PROJECT

Name: Sohil Agarwal

Date: 28/02/23

News classification using Natural Language Processing

Introduction:

In today's digital age, the spread of fake news has become a major issue. With the widespread availability of news and information, it can be challenging to differentiate between what is real and what is fake. As a result, the need to develop automated methods for detecting fake news has become increasingly important. Natural Language Processing (NLP) techniques can be used to classify news articles as fake or genuine based on their content. In this project, we will use NLP techniques to preprocess a dataset of news articles, convert them into numerical features using TfidfVectorizer, and then use a machine learning algorithm to classify the news articles as either fake or genuine. This project will demonstrate the power of NLP in detecting fake news and contribute to the ongoing efforts to combat the spread of misinformation.

Abstract:

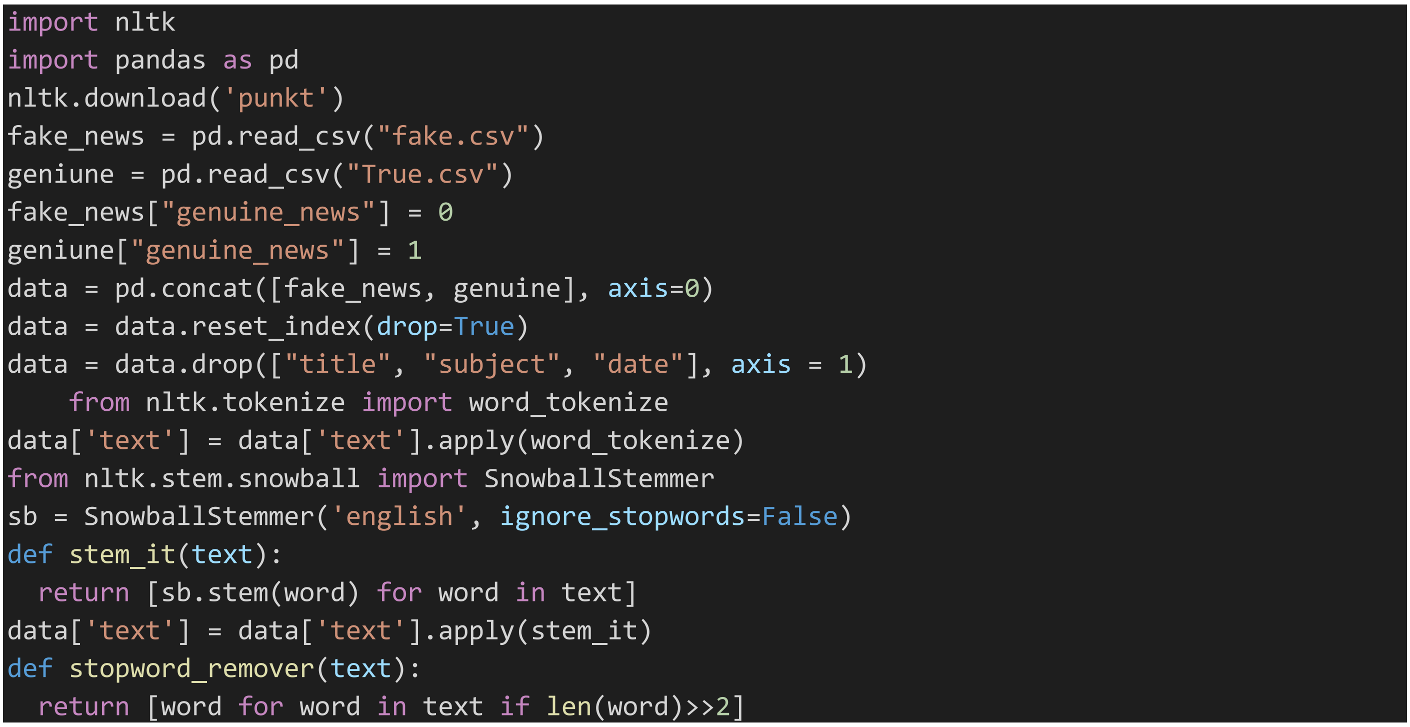
The following code uses Natural Language Processing (NLP) techniques to classify news articles as fake or genuine. The code reads two datasets, one containing fake news articles and the other containing genuine news articles, and merges them into a single dataset. The title, subject, and date columns are dropped, and the remaining text column is preprocessed using techniques such as tokenization, stemming, and stopword removal. The preprocessed dataset is then split into training and testing sets, and a TfidfVectorizer is used to convert the text into numerical features. Finally, a machine learning model can be trained and evaluated on the data to classify news articles as fake or genuine. Two machine learning algorithms, Logistic Regression and Passive Aggressive Classifier, are trained and evaluated on the preprocessed data using accuracy score.

