fake-news-predictions

January 10, 2024

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

[2]:
```
true_news = pd.read_csv('true.csv')
```

[3]:
```
true_news.head(5)
```

[3]:
```
                                               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
```

[4]:
```
true_news.shape
```

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

[5]:
```
fake_news = pd.read_csv('fake.csv')
```

[6]:
```
fake_news.shape
```

[6]:
```
(23481, 4)
```

[7]:
```
true_news['label' ] = 1
fake_news['label' ] = 0
```

[8]:
```
frames = [true_news.loc[:5000][:], fake_news.loc[:5000][:]]
```

[9]:
```
df = pd.concat(frames)
```

[10]:
```
df.head(5)
```

[10]:

	title	text	subject	date	label
0	As U.S. budget fight looms, Republicans flip t…	WASHINGTON (Reuters) - The head of a conservat…	politicsNews	December 31, 2017	1
1	U.S. military to accept transgender recruits o…	WASHINGTON (Reuters) - Transgender people will…	politicsNews	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Muell…	WASHINGTON (Reuters) - The special counsel inv…	politicsNews	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat…	WASHINGTON (Reuters) - Trump campaign adviser …	politicsNews	December 30, 2017	1
4	Trump wants Postal Service to charge 'much mor…	SEATTLE/WASHINGTON (Reuters) - President Donal…	politicsNews	December 29, 2017	1

[11]:
```
df.tail(5)
```

[11]:

	title	text	subject	date	label
4996	Justice Department Announces It Will No Longe…	Republicans are about to lose a huge source of…	News		
4997	WATCH: S.E. Cupp Destroys Trump Adviser's 'Fa…	A pawn working for Donald Trump claimed that w…	News		
4998	WATCH: Fox Hosts Claim Hillary Has Brain Dama…	Fox News is desperate to sabotage Hillary Clin…	News		
4999	CNN Panelist LAUGHS In Corey Lewandowski's Fa…	As Donald Trump s campaign continues to sink d…	News		
5000	Trump Supporter Who Wants To Shoot Black Kids…	Hi folks, John Harper here, at least if you as…	News		

```
4996  August 18, 2016      0
4997  August 18, 2016      0
4998  August 18, 2016      0
4999  August 18, 2016      0
5000  August 18, 2016      0
```

[12]:
```
df.dtypes
```

[12]:
```
title      object
text       object
subject    object
date       object
label       int64
dtype: object
```

[13]:
```
x = df. drop('label', axis=1)
y = df['label']
```

[14]:
```
df = df.dropna()
df_copy = df.copy()
```

[15]:
```
df_copy.reset_index(inplace=True)
df_copy.head()
```

[15]:
```
   index                                              title  \
0      0  As U.S. budget fight looms, Republicans flip t…
1      1  U.S. military to accept transgender recruits o…
2      2  Senior U.S. Republican senator: 'Let Mr. Muell…
3      3  FBI Russia probe helped by Australian diplomat…
4      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  label
0  December 31, 2017       1
1  December 29, 2017       1
2  December 31, 2017       1
3  December 30, 2017       1
4  December 29, 2017       1
```

0.0.1 Data Preprocessing

[16]:
```
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, classification_report,␣
↪confusion_matrix
```

[17]:
```
df_copy['combined_text'] =  df_copy['title'] + ' ' + df_copy['text']
```

[18]:
```
x_train, x_test, y_train, y_test = train_test_split(df_copy['combined_text'],␣
↪df_copy['label'], test_size=0.3, random_state=42)
```

[19]:
```
# TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer is commonly␣
↪used in text classification tasks

tfidf_vectorizer = TfidfVectorizer(stop_words='english', max_features=5000)
x_train_tfidf = tfidf_vectorizer.fit_transform(x_train)
x_test_tfidf = tfidf_vectorizer.transform(x_test)
```

[41]:
```
# print tfidf_vectorizer

#terms = tfidf_vectorizer.get_feature_names_out()
#idf_values = tfidf_vectorizer.idf_

#for term, idf in zip(terms, idf_values):
    #print(f'Term: {term}, IDF: {idf}')
```

0.0.2 Model Building - Passive Aggresive Classifier

0.0.3 Using Naives Bayes

[20]:
```
from sklearn.naive_bayes import MultinomialNB
```

[21]:
```
model = MultinomialNB()
model.fit(x_train_tfidf, y_train)
```

[21]:
```
MultinomialNB()
```

[22]:
```
predictions = model.predict(x_test_tfidf)
```

[43]:
```
print("Accuracy:", accuracy_score(y_test, predictions))
print("\nConfusion Matrix:\n", confusion_matrix(y_test, predictions))
```

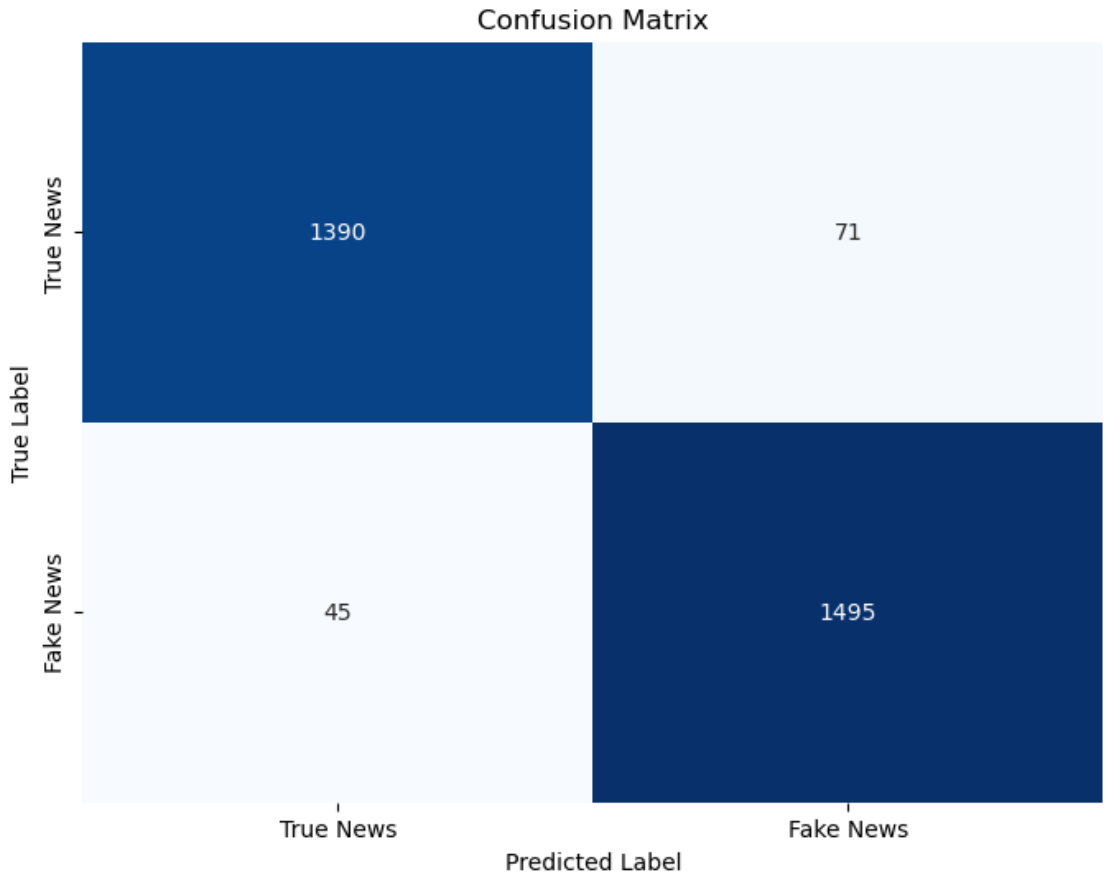
```
Accuracy: 0.9613462179273575

Confusion Matrix:
```

```
[[1390   71]
[  45 1495]]
```

Visualizations

[24]:
```
# pip install matplotlib seaborn
```

[25]:
```
import matplotlib.pyplot as plt
import seaborn as sns

conf_matrix = confusion_matrix(y_test, predictions)

# Plot confusion matrix as a heatmap

plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues', cbar=False,
            xticklabels=['True News', 'Fake News'], yticklabels=['True News',␣
 ↪'Fake News'])
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.title('Confusion Matrix')
plt.show()
```



0.0.4 Using SVM

[26]:
```
from sklearn.svm import SVC
```

[27]:
```
svm_model = SVC(kernel='linear', C=1.0)
svm_model.fit(x_train_tfidf, y_train)
```

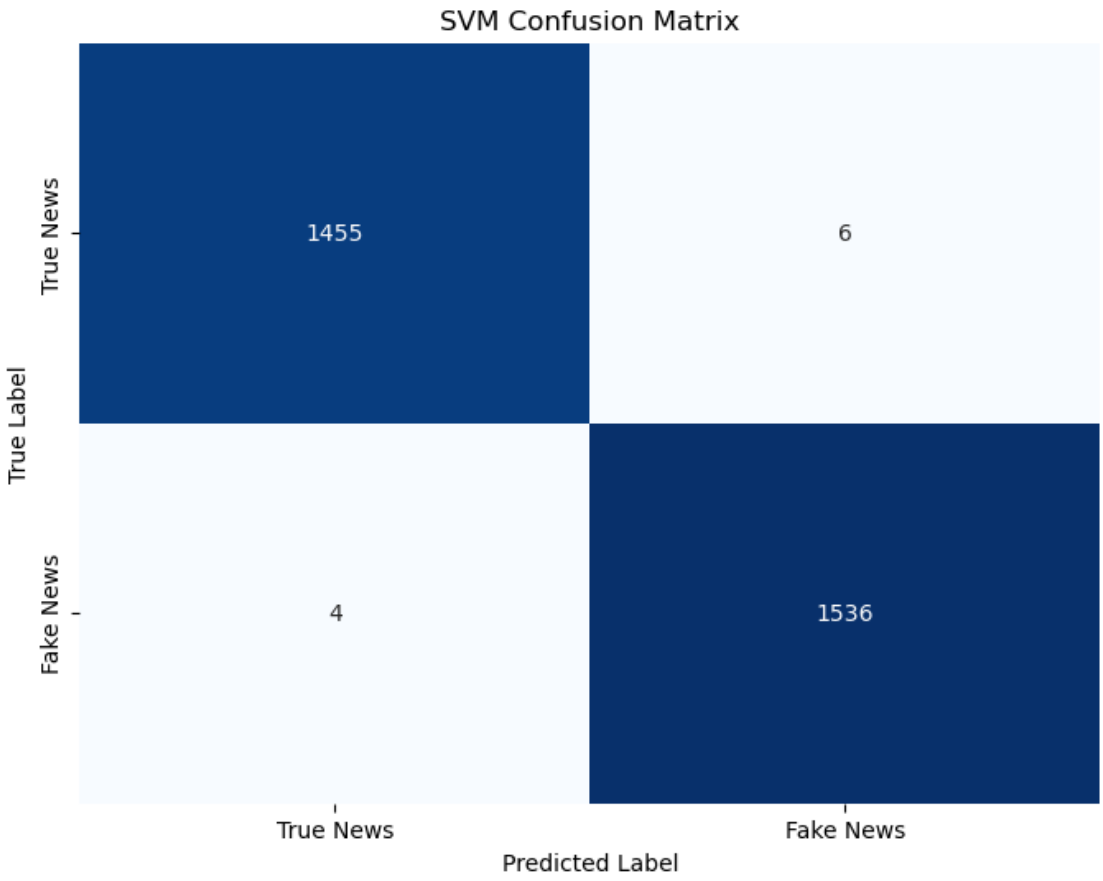
[27]:
```
SVC(kernel='linear')
```

[28]:
```
# making predictions

svm_predictions = svm_model.predict(x_test_tfidf)
```

[42]:
```
print("Accuracy:", accuracy_score(y_test, svm_predictions))
print("\nConfusion Matrix:\n", confusion_matrix(y_test, svm_predictions))
```

```
Accuracy: 0.9966677774075309

Confusion Matrix:
 [[1455    6]
 [   4 1536]]
```

[30]:
```
# Plot confusion matrix as a heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(confusion_matrix(y_test, svm_predictions), annot=True, fmt='d',␣
 ↪cmap='Blues', cbar=False,
           xticklabels=['True News', 'Fake News'], yticklabels=['True News',␣
 ↪'Fake News'])
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.title('SVM Confusion Matrix')
plt.show()
```



0.0.5 Using Random Forest Regression

[31]:
```
from sklearn.ensemble import RandomForestClassifier
```

[33]:
```
# Train the model
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(x_train_tfidf, y_train)
```

[33]:
```
RandomForestClassifier(random_state=42)
```

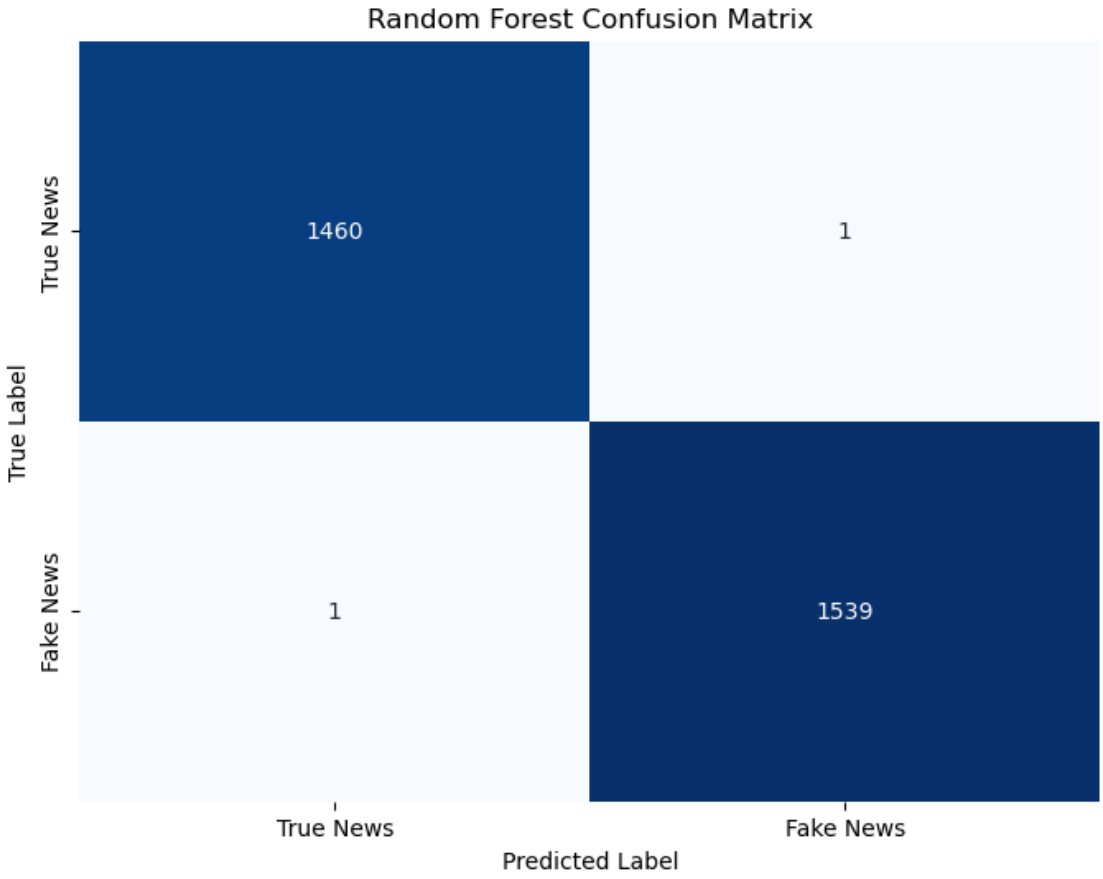
[34]:
```
rf_predictions = rf_model.predict(x_test_tfidf)
```

[35]:
```
print("Accuracy:", accuracy_score(y_test, rf_predictions))
print("\nConfusion Matrix:\n", confusion_matrix(y_test, rf_predictions))
```

```
Accuracy: 0.9993335554815062

Confusion Matrix:
 [[1460    1]
```

```
 [   1 1539]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1461
           1       1.00      1.00      1.00      1540

    accuracy                           1.00      3001
   macro avg       1.00      1.00      1.00      3001
weighted avg       1.00      1.00      1.00      3001
```

[36]:
```
# Plot confusion matrix as a heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(confusion_matrix(y_test, rf_predictions), annot=True, fmt='d',␣
↪cmap='Blues', cbar=False,
            xticklabels=['True News', 'Fake News'], yticklabels=['True News',␣
↪'Fake News'])
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.title('Random Forest Confusion Matrix')
plt.show()
```



Save model and Vectorize

Random Forest Regression has high Accuracy compared to naives bayes and Support vector machine

[37]:
```
import pickle

# Save the model
with open('fake_news_dectector_rf_model.pkl', 'wb') as model_file:
    pickle.dump(model, model_file)

# Save the TF-IDF vectorizer
with open('fake_news_dectector_rf_model_vectorizer.pkl', 'wb') as␣
 ↪vectorizer_file:
    pickle.dump(tfidf_vectorizer, vectorizer_file)
```

[38]:
```
# Can use the below code to load the model for future predictions

# Load the model
with open('fake_news_dectector_rf_model.pkl', 'rb') as model_file:
    loaded_model = pickle.load(model_file)
```

```
# Load the TF-IDF vectorizer
with open('fake_news_dectector_rf_model_vectorizer.pkl', 'rb') as␣
 ↪vectorizer_file:
    loaded_vectorizer = pickle.load(vectorizer_file)
```

Camparing pickle and joblib joblib has better performance specially when using in HDFS but pickle is the convinient way to save and load models

[ ]: