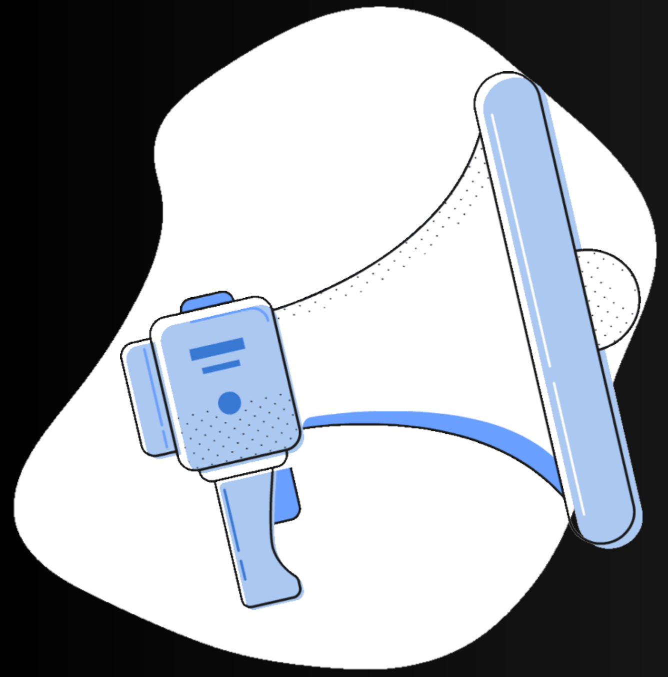
NewsVerifier

Data Mining Project

Group 6

Patel Hetvi
Vinay Mehra
Achini Nanayakkara

Introduction



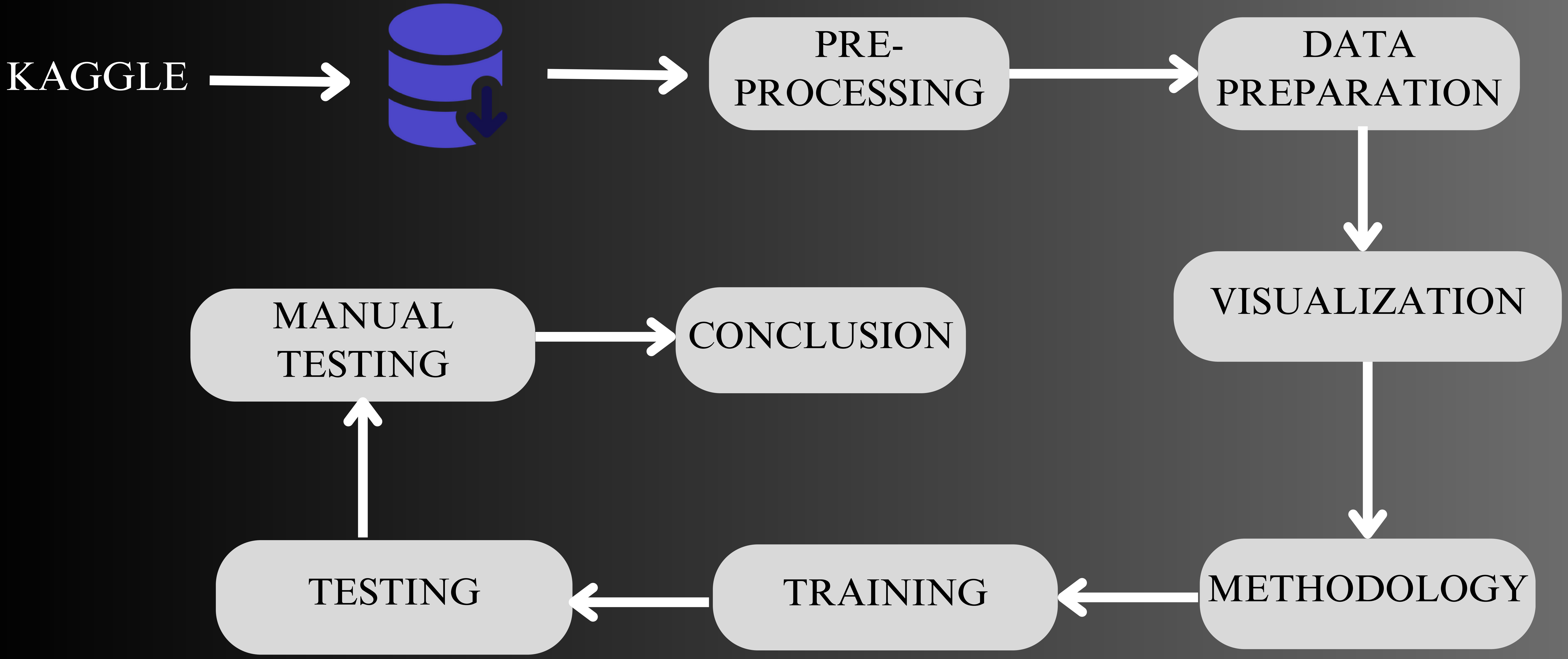
Why do we need this technology ?

- The internet has become essential in our lives, making news more accessible, especially for young people who prefer it over traditional media.
- In world, where internet access has recently become affordable, many people get their news online. However, this rapid spread of information can lead to issues with fake news, which can harm individuals or groups for political, religious, or other purposes.
- Research by the BBC on the 2014 Indian general election found significant polarization of fake news on social media, with 72% of Indians unable to distinguish real news from fake.
- This project aims to help citizens identify and expose fake news, promoting digital literacy to mitigate its harmful effects.

PROBLEM DEFINITION

How can we effectively identify fake news with the attributes of a news article?

WORK FLOW



KAGGLE

PRE-PROCESSING

DATA PREPARATION

VISUALIZATION

METHODOLOGY

TRAINING

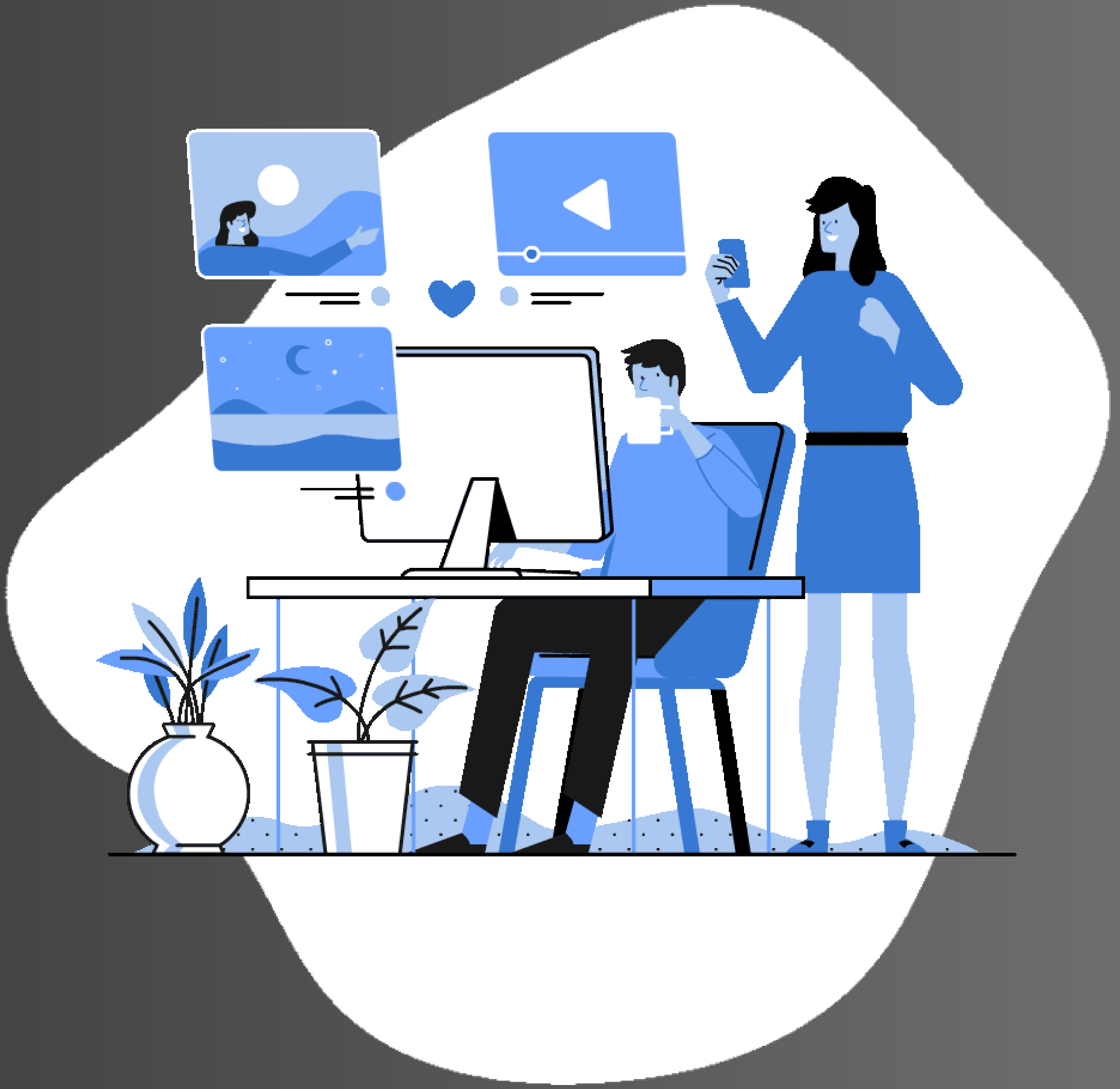
TESTING

MANUAL TESTING

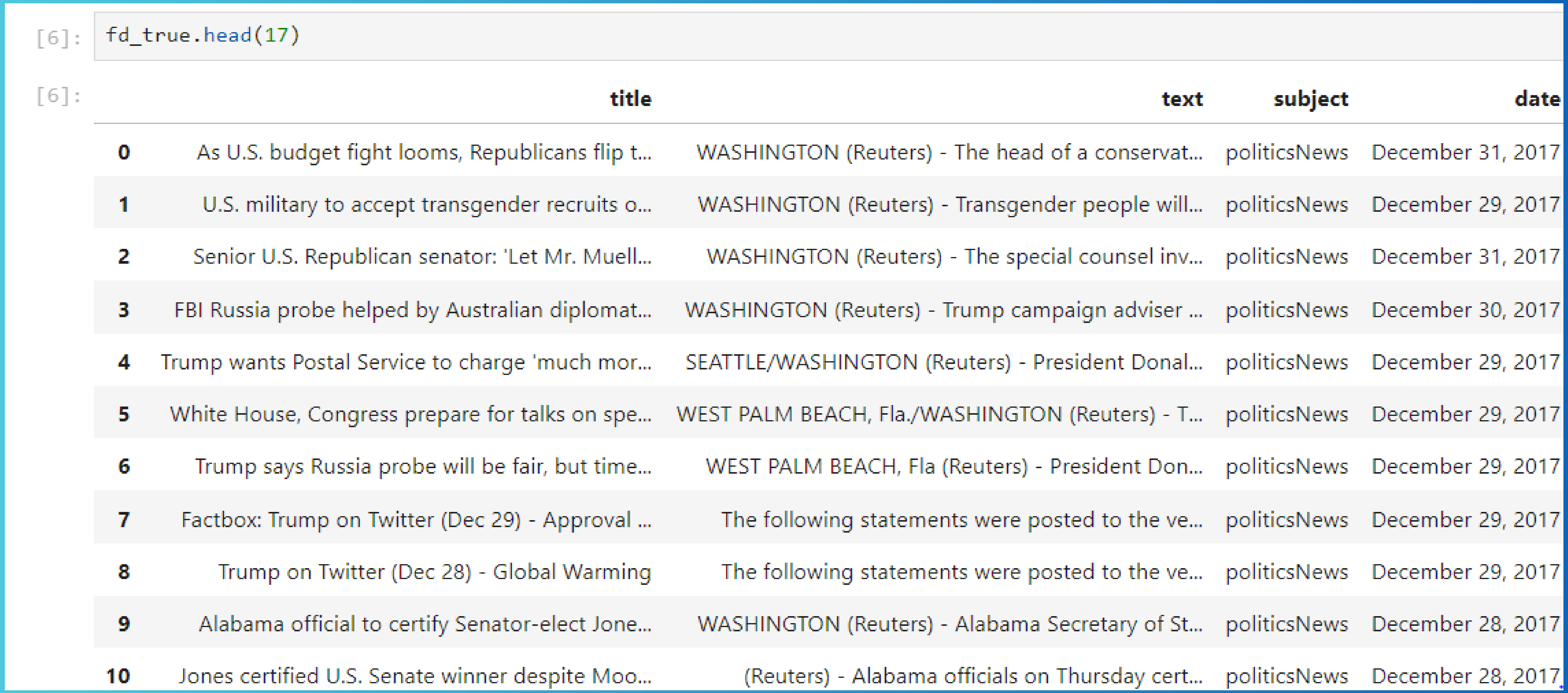
CONCLUSION

DATA SET

- The dataset was obtained from kaggle. data set. (ISOT fake news detection Dataset)
- The dataset consists of two separate CSV files
  --> One containing fake news articles
- Attributes: Title, Text, Subject, Date
- Sample Size: 23,481 articles
  --> Second containing true news articles
- Attributes: Title, Text, Subject, Date
- Sample Size: 21,417 articles
- Each dataset includes attributes such as the title, text, subject, and date of the news articles.



DATA IMPORT



[6]:
```
fd_true.head(17)
```

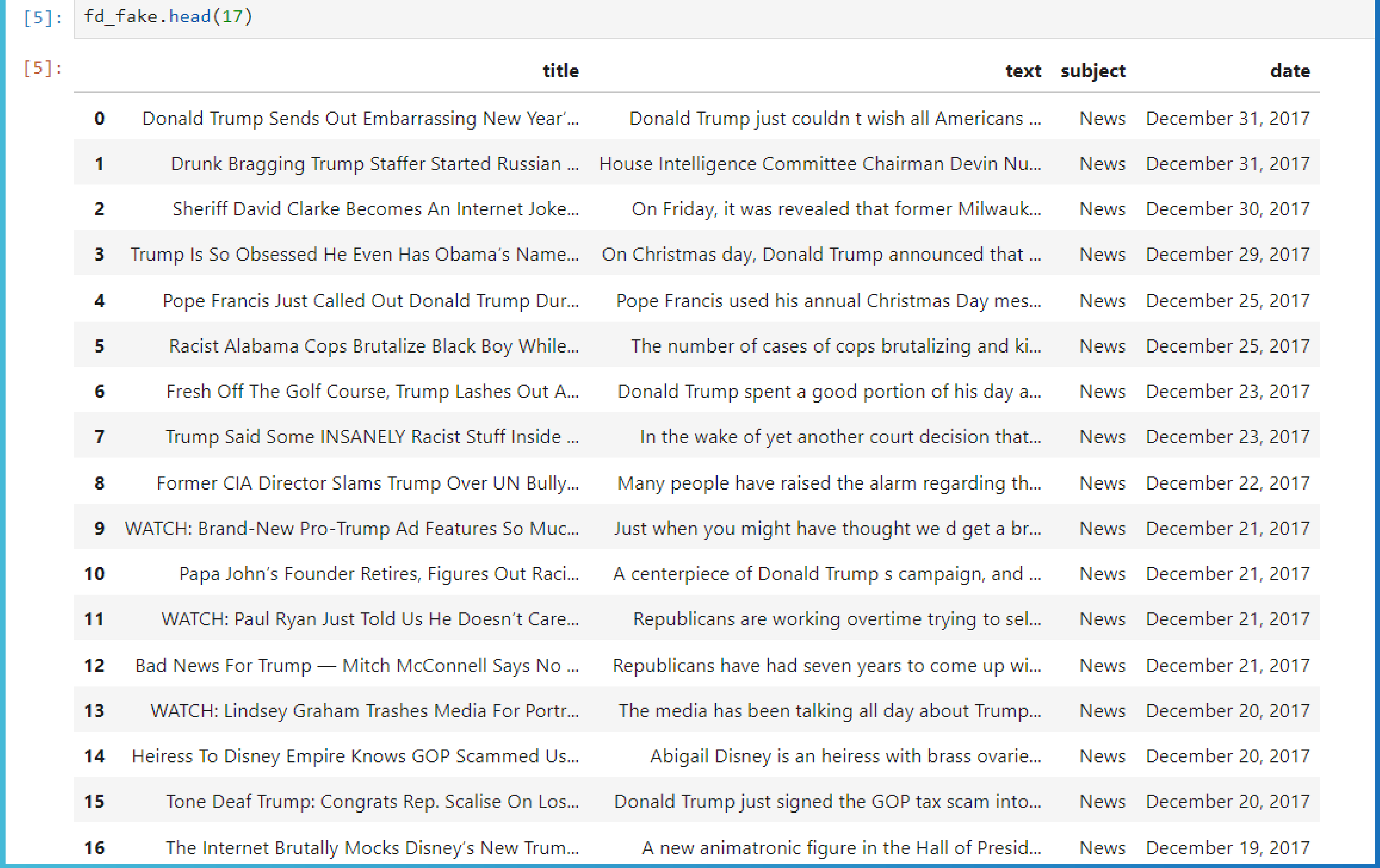
[6]:

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017
5	White House, Congress prepare for talks on spe...	WEST PALM BEACH, Fla./WASHINGTON (Reuters) - T...	politicsNews	December 29, 2017
6	Trump says Russia probe will be fair, but time...	WEST PALM BEACH, Fla (Reuters) - President Don...	politicsNews	December 29, 2017
7	Factbox: Trump on Twitter (Dec 29) - Approval ...	The following statements were posted to the ve...	politicsNews	December 29, 2017
8	Trump on Twitter (Dec 28) - Global Warming	The following statements were posted to the ve...	politicsNews	December 29, 2017
9	Alabama official to certify Senator-elect Jone...	WASHINGTON (Reuters) - Alabama Secretary of St...	politicsNews	December 28, 2017
10	Jones certified U.S. Senate winner despite Moo...	(Reuters) - Alabama officials on Thursday cert...	politicsNews	December 28, 2017

True.csv file

- **Attributes**: Title, Text, Subject, Date
- **Sample Size**: 21,417 articles

DATA IMPORT



```
[5]: fd_fake.head(17)
```

[5]:

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017
5	Racist Alabama Cops Brutalize Black Boy While...	The number of cases of cops brutalizing and ki...	News	December 25, 2017
6	Fresh Off The Golf Course, Trump Lashes Out A...	Donald Trump spent a good portion of his day a...	News	December 23, 2017
7	Trump Said Some INSANELY Racist Stuff Inside ...	In the wake of yet another court decision that...	News	December 23, 2017
8	Former CIA Director Slams Trump Over UN Bully...	Many people have raised the alarm regarding th...	News	December 22, 2017
9	WATCH: Brand-New Pro-Trump Ad Features So Muc...	Just when you might have thought we d get a br...	News	December 21, 2017
10	Papa John's Founder Retires, Figures Out Raci...	A centerpiece of Donald Trump s campaign, and ...	News	December 21, 2017
11	WATCH: Paul Ryan Just Told Us He Doesn't Care...	Republicans are working overtime trying to sel...	News	December 21, 2017
12	Bad News For Trump — Mitch McConnell Says No ...	Republicans have had seven years to come up wi...	News	December 21, 2017
13	WATCH: Lindsey Graham Trashes Media For Portr...	The media has been talking all day about Trump...	News	December 20, 2017
14	Heiress To Disney Empire Knows GOP Scammed Us...	Abigail Disney is an heiress with brass ovarie...	News	December 20, 2017
15	Tone Deaf Trump: Congrats Rep. Scalise On Los...	Donald Trump just signed the GOP tax scam into...	News	December 20, 2017
16	The Internet Brutally Mocks Disney's New Trum...	A new animatronic figure in the Hall of Presid...	News	December 19, 2017

