In [1]:

```
#importing pandas,numpy,matplotlib,seaborn and sklearn

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn import feature_extraction, linear_model, model_selection, preprocessi
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
```

In [2]:

```
fake = pd.read_csv("Fake.csv")
true = pd.read_csv("True.csv")
```

In [3]:

```
fake.shape
```

Out[3]:

```
(23481, 4)
```

In [4]:

```
true.shape
```

Out[4]:

```
(21417, 4)
```

In [5]:

```
# Add flag to track fake and real
fake['target'] = 'fake'
true['target'] = 'true'
```

In [6]:

```
fake.head()
```

Out[6]:

	title	text	subject	date	target
0	Donald Trump Sends Out Embarrassing New Year'…	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	fake
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	fake
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	fake
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	fake
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	fake

In [7]:

```
true.head()
```

Out[7]:

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

In [8]:

```python
# Concatenate dataframes
data = pd.concat([fake, true]).reset_index(drop = True)
data.shape
```

Out[8]:

```
(44898, 5)
```

In [9]:

```python
data.head(5)
```

Out[9]:

	title	text	subject	date	target
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	fake
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	fake
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	fake
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	fake
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	fake

In [10]:

```python
data.tail(5)
```

Out[10]:

	title	text	subject	date	target
44893	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	true
44894	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	true
44895	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	true
44896	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	true
44897	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	true

In [11]:

```python
# Shuffle the data
from sklearn.utils import shuffle
```

```
data = shuffle(data)
data = data.reset_index(drop=True)
```

In [12]:
```
# Check the data
data.head()
```

Out[12]:

	title	text	subject	date	target
0	Trump Is Giddy About His Upcoming Meeting Wit...	For the first time since taking office, Donald...	News	June 26, 2017	fake
1	Brexit 'no deal' is massively less probable af...	LONDON (Reuters) - Britain is less likely to l...	worldnews	December 14, 2017	true
2	Trump Goes FULL Propaganda Declaring He Knows...	Who needs experts, really, when one can rely o...	News	April 4, 2016	fake
3	Trump's Central America plan will not boost mi...	(Reuters) - The Trump administration s effort ...	worldnews	September 20, 2017	true
4	Donald Trump’s Son Bashes His Father’s Campai...	If you ve never had to suffer through any of T...	News	August 17, 2016	fake

In [13]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 44898 entries, 0 to 44897
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   title    44898 non-null  object
 1   text     44898 non-null  object
 2   subject  44898 non-null  object
 3   date     44898 non-null  object
 4   target   44898 non-null  object
dtypes: object(5)
memory usage: 1.7+ MB
```

In [14]:
```
# Removing the date
data.drop(["date"],axis=1,inplace=True)
data.head()
```

Out[14]:

	title	text	subject	target
0	Trump Is Giddy About His Upcoming Meeting Wit...	For the first time since taking office, Donald...	News	fake
1	Brexit 'no deal' is massively less probable af...	LONDON (Reuters) - Britain is less likely to l...	worldnews	true
2	Trump Goes FULL Propaganda Declaring He Knows...	Who needs experts, really, when one can rely o...	News	fake
3	Trump's Central America plan will not boost mi...	(Reuters) - The Trump administration s effort ...	worldnews	true
4	Donald Trump’s Son Bashes His Father’s Campai...	If you ve never had to suffer through any of T...	News	fake

In [15]:
```
# Removing the title
data.drop(["title"],axis=1,inplace=True)
data.head()
```

Out[15]:

	text	subject	target
0	For the first time since taking office, Donald...	News	fake
1	LONDON (Reuters) - Britain is less likely to l...	worldnews	true
2	Who needs experts, really, when one can rely o...	News	fake
3	(Reuters) - The Trump administration s effort ...	worldnews	true
4	If you ve never had to suffer through any of T...	News	fake

In [16]:

```
# Convert to lowercase

data['text'] = data['text'].apply(lambda x: x.lower())
data.head()
```

Out[16]:

	text	subject	target
0	for the first time since taking office, donald...	News	fake
1	london (reuters) - britain is less likely to l...	worldnews	true
2	who needs experts, really, when one can rely o...	News	fake
3	(reuters) - the trump administration s effort ...	worldnews	true
4	if you ve never had to suffer through any of t...	News	fake

In [17]:

```
# Remove punctuation

import string

def punctuation_removal(text):
    all_list = [char for char in text if char not in string.punctuation]
    clean_str = ''.join(all_list)
    return clean_str

data['text'] = data['text'].apply(punctuation_removal)
```

In [18]:

```
# Check
data.head()
```

Out[18]:

	text	subject	target
0	for the first time since taking office donald ...	News	fake
1	london reuters britain is less likely to leav...	worldnews	true
2	who needs experts really when one can rely on ...	News	fake
3	reuters the trump administration s effort to ...	worldnews	true
4	if you ve never had to suffer through any of t...	News	fake

In [20]:

```
# Removing stopwords
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
stop = stopwords.words('english')

data['text'] = data['text'].apply(lambda x: ' '.join([word for word in x.split() i
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\sudha\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [21]:

```
data.head()
```

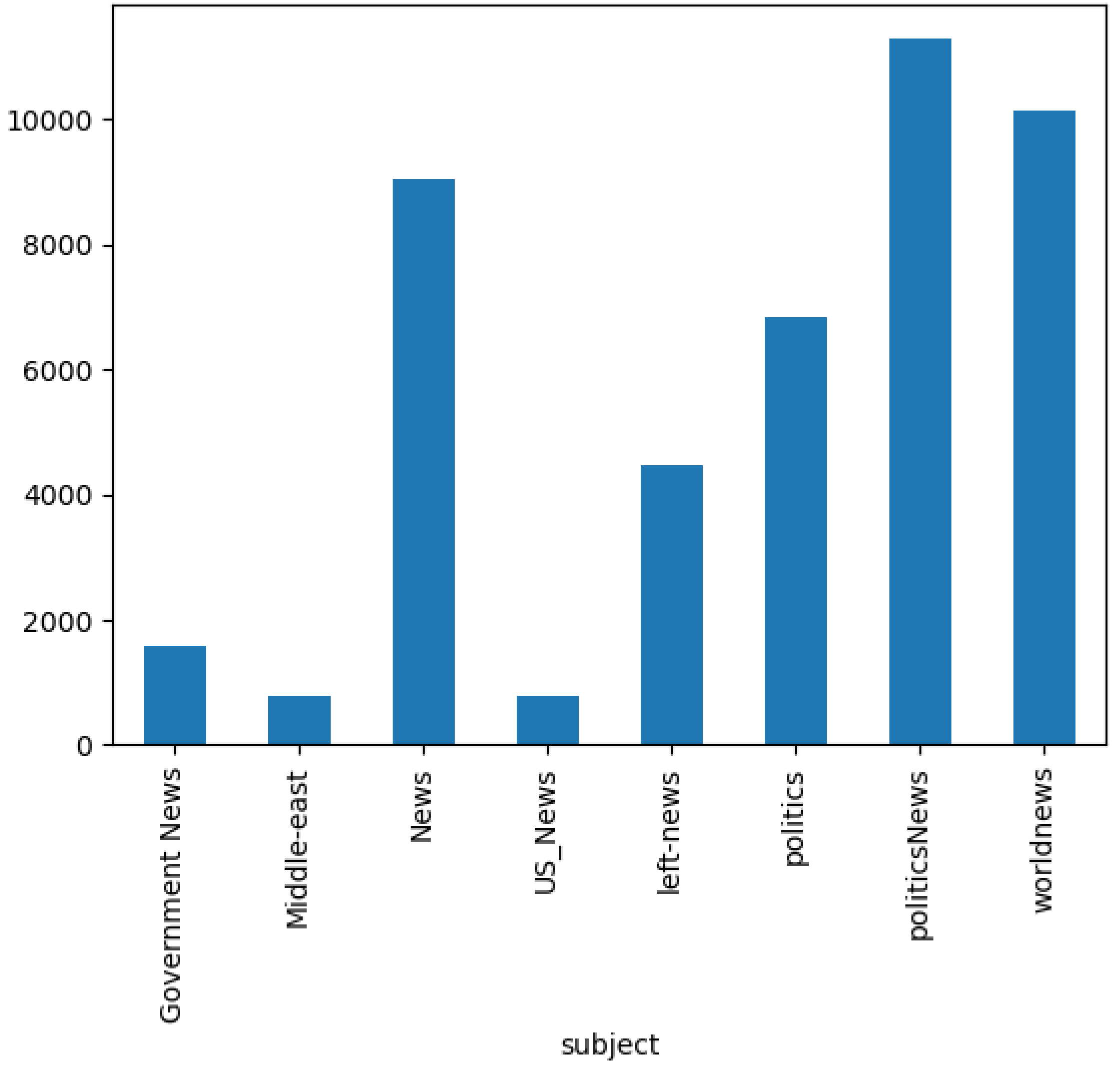
Out[21]:

	text	subject	target
0	first time since taking office donald trump re...	News	fake
1	london reuters britain less likely leave europ...	worldnews	true
2	needs experts really one rely intuition mean d...	News	fake
3	reuters trump administration effort combat vio...	worldnews	true
4	never suffer trump surrogates trying participa...	News	fake

In [22]:

```
# How many articles per subject?
print(data.groupby(['subject'])['text'].count())
data.groupby(['subject'])['text'].count().plot(kind="bar")
plt.show()
```

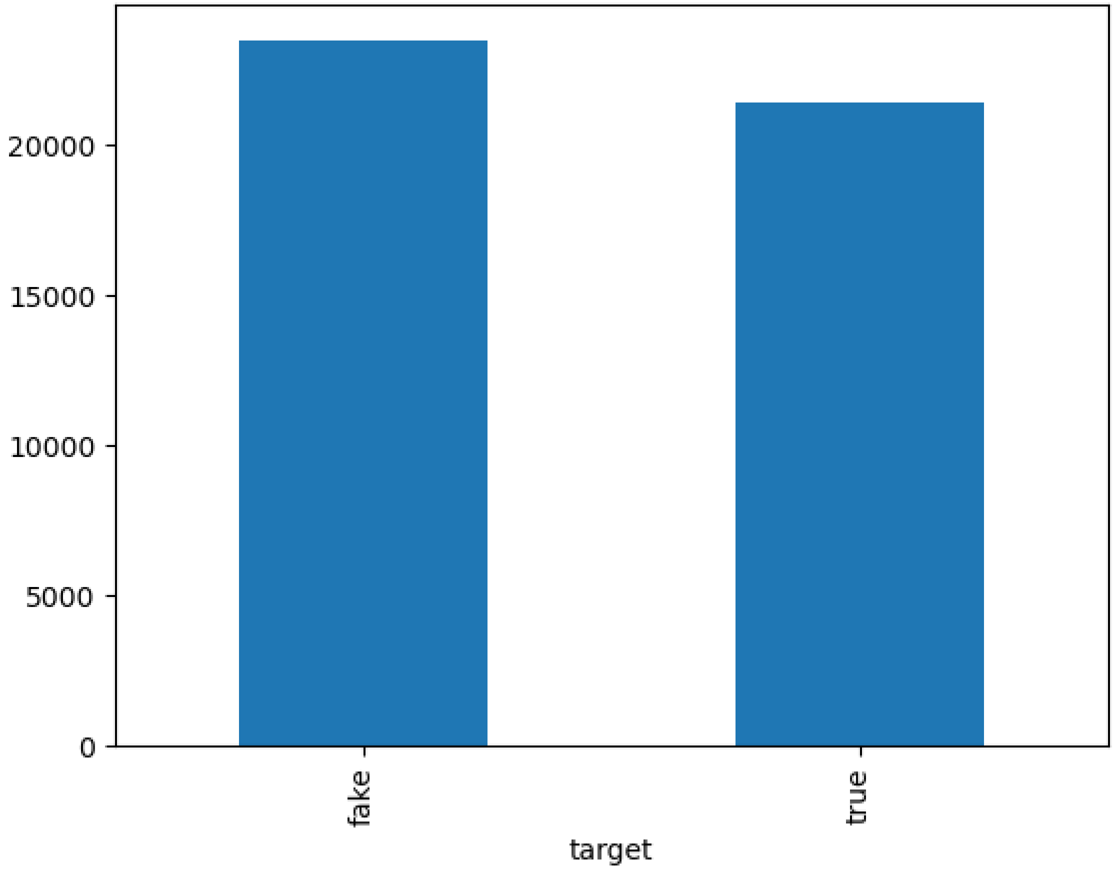
```
subject
Government News     1570
Middle-east          778
News                9050
US_News              783
left-news           4459
politics            6841
politicsNews       11272
worldnews          10145
Name: text, dtype: int64
```