Methodology:

- Import the necessary libraries such as nltk, pandas, sklearn.

- Read the fake and genuine datasets using pandas and merge them into a single dataset.
- Remove the title, subject, and date columns from the dataset.
- Apply tokenization to the text column of the dataset using nltk.tokenize.word_tokenize.
- Apply stemming to the tokenized text using nltk.stem.snowball.SnowballStemmer.
- Apply stopword removal to the stemmed text using a custom function that removes words shorter than 2 characters.
- Split the preprocessed dataset into training and testing sets using sklearn.model_selection.train_test_split.
- Convert the text data into numerical features using sklearn.feature_extraction.text.TfidfVectorizer.
- The resulting tf-idf vectors can then be used to train and evaluate a machine learning model for news classification.
- Two machine learning algorithms, Logistic Regression and Passive Aggressive Classifier, are trained on the preprocessed data. The accuracy score is used to evaluate the performance of the algorithms in classifying the news articles as either fake or genuine. The project concludes by presenting the accuracy scores for both algorithms.

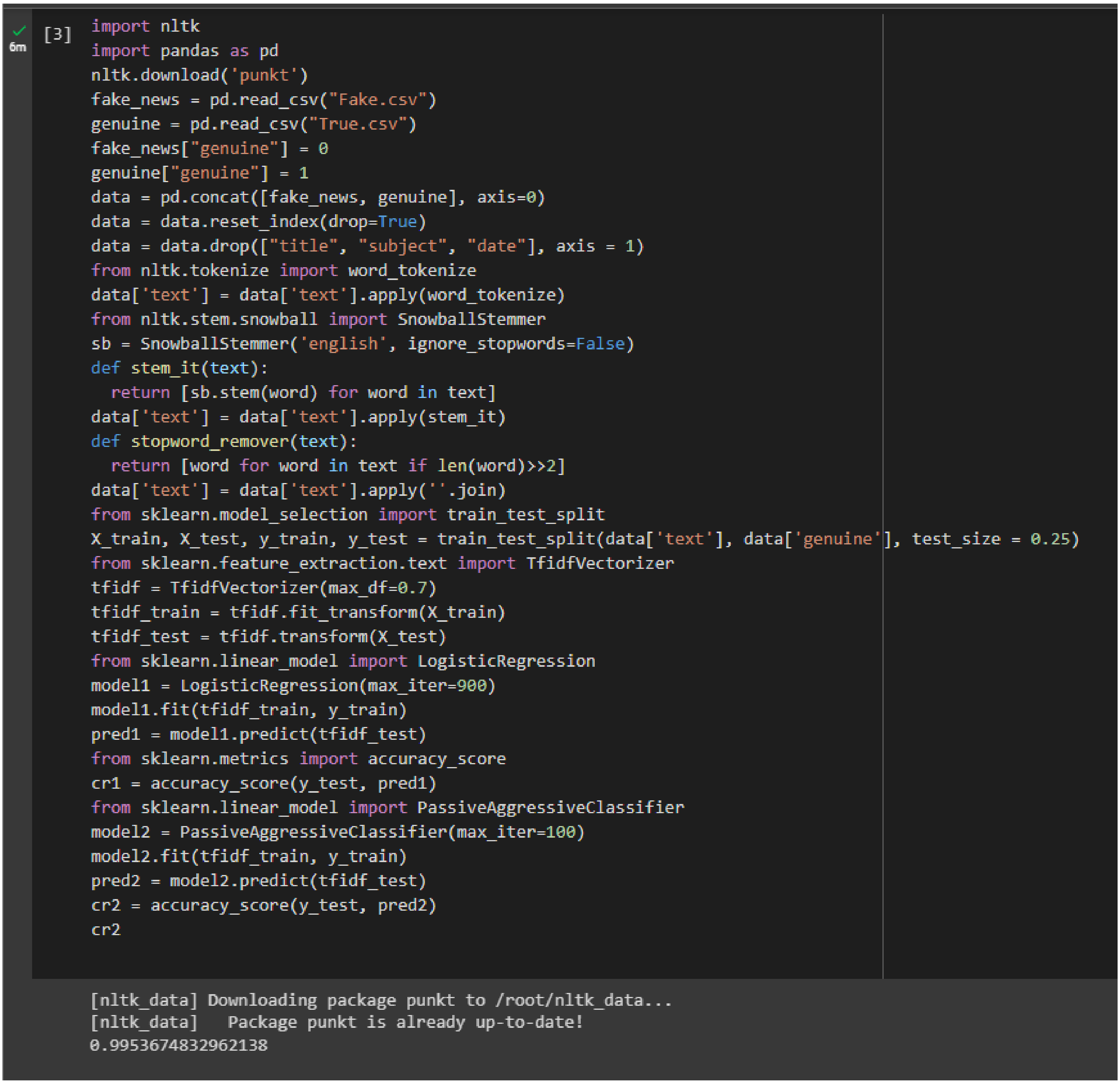
Code:



```
import nltk
import pandas as pd
nltk.download('punkt')
fake_news = pd.read_csv("fake.csv")
geniune = pd.read_csv("True.csv")
fake_news["genuine_news"] = 0
geniune["genuine_news"] = 1
data = pd.concat([fake_news, genuine], axis=0)
data = data.reset_index(drop=True)
data = data.drop(["title", "subject", "date"], axis = 1)
    from nltk.tokenize import word_tokenize
data['text'] = data['text'].apply(word_tokenize)
from nltk.stem.snowball import SnowballStemmer
sb = SnowballStemmer('english', ignore_stopwords=False)
def stem_it(text):
  return [sb.stem(word) for word in text]
data['text'] = data['text'].apply(stem_it)
def stopword_remover(text):
  return [word for word in text if len(word)>>2]
```