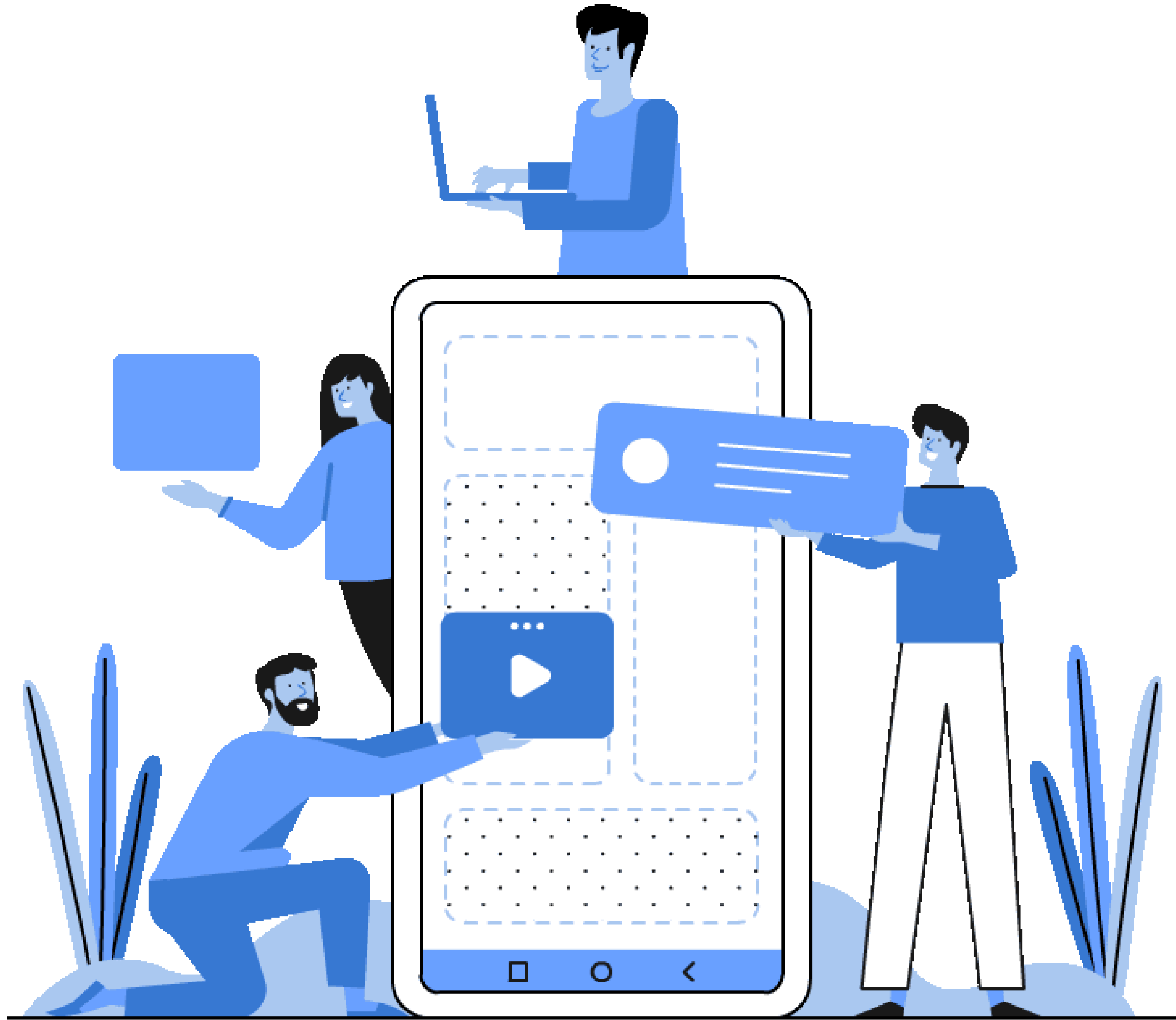
Fake.csv file

- Attributes: Title, Text, Subject, Date
- Sample Size: 23,481 articles

DATA PREPROCESSING

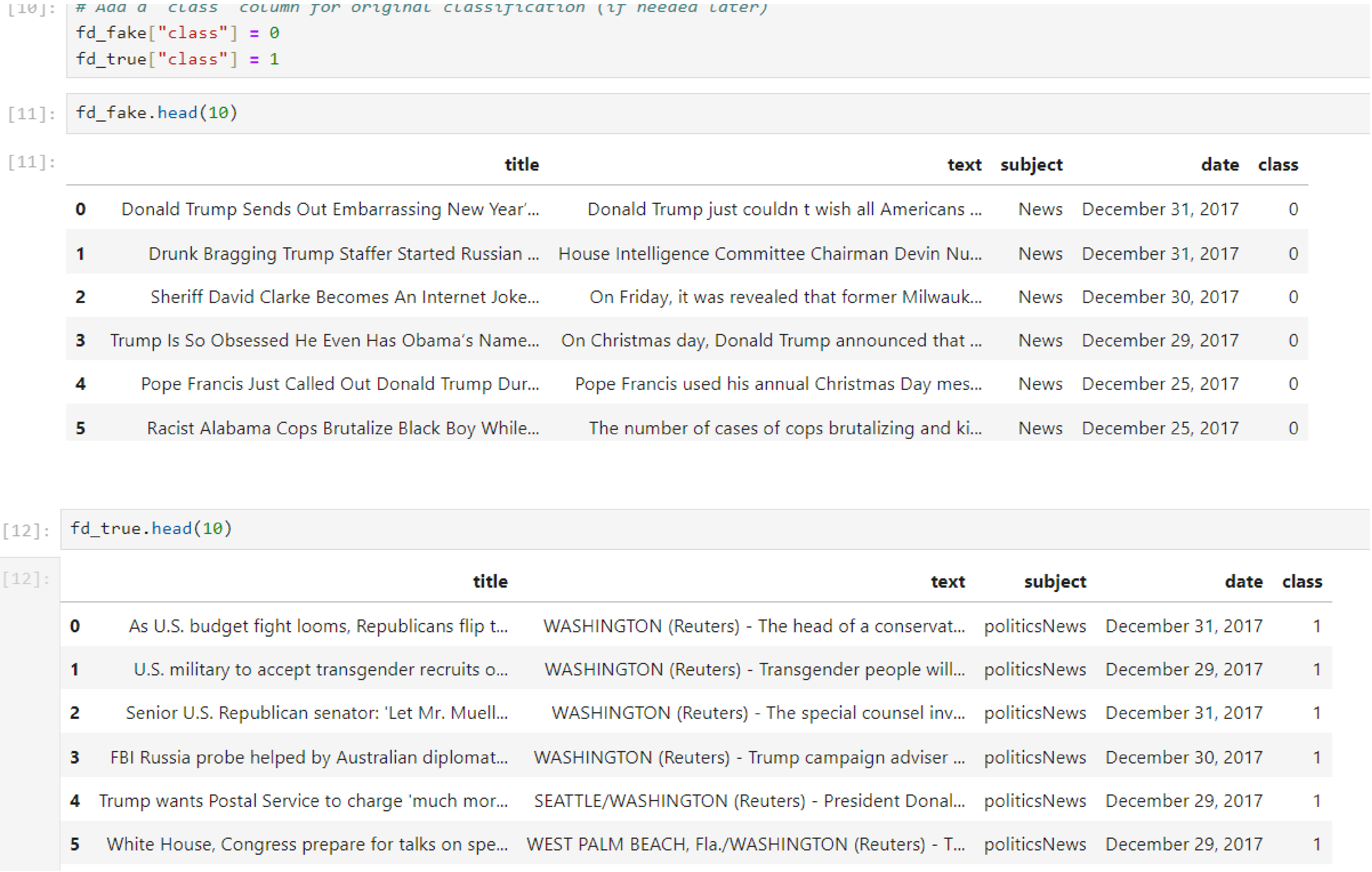
Perform various text cleaning steps :

- Merging Datasets
- Converting the string into lower case
- Removing hyper links from news
- Removing special characters from news
- Remove stopwords from news



Label data

Combine the datasets with an additional column Class Label where 1 represents true news and 0 represents fake news.



[10]:
```python
# Add a "class" column for original classification (if needed later)
fd_fake["class"] = 0
fd_true["class"] = 1
```

[11]:
```python
fd_fake.head(10)
```

[11]:

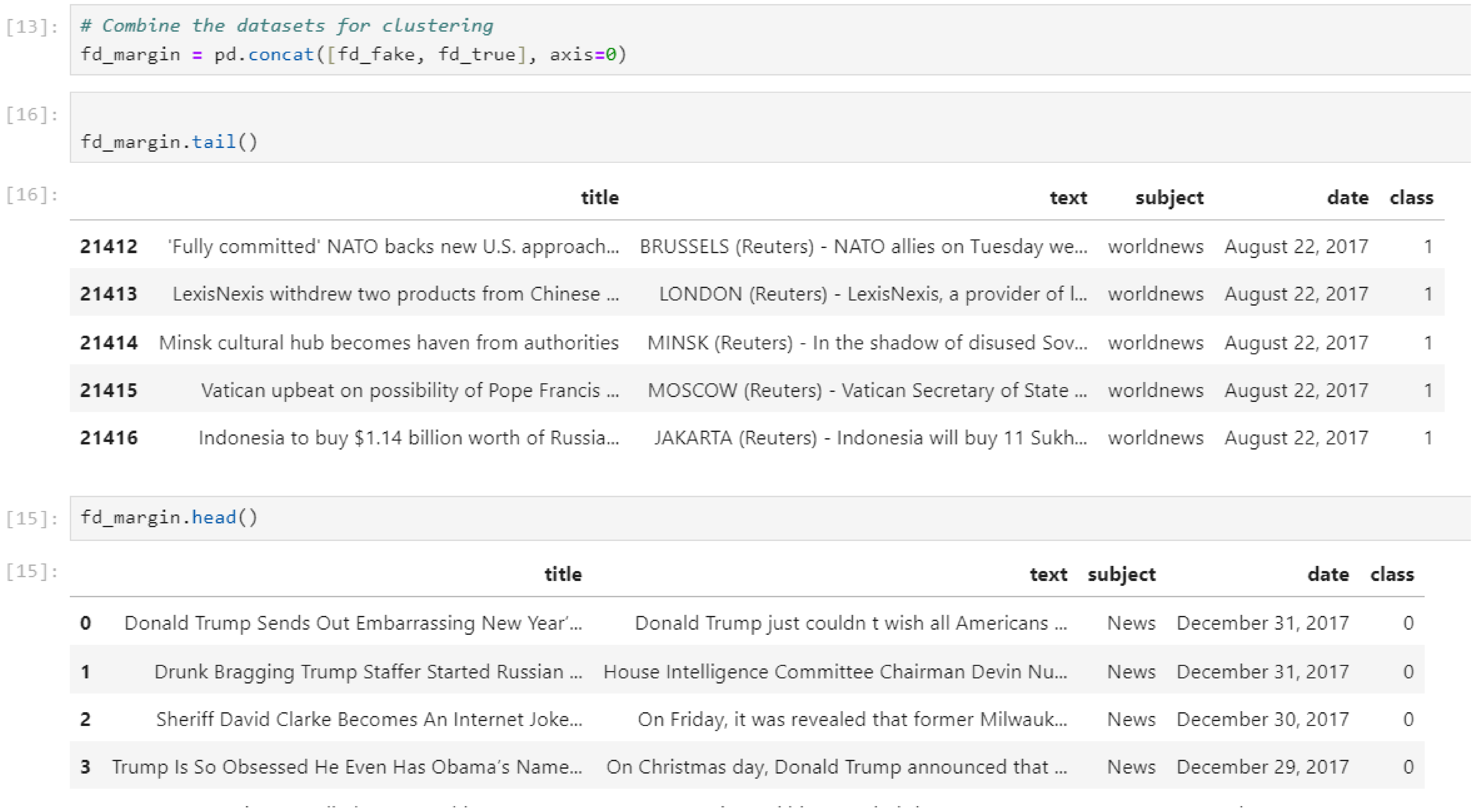
	title	text	subject	date	class
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke 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
5	Racist Alabama Cops Brutalize Black Boy While...	The number of cases of cops brutalizing and ki...	News	December 25, 2017	0