In [23]:

```
# How many fake and real articles?
print(data.groupby(['target'])['text'].count())
data.groupby(['target'])['text'].count().plot(kind="bar")
plt.show()
```

```
target
fake    23481
true    21417
Name: text, dtype: int64
```

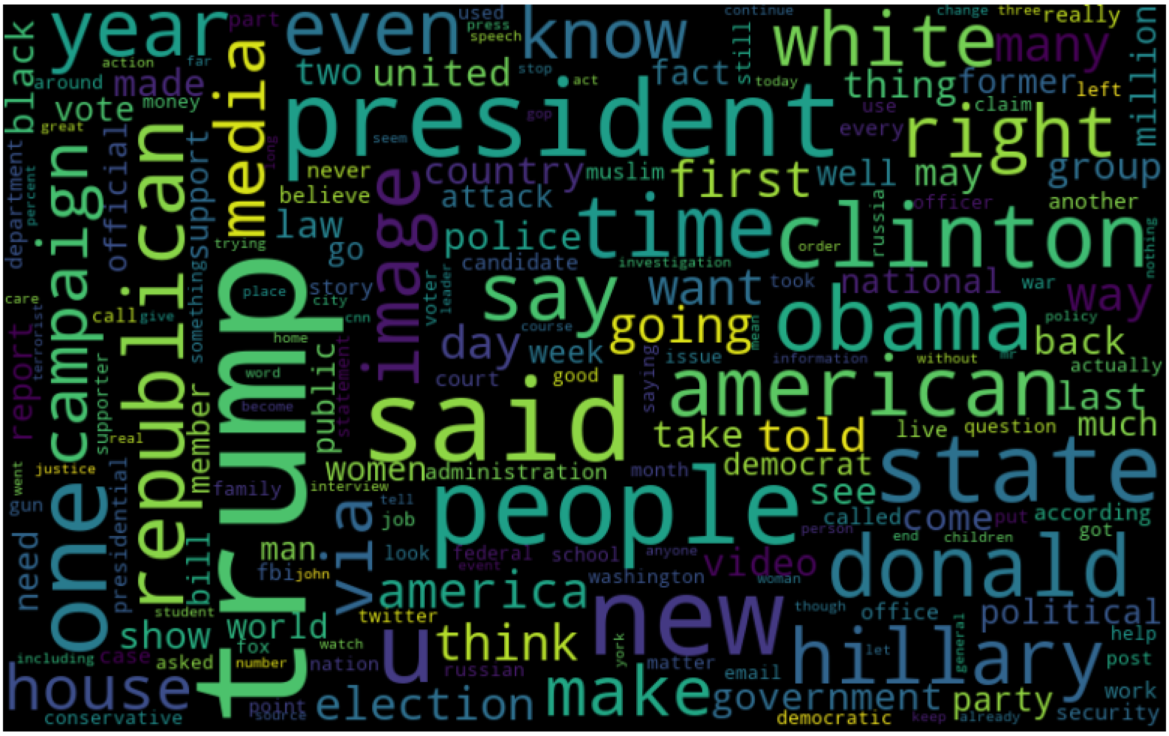


In [24]:

```
#!pip install wordcloud
```

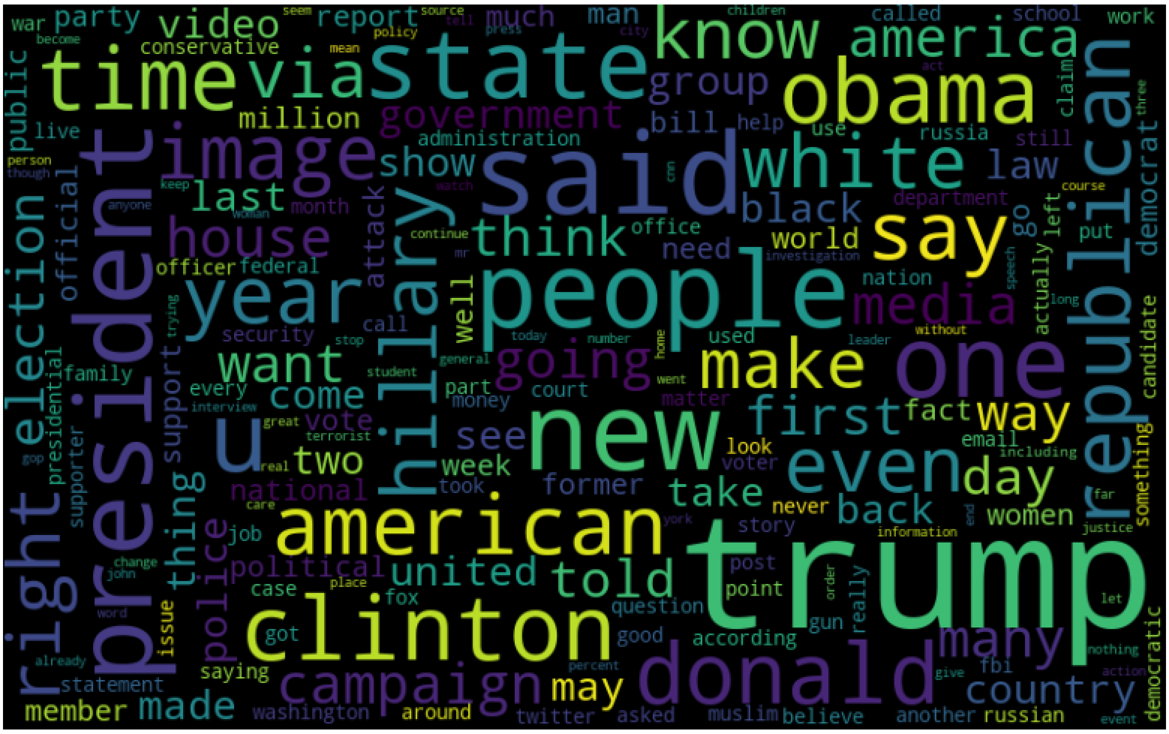
In [25]:

```
# Word cloud for fake news
from wordcloud import WordCloud

fake_data = data[data["target"] == "fake"]
all_words = ' '.join([text for text in fake_data.text])

wordcloud = WordCloud(width= 800, height= 500,
                          max_font_size = 110,
                          collocations = False).generate(all_words)

plt.figure(figsize=(10,7))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```



In [26]:

```
# Word cloud for real news
from wordcloud import WordCloud

real_data = data[data["target"] == "true"]
all_words = ' '.join([text for text in fake_data.text])

wordcloud = WordCloud(width= 800, height= 500,
                      max_font_size = 110,
                      collocations = False).generate(all_words)

plt.figure(figsize=(10,7))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

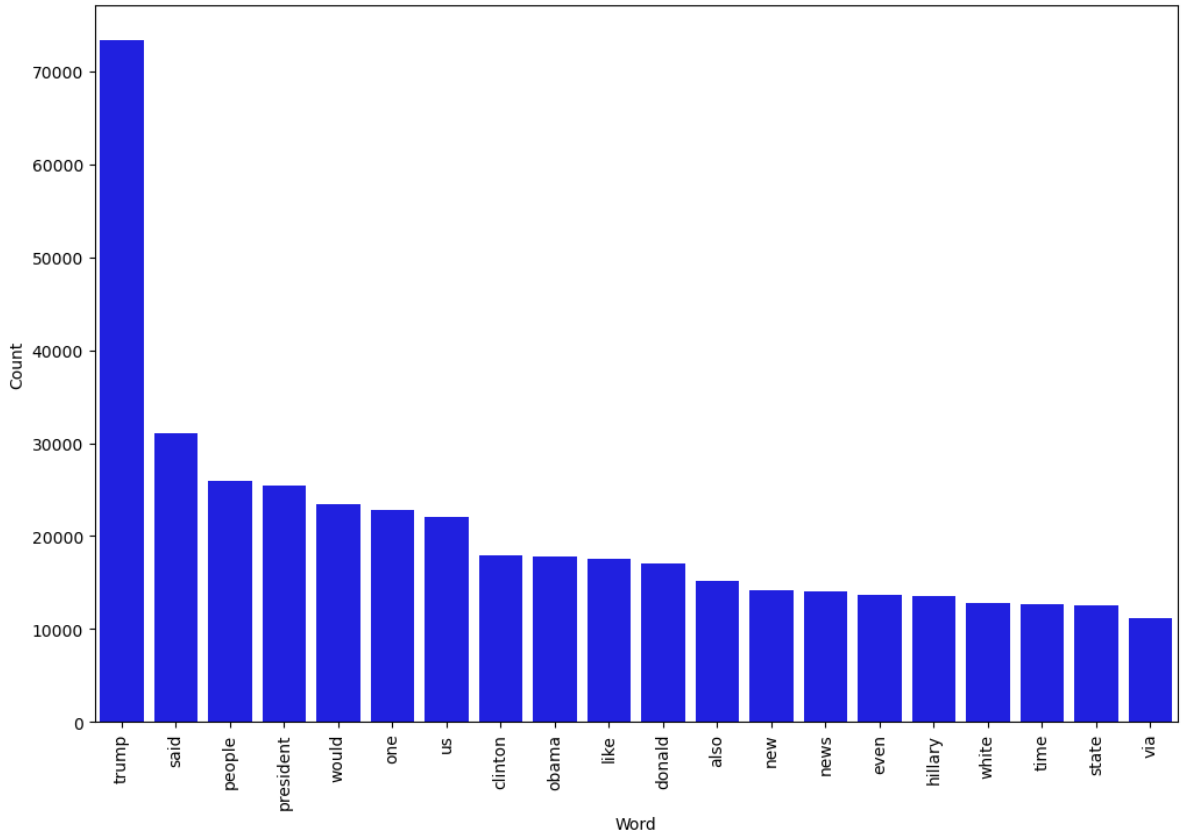


In [27]:

```
# Most frequent words counter
from nltk import tokenize

token_space = tokenize.WhitespaceTokenizer()

def counter(text, column_text, quantity):
    all_words = ' '.join([text for text in text[column_text]])
    token_phrase = token_space.tokenize(all_words)
    frequency = nltk.FreqDist(token_phrase)
    df_frequency = pd.DataFrame({"Word": list(frequency.keys()),
                                   "Frequency": list(frequency.values())})
    df_frequency = df_frequency.nlargest(columns = "Frequency", n = quantity)
    plt.figure(figsize=(12,8))
    ax = sns.barplot(data = df_frequency, x = "Word", y = "Frequency", color = 'bl
    ax.set(ylabel = "Count")
    plt.xticks(rotation='vertical')
    plt.show()
```

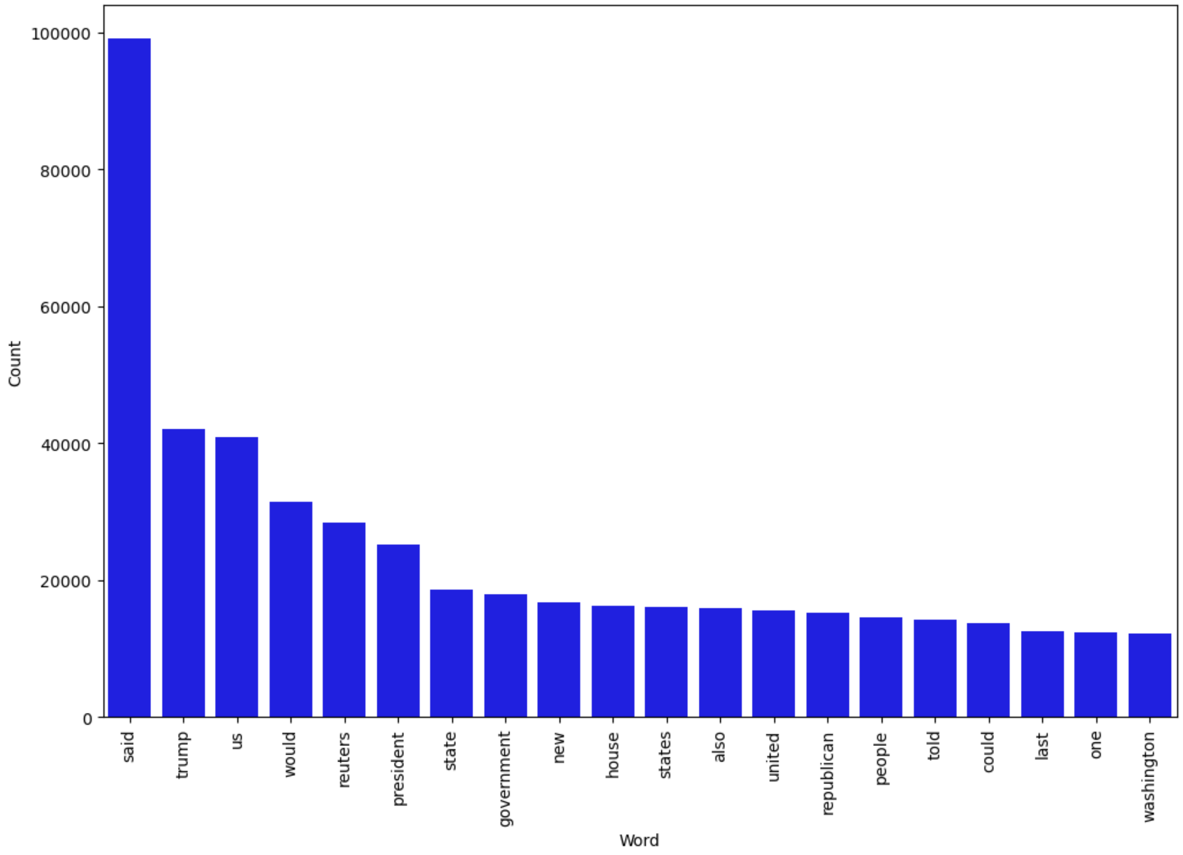
In [28]:

```
# Most frequent words in fake news
counter(data[data["target"] == "fake"], "text", 20)
```



In [29]:

```
# Most frequent words in real news
counter(data[data["target"] == "true"], "text", 20)
```



In [30]:

```
# Function to plot the confusion matrix
from sklearn import metrics
import itertools

def plot_confusion_matrix(cm, classes,
                          normalize=False,
                          title='Confusion matrix',
                          cmap=plt.cm.Blues):

    plt.imshow(cm, interpolation='nearest', cmap=cmap)
    plt.title(title)
    plt.colorbar()
    tick_marks = np.arange(len(classes))
    plt.xticks(tick_marks, classes, rotation=45)
    plt.yticks(tick_marks, classes)