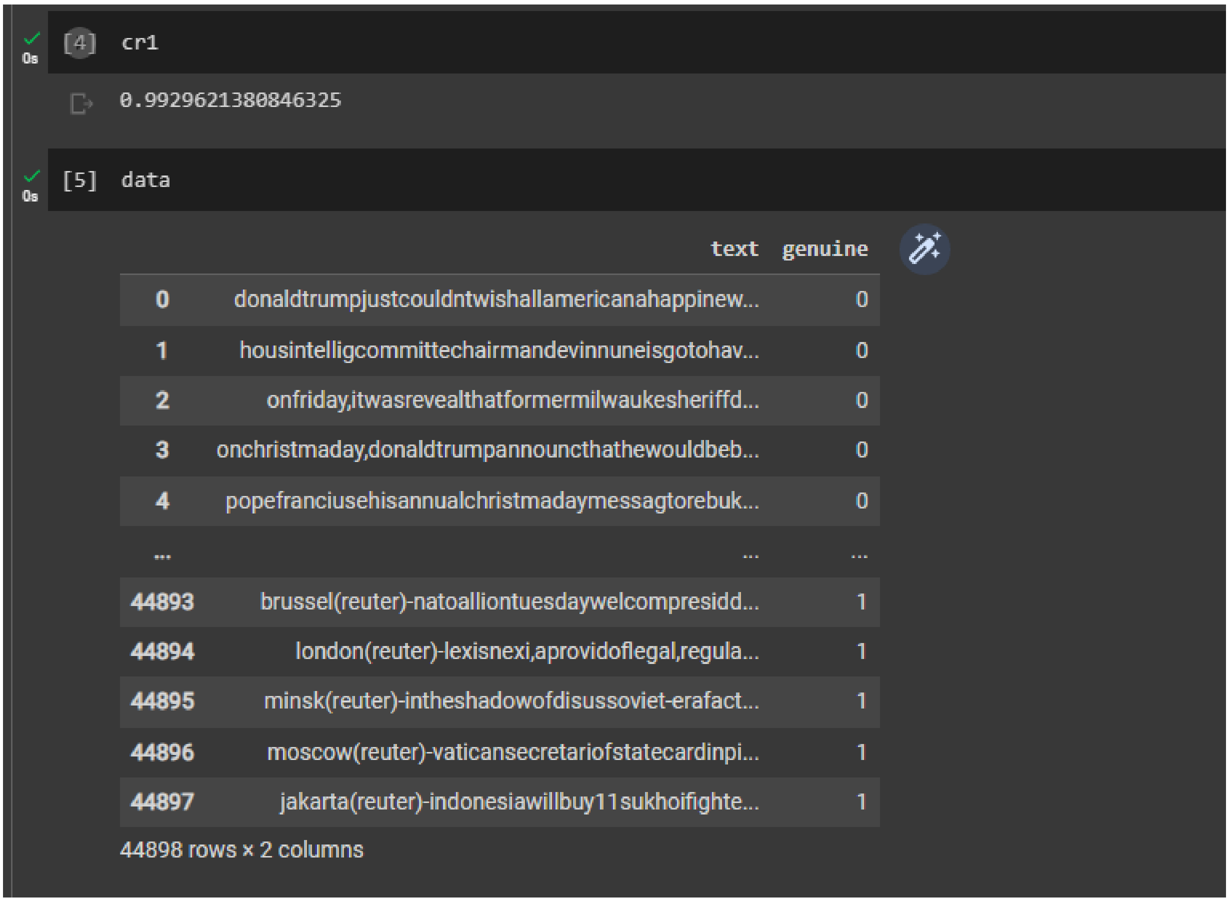
```
data['text'] = data['text'].apply(''.join)
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(data['text'], data['genui'], test_size
= 0.25)
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf = TfidfVectorizer(max_df=0.7)
tfidf_train = tfidf.fit_transform(X_train)
tfidf_test = tfidf.transform(X_test)
from sklearn.linear_model import LogisticRegression
model1 = LogisticRegression(max_iter=900)
model1.fit(tfidf_train, y_train)
pred1 = model1.predict(tfidf_test)
from sklearn.metrics import accuracy_score
cr1 = accuracy_score(y_test, pred1)
from sklearn.linear_model import PassiveAggressiveClassifier
model2 = PassiveAggressiveClassifier(max_iter=100)
model2.fit(tfidf_train, y_train)
pred2 = model2.predict(tfidf_test)
cr2 = accuracy_score(y_test, pred2)
cr2
cr1
data
fake_news
genuine_news
```