[12]:
```python
fd_true.head(10)
```

[12]:

	title	text	subject	date	class
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
5	White House, Congress prepare for talks on spe...	WEST PALM BEACH, Fla./WASHINGTON (Reuters) - T...	politicsNews	December 29, 2017	1

merged data



[13]:

```
# Combine the datasets for clustering
fd_margin = pd.concat([fd_fake, fd_true], axis=0)
```

[16]:

```

fd_margin.tail()
```

[16]:

	title	text	subject	date	class
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
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	1

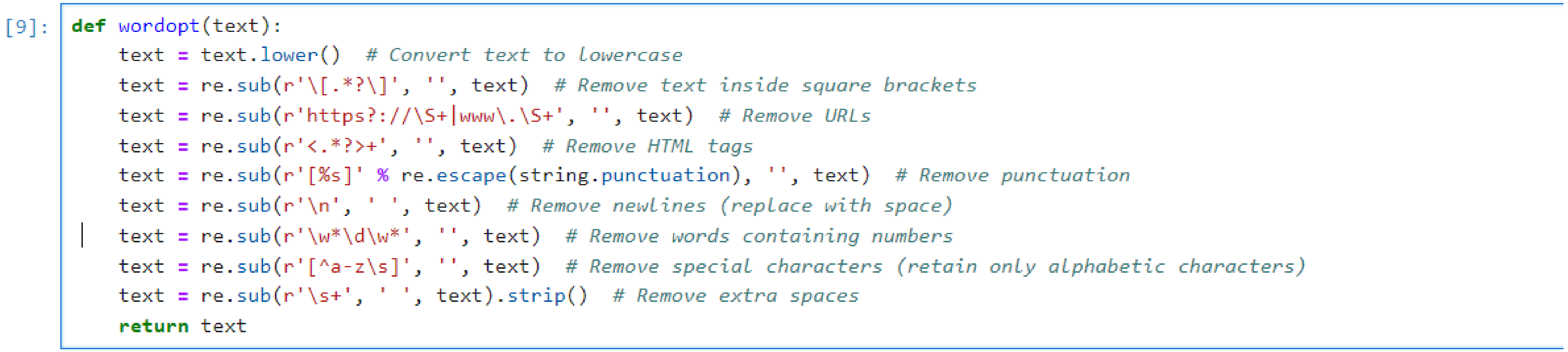
[15]:

```
fd_margin.head()
```

[15]:

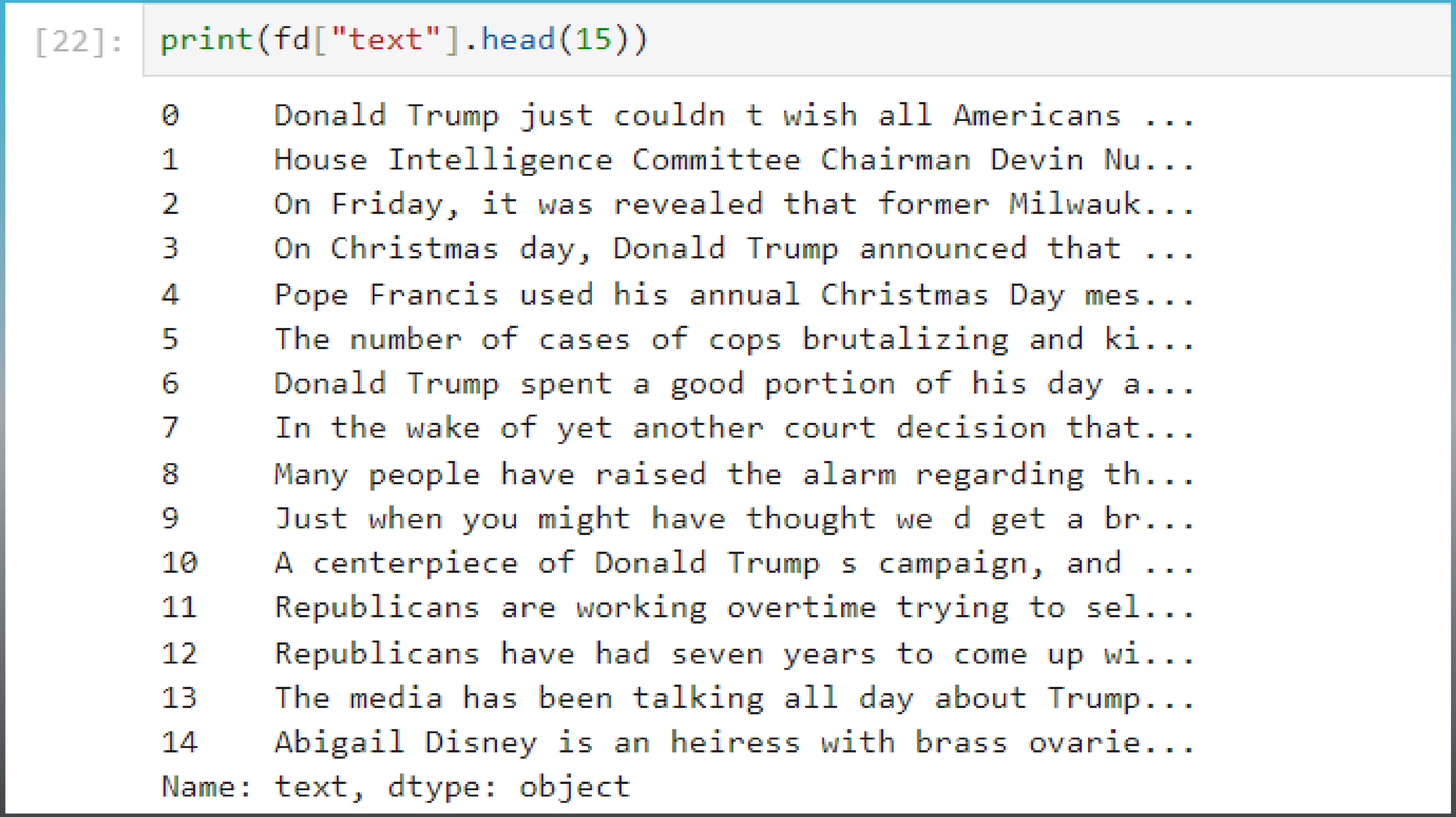
	title	text	subject	date	class
0	Donald Trump Sends Out Embarrassing New Year’...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke 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

DATA PROCESSING



```
[9]: def wordopt(text):
         text = text.lower()  # Convert text to lowercase
         text = re.sub(r'\[.*?\]', '', text)  # Remove text inside square brackets
         text = re.sub(r'https?://\S+|www\.\S+', '', text)  # Remove URLs
         text = re.sub(r'<.*?>+', '', text)  # Remove HTML tags
         text = re.sub(r'[%s]' % re.escape(string.punctuation), '', text)  # Remove punctuation
         text = re.sub(r'\n', ' ', text)  # Remove newlines (replace with space)
         text = re.sub(r'\w*\d\w*', '', text)  # Remove words containing numbers
         text = re.sub(r'[^a-z\s]', '', text)  # Remove special characters (retain only alphabetic characters)
         text = re.sub(r'\s+', ' ', text).strip()  # Remove extra spaces
         return text
```

RESULT



```
[22]: print(fd["text"].head(15))
```

```
0     Donald Trump just couldn t wish all Americans ...
1     House Intelligence Committee Chairman Devin Nu...
2     On Friday, it was revealed that former Milwauk...
3     On Christmas day, Donald Trump announced that ...
4     Pope Francis used his annual Christmas Day mes...
5     The number of cases of cops brutalizing and ki...
6     Donald Trump spent a good portion of his day a...
7     In the wake of yet another court decision that...
8     Many people have raised the alarm regarding th...
9     Just when you might have thought we d get a br...
10    A centerpiece of Donald Trump s campaign, and ...
11    Republicans are working overtime trying to sel...
12    Republicans have had seven years to come up wi...
13    The media has been talking all day about Trump...
14    Abigail Disney is an heiress with brass ovarie...
Name: text, dtype: object
```

IDENTIFICATION OF ISSUES

Missing Data :

- Analysis: Checked for missing values in all columns.
- Finding: No missing values were found in the dataset.

