing sets

# Fit a model to the training data. This was missing in the original code.
vectorizer = TfidfVectorizer()
x_train = vectorizer.fit_transform(x_train)
model = LogisticRegression()
model.fit(x_train, y_train)

# Transform the test data using the same vectorizer
x_test = vectorizer.transform(x_test)

# Make predictions on the test set
y_pred = model.predict(x_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

# Print classification report for more detailed evaluation
print(classification_report(y_test, y_pred))
```

```
Accuracy: 0.0
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       1.0
           1       0.00      0.00      0.00       0.0

    accuracy                           0.00       1.0
   macro avg       0.00      0.00      0.00       1.0
weighted avg       0.00      0.00      0.00       1.0


/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
```

```
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to
0.0 in labels with no predicted samples. Use `zero_division` parameter to
control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Recall and F-score are ill-defined and being set to 0.0
in labels with no true samples. Use `zero_division` parameter to control this
behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to
0.0 in labels with no predicted samples. Use `zero_division` parameter to
control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Recall and F-score are ill-defined and being set to 0.0
in labels with no true samples. Use `zero_division` parameter to control this
behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to
0.0 in labels with no predicted samples. Use `zero_division` parameter to
control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Recall and F-score are ill-defined and being set to 0.0
in labels with no true samples. Use `zero_division` parameter to control this
behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

[ ]:

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report

# Sample data (replace with your actual data)
# Increased the size of the dataset to include more samples and ensure both␣
 ↪classes are present in the training set.
data = {'text': ['This is a positive sentence.', 'This is a negative sentence.
 ↪', 'Another positive one.', 'And a negative one.'],
        'class': [1, 0, 1, 0]}
data = pd.DataFrame(data) # Convert the dictionary to a DataFrame

# Now 'x' and 'y' can be defined
x=data['text'] #ind
y=data['class'] #dep
```

```python
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42) # Split the data into training and testing sets

# Fit a model to the training data. This was missing in the original code.
vectorizer = TfidfVectorizer()
x_train = vectorizer.fit_transform(x_train) # Fit the vectorizer to the training data and transform it.
model = LogisticRegression()
model.fit(x_train, y_train)

# Transform the test data using the same vectorizer
x_test = vectorizer.transform(x_test)

# Make predictions on the test set
y_pred = model.predict(x_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

# Print classification report for more detailed evaluation
print(classification_report(y_test, y_pred))

# Define the filtering function here if it was not defined previously
def filtering(text):
    # Implement your text filtering logic here
    # For example, you might want to remove punctuation, convert to lowercase, etc.
    return text.lower()

def predict_news(text):
    text_vectorized = vectorizer.transform([filtering(text)])
    prediction = model.predict(text_vectorized)
    if prediction == 1:
        return "This news is likely fake."
    else:
        return "This news is likely true."

user_input = input("Enter news text: ")
result = predict_news(user_input)
print(result)
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to
0.0 in labels with no predicted samples. Use `zero_division` parameter to
control this behavior.
```

```
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Recall and F-score are ill-defined and being set to 0.0
in labels with no true samples. Use `zero_division` parameter to control this
behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to
0.0 in labels with no predicted samples. Use `zero_division` parameter to
control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Recall and F-score are ill-defined and being set to 0.0
in labels with no true samples. Use `zero_division` parameter to control this
behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to
0.0 in labels with no predicted samples. Use `zero_division` parameter to
control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Recall and F-score are ill-defined and being set to 0.0
in labels with no true samples. Use `zero_division` parameter to control this
behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

```
Accuracy: 0.0
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       1.0
           1       0.00      0.00      0.00       0.0

    accuracy                           0.00       1.0
   macro avg       0.00      0.00      0.00       1.0
weighted avg       0.00      0.00      0.00       1.0

Enter news text: modi died
This news is likely fake.
```

[ ]:

```python
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score, classification_report

# Load the dataset
data_true=pd.read_csv("/content/drive/MyDrive/ml proj/True.csv")
```

```
data_fake=pd.read_csv("/content/drive/MyDrive/ml proj/Fake.csv")

# Preprocess the data
x = data['text']
y = data['class']

# Vectorize the text data
vectorizer = TfidfVectorizer(max_features=1000)
x_vectorized = vectorizer.fit_transform(x)

# Split the data into training and testing sets
x_train, x_test, y_train, y_test = train_test_split(x_vectorized, y,␣
 ↪test_size=0.2, random_state=42)

# Train Decision Tree model
model = DecisionTreeClassifier()
model.fit(x_train, y_train)
y_pred = model.predict(x_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)

print("Accuracy:", accuracy)
print("Classification Report:\n", report)

# Function to get user input and predict output
def get_user_input():
    user_input = input("Enter news text: ")
    user_input_vectorized = vectorizer.transform([user_input])
    return user_input_vectorized

# Get user input and predict
user_input_vectorized = get_user_input()
user_prediction = model.predict(user_input_vectorized)

print("Prediction:", "Fake news" if user_prediction[0] == 0 else "True news")
```

```
Accuracy: 1.0
Classification Report:
               precision    recall  f1-score   support

           0       1.00      1.00      1.00         1

    accuracy                           1.00         1
   macro avg       1.00      1.00      1.00         1
weighted avg       1.00      1.00      1.00         1
```

```
Enter news text: india is a country
Prediction: True news
```

[ ]:
```
from sklearn.ensemble import RandomForestClassifier
```

[ ]:
```
RFC = RandomForestClassifier(random_state=0)
RFC.fit(x_train,y_train)
```

[ ]:
```
RandomForestClassifier(random_state=0)
```

[ ]:
```
# Import necessary libraries
from sklearn.ensemble import RandomForestClassifier

# Split the data into training and testing sets
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2,␣
 ↪random_state=42)

# Vectorize the text data using the fitted vectorizer
x_train_transformed = vectorizer.transform(x_train) # Transform x_train using␣
 ↪the fitted vectorizer
x_test_transformed = vectorizer.transform(x_test) # Transform x_test using the␣
 ↪fitted vectorizer

# Create a Random Forest classifier
RF = RandomForestClassifier()

# Train the classifier using the transformed data
RF.fit(x_train_transformed, y_train) # Use transformed x_train

# Make predictions on the test set (using transformed data)
y_pred_rf = RF.predict(x_test_transformed)

# Evaluate the accuracy of the predictions
print("Random Forest Accuracy:", accuracy_score(y_test, y_pred_rf))

# Get user input
user_input = input("Enter some text: ")

# Transform the user input using the fitted vectorizer
user_input_transformed = vectorizer.transform([user_input])

# Make prediction using the trained Random Forest
prediction_rf = RF.predict(user_input_transformed)[0] # Get the prediction␣
 ↪result

# Print prediction result
```

```
if prediction_rf == 1:
    print("The news is true")
else:
    print("The news is fake")
```

```
Random Forest Accuracy: 0.0
Enter some text: india is a country
The news is true
```