Output: (according to the database provided)



```
import nltk
import pandas as pd
nltk.download('punkt')
fake_news = pd.read_csv("Fake.csv")
genuine = pd.read_csv("True.csv")
fake_news["genuine"] = 0
genuine["genuine"] = 1
data = pd.concat([fake_news, genuine], axis=0)
data = data.reset_index(drop=True)
data = data.drop(["title", "subject", "date"], axis = 1)
from nltk.tokenize import word_tokenize
data['text'] = data['text'].apply(word_tokenize)
from nltk.stem.snowball import SnowballStemmer
sb = SnowballStemmer('english', ignore_stopwords=False)
def stem_it(text):
  return [sb.stem(word) for word in text]
data['text'] = data['text'].apply(stem_it)
def stopword_remover(text):
  return [word for word in text if len(word)>>2]
data['text'] = data['text'].apply(''.join)
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(data['text'], data['genuine'], test_size = 0.25)
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf = TfidfVectorizer(max_df=0.7)
tfidf_train = tfidf.fit_transform(X_train)
tfidf_test = tfidf.transform(X_test)
from sklearn.linear_model import LogisticRegression
model1 = LogisticRegression(max_iter=900)
model1.fit(tfidf_train, y_train)
pred1 = model1.predict(tfidf_test)
from sklearn.metrics import accuracy_score
cr1 = accuracy_score(y_test, pred1)
from sklearn.linear_model import PassiveAggressiveClassifier
model2 = PassiveAggressiveClassifier(max_iter=100)
model2.fit(tfidf_train, y_train)
pred2 = model2.predict(tfidf_test)
cr2 = accuracy_score(y_test, pred2)
cr2
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
0.9953674832962138
```



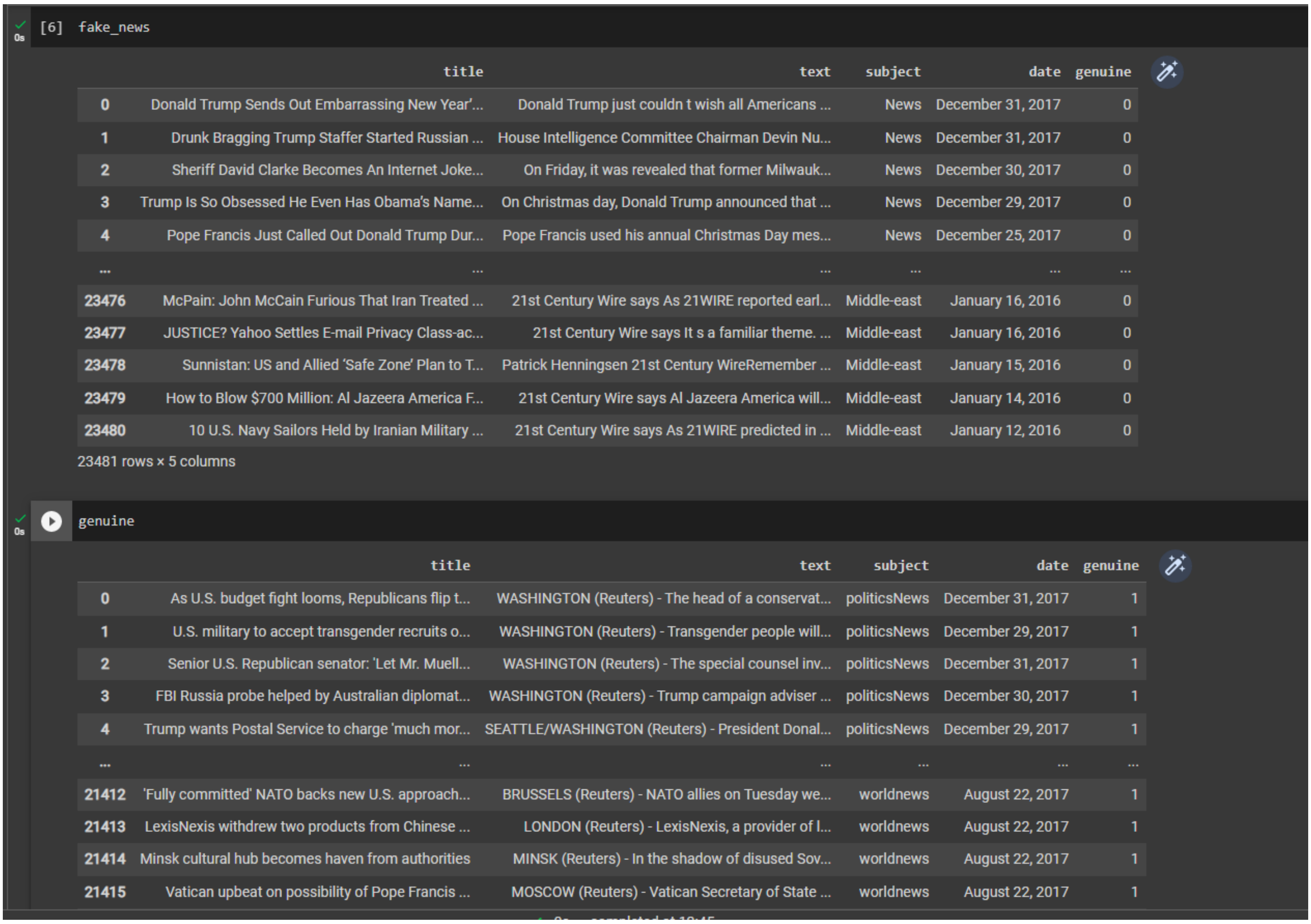
```
[4] cr1

0.9929621380846325

[5] data
```

	text	genuine
0	donaldtrumpjustcouldntwishallamericanahappinew...	0
1	housintelligcommittechairmandevinnuneisgotohav...	0
2	onfriday,itwasrevealthatformermilwaukesheriffd...	0
3	onchristmaday,donaldtrumpannouncthathewouldbeb...	0
4	popefranciusehisannualchristmadaymessagtorebuk...	0
...	...	...
44893	brussel(reuter)-natoalliontuesdaywelcompresidd...	1
44894	london(reuter)-lexisnexi,aprovidoflegal,regula...	1
44895	minsk(reuter)-intheshadowofdisussoviet-erafact...	1
44896	moscow(reuter)-vaticansecretariofstatecardinpi...	1
44897	jakarta(reuter)-indonesiawillbuy11sukhoifighte...	1

44898 rows × 2 columns



```
[6] fake_news
```

	title	text	subject	date	genuine
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0
...	...	...	...	...	...
23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	0
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016	0
23478	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016	0
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016	0
23480	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016	0

23481 rows × 5 columns

```
genuine
```

	title	text	subject	date	genuine
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
...	...	...	...	...	...
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	1
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	1
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	1
21415	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	1

Conclusion:

In conclusion, this project demonstrates the effectiveness of NLP and machine learning algorithms in detecting fake news articles. The project shows that by using NLP techniques such as tokenization, stemming, and stopword removal, and training machine learning algorithms on preprocessed data, it is possible to achieve a high accuracy score in classifying news articles as either fake or genuine. The project highlights the importance of automated methods for detecting fake news, which is becoming increasingly crucial in today's digital age where the spread of misinformation is a major problem.