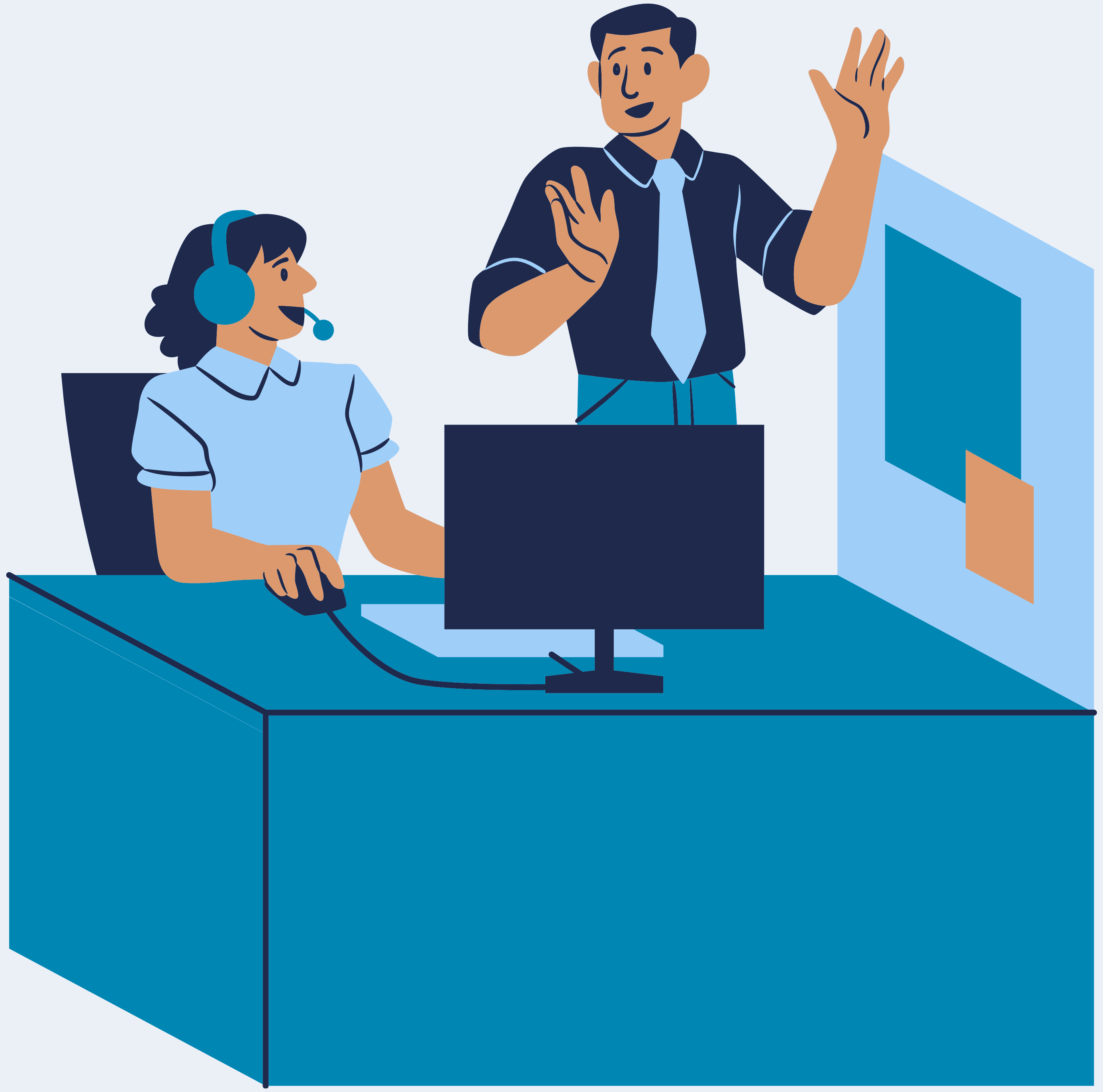
Duplicates :

- Analysis: Checked for duplicate rows based on the text content.
- Finding: 1,234 duplicate articles were found and removed.

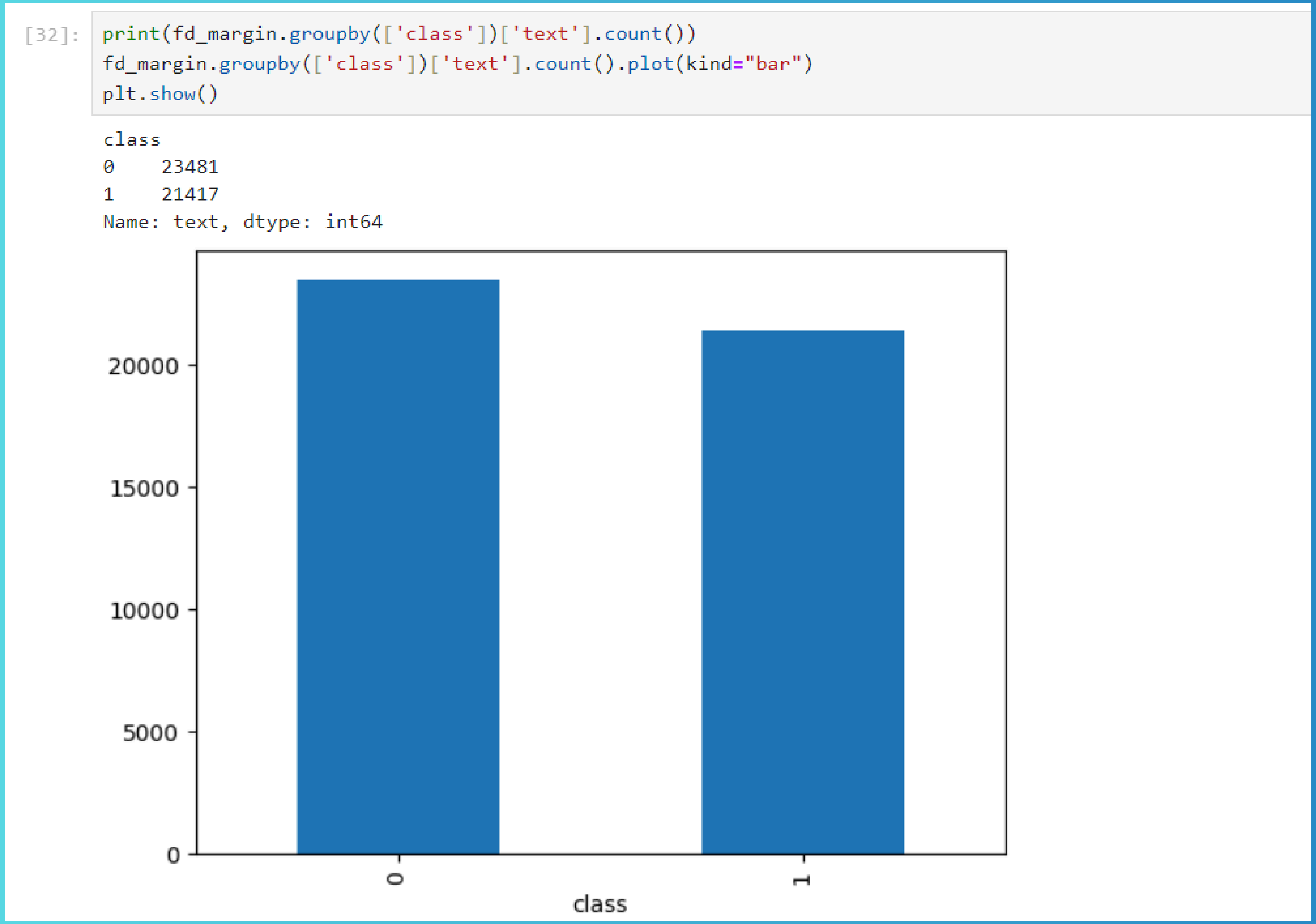
Imbalanced Data :

- Analysis: Compared the number of articles in true and fake news.
- Finding: The dataset is slightly imbalanced with more fake news articles than true news articles.

DESCRIPTIVE STATISTICS AND VISUALIZATIONS



Bar Plot for data type

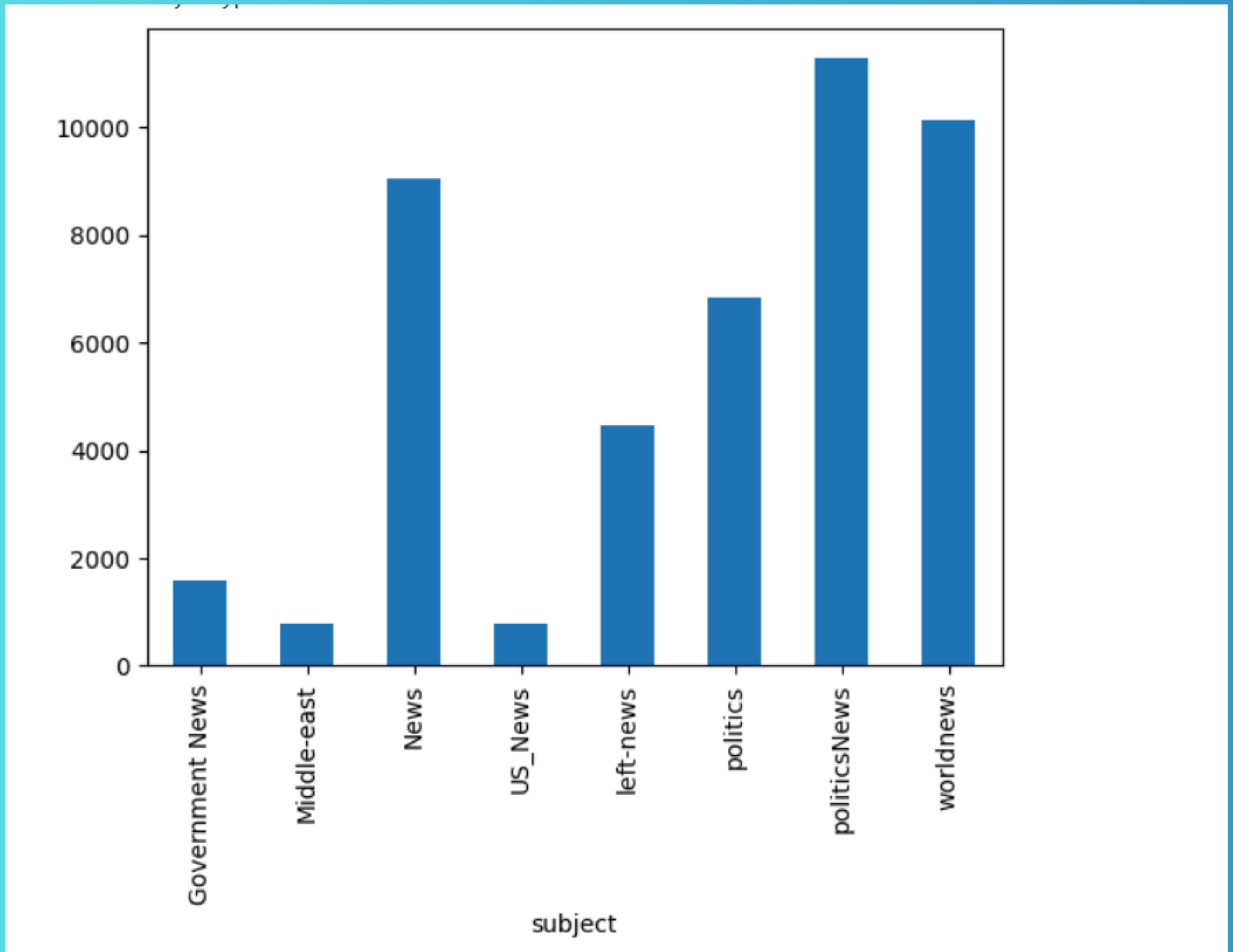


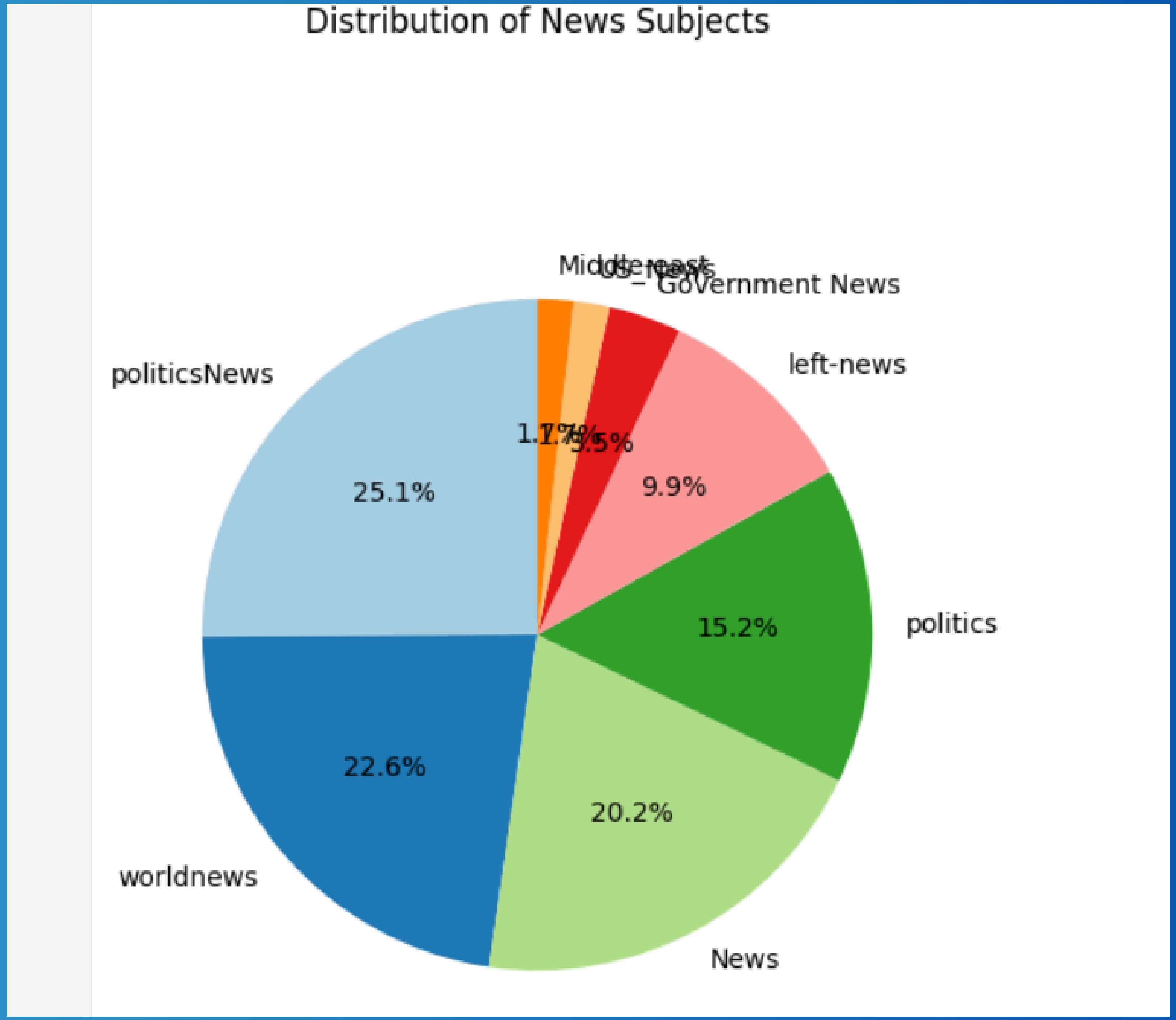
```
[32]: print(fd_margin.groupby(['class'])['text'].count())
      fd_margin.groupby(['class'])['text'].count().plot(kind="bar")
      plt.show()
```

```
class
0    23481
1    21417
Name: text, dtype: int64
```

- Bar plot of the count of text by class for fd_margin data set. The fake data is represented by 0 and the true data is represented by 1.

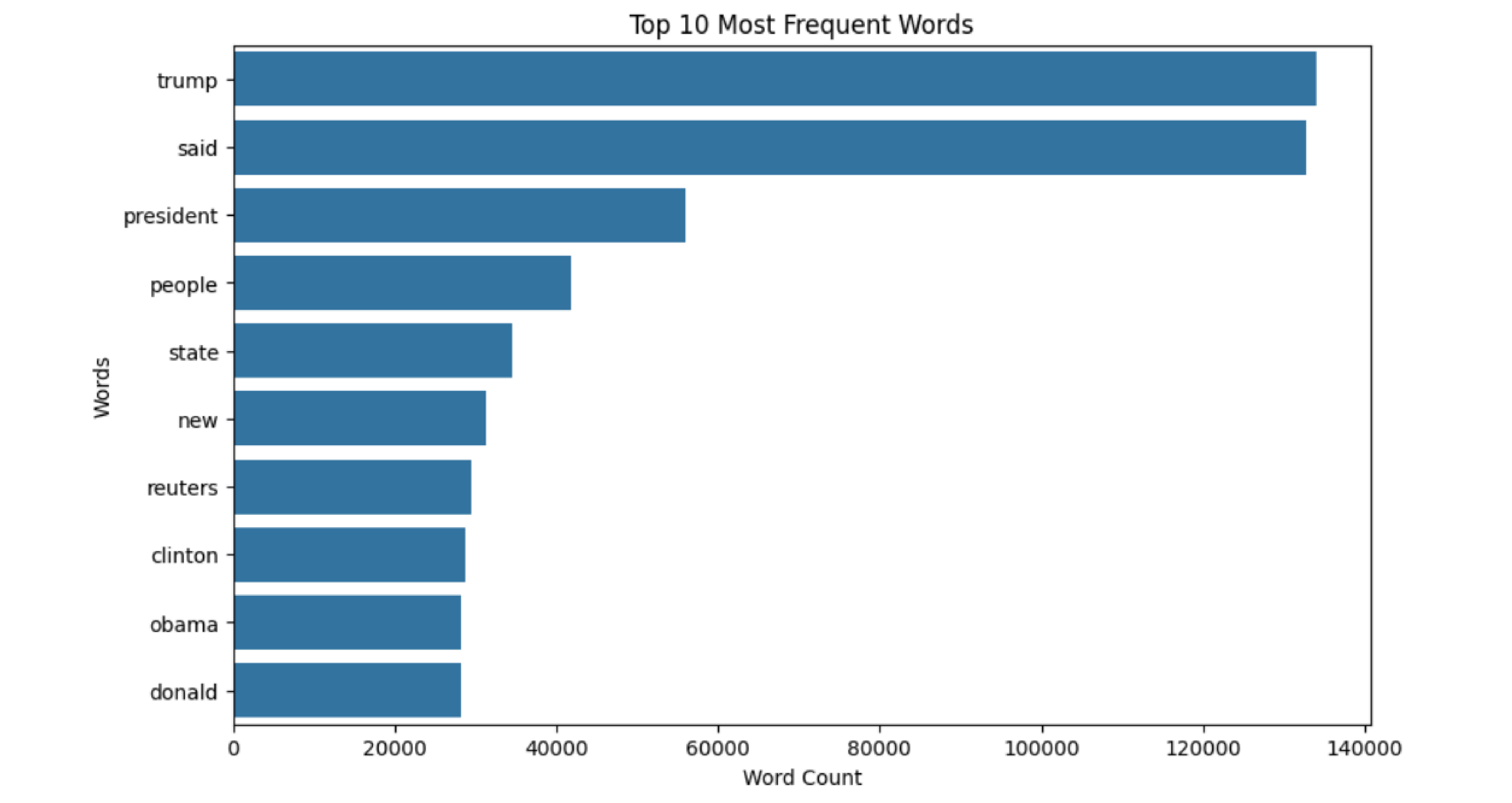
Visualization of news category in Histogtram and Pie Chart





The x-axis represents the subject categories, The y-axis represents the count or frequency of occurrences for each subject,

Top 10 most frequent words in the dataset

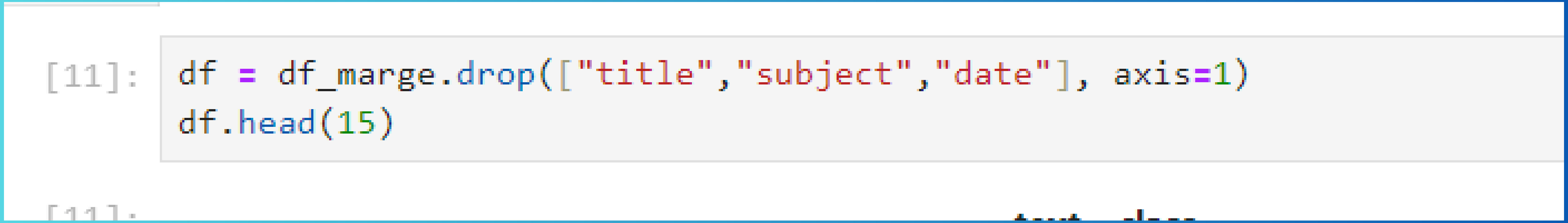


According to the bar chart we can identify ‘trump’ is the most frequent word in this dataset and ‘Donald’ is the least frequent word from top 10 most frequent words in the dataset.

DATA PREPARATION

After Visualization of our data set we plan to use only two columns for further processing.

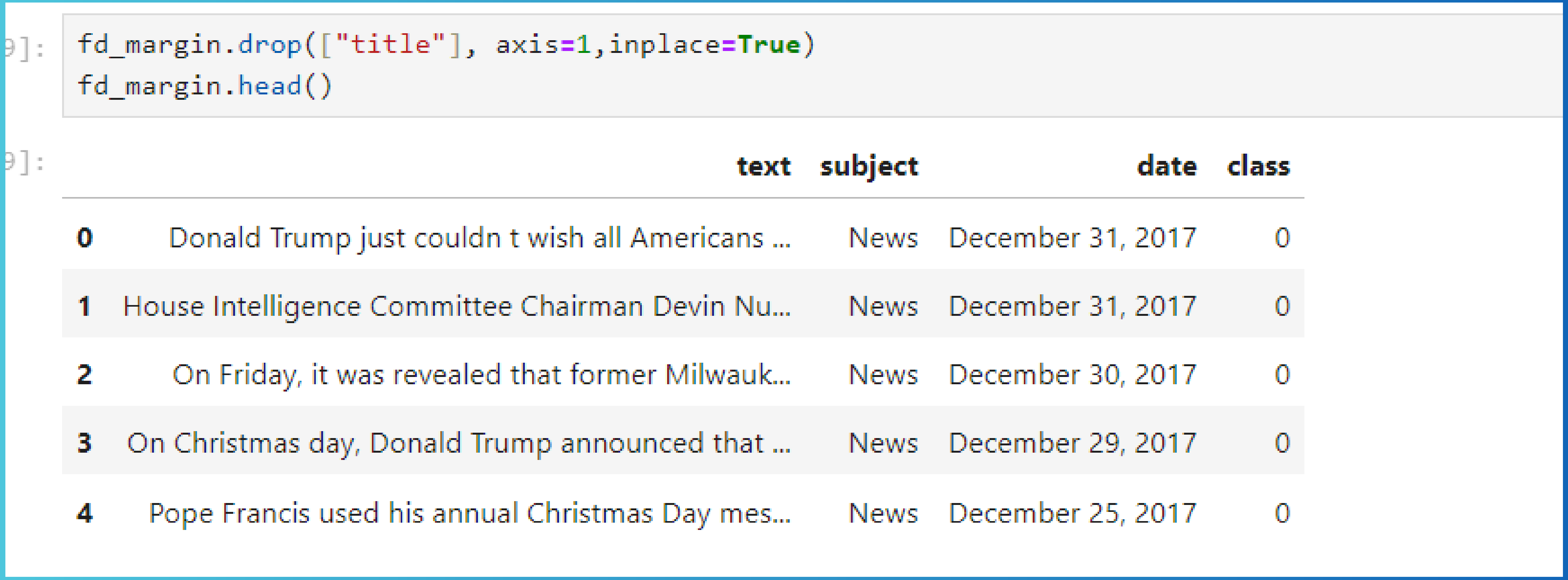
We dropped other unwanted columns such as “title”, “date” and “Subject”.



```
[11]: df = df_marge.drop(["title","subject","date"], axis=1)
      df.head(15)
```

```
[11]:                                         text    class
```

RESULT

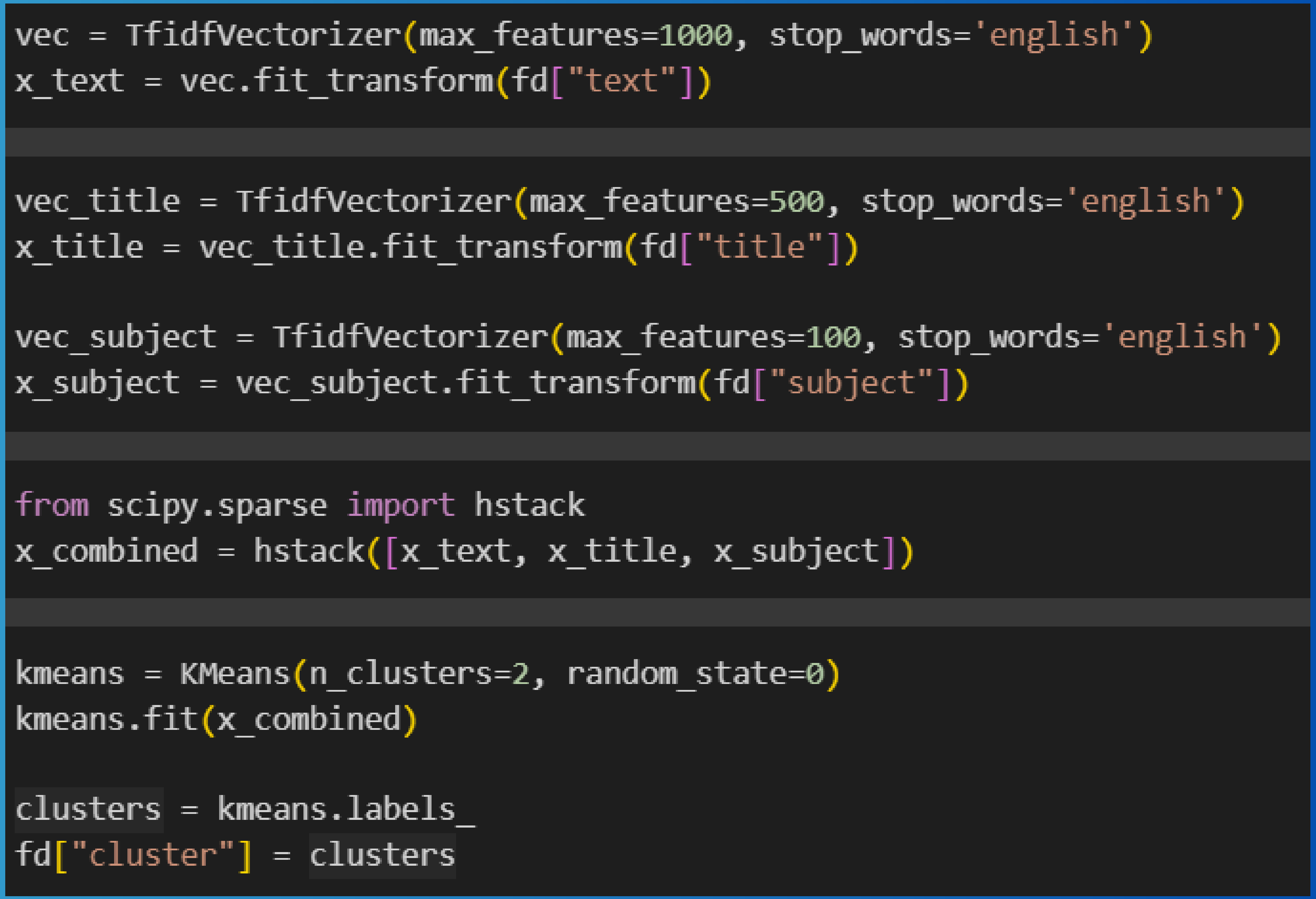


```
fd_margin.drop(["title"], axis=1,inplace=True)
fd_margin.head()
```

	text	subject	date	class
0	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
3	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
4	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0

Methodology

This code applies TF-IDF vectorization to three columns of text data: `text`, `title`, and `subject` from the dataset. Each column is transformed into numerical feature vectors using different configurations of TfidfVectorizer (with `max_features` specifying the number of features). The three resulting feature matrices are then combined using hstack() to create a single feature matrix. This combined matrix is fed into a KMeans clustering model to group the articles into 2 clusters. The cluster assignments are stored in the `clusters` array and added to the DataFrame as a new column, "cluster".



```
vec = TfidfVectorizer(max_features=1000, stop_words='english')
x_text = vec.fit_transform(fd["text"])
```

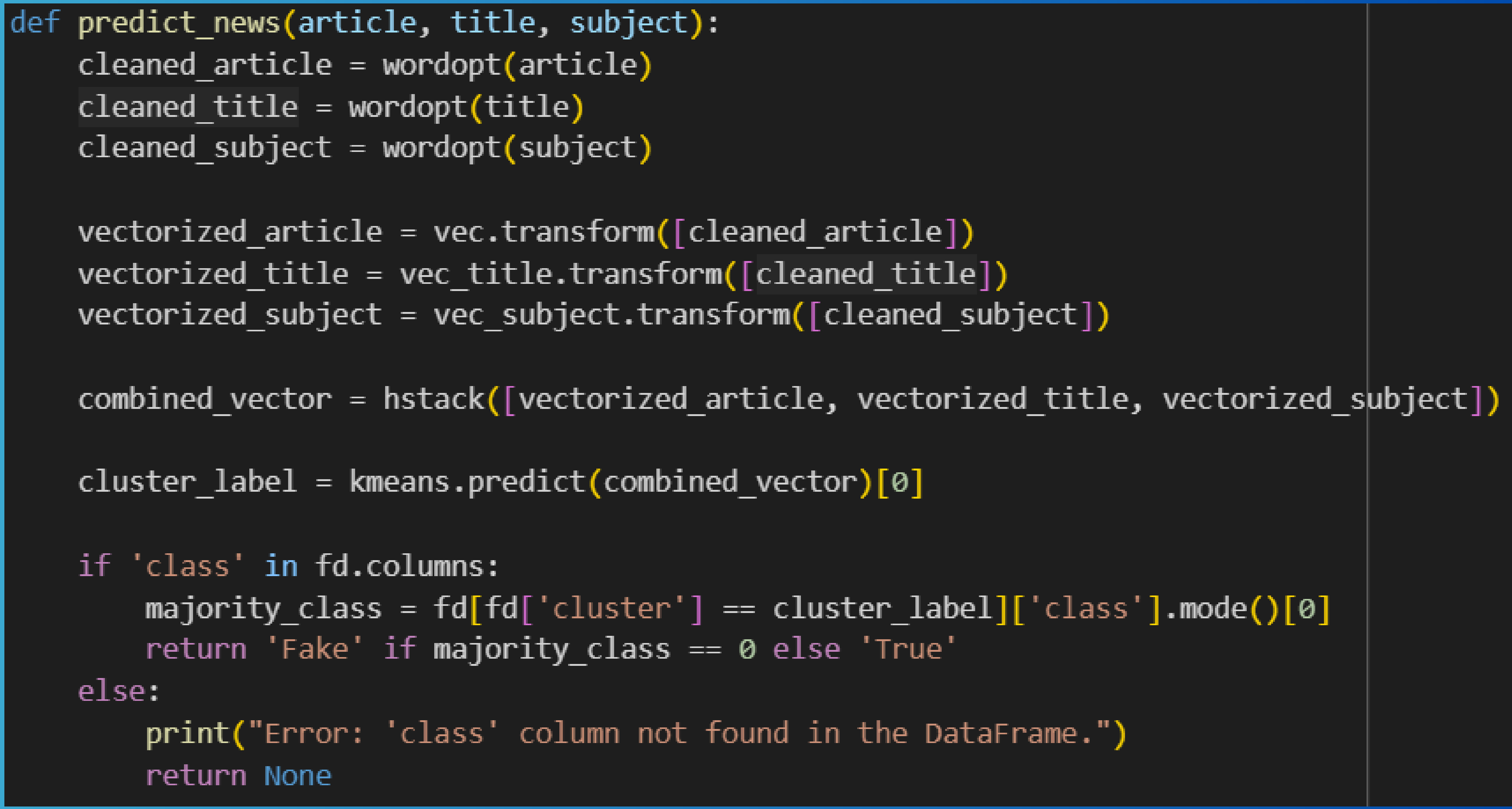
```
vec_title = TfidfVectorizer(max_features=500, stop_words='english')
x_title = vec_title.fit_transform(fd["title"])

vec_subject = TfidfVectorizer(max_features=100, stop_words='english')
x_subject = vec_subject.fit_transform(fd["subject"])
```

```
from scipy.sparse import hstack
x_combined = hstack([x_text, x_title, x_subject])
```

```
kmeans = KMeans(n_clusters=2, random_state=0)
kmeans.fit(x_combined)

clusters = kmeans.labels_
fd["cluster"] = clusters
```

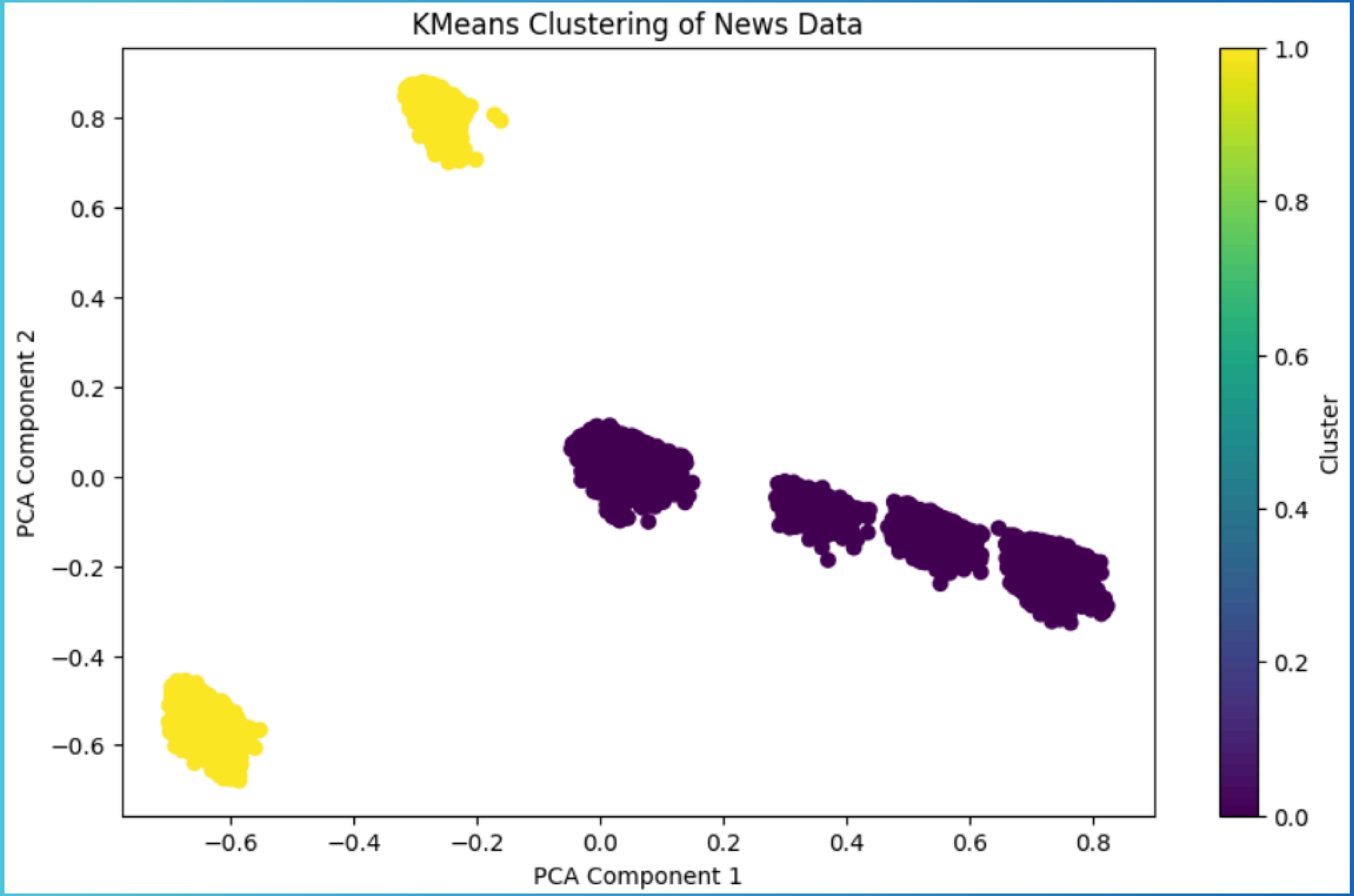
Training

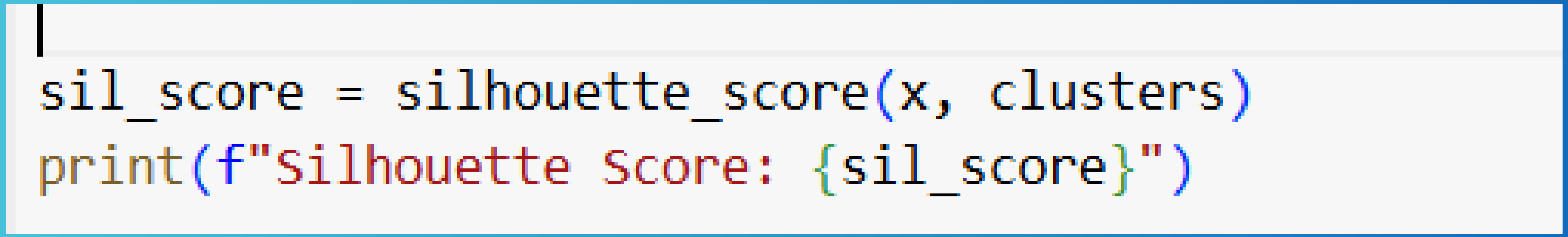
The `predict_news` function takes an article's text, title, and subject as inputs, cleans them using a text preprocessing function, and converts them into numerical vectors using TF-IDF. These vectors are combined and fed into a trained KMeans clustering model, which predicts the cluster the article belongs to. The function then checks the majority label (Fake or True) of articles in that cluster and returns "Fake" if the majority are fake, or "True" otherwise. If the necessary columns are missing, it returns an error message.



```
def predict_news(article, title, subject):
    cleaned_article = wordopt(article)
    cleaned_title = wordopt(title)
    cleaned_subject = wordopt(subject)

    vectorized_article = vec.transform([cleaned_article])
    vectorized_title = vec_title.transform([cleaned_title])
    vectorized_subject = vec_subject.transform([cleaned_subject])

    combined_vector = hstack([vectorized_article, vectorized_title, vectorized_subject])

    cluster_label = kmeans.predict(combined_vector)[0]

    if 'class' in fd.columns:
        majority_class = fd[fd['cluster'] == cluster_label]['class'].mode()[0]
        return 'Fake' if majority_class == 0 else 'True'
    else:
        print("Error: 'class' column not found in the DataFrame.")
        return None
```

Testing

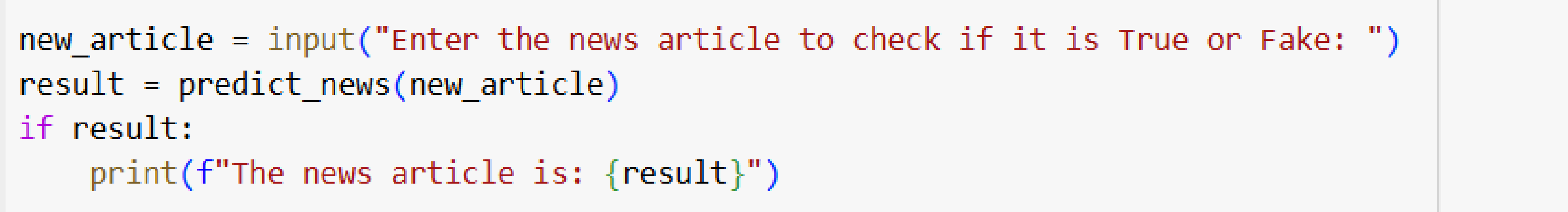

KMeans Clustering of News Data
PCA Component 1
PCA Component 2
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
Cluster
1.0
0.8
0.6
0.4
0.2
0.0



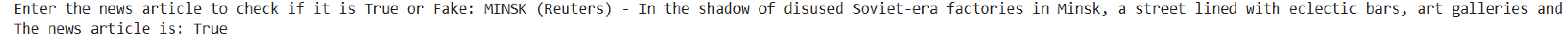
```python
sil_score = silhouette_score(x, clusters)
print(f"Silhouette Score: {sil_score}")
```

The output of the silhouette score is: 0.09

Manual Testing



```
new_article = input("Enter the news article to check if it is True or Fake: ")
result = predict_news(new_article)
if result:
    print(f"The news article is: {result}")
```



```
Enter the news article to check if it is True or Fake: MINSK (Reuters) - In the shadow of disused Soviet-era factories in Minsk, a street lined with eclectic bars, art galleries and
The news article is: True
```

Conclusion

We learned to detect fake news with Python. We took a political dataset, implemented a TF-IDF Vectorizer, initialized a Cluster analysis, and fit our model. We ended 34 up obtaining an accuracy of 70 in magnitude And we tested the result manually that the news is true or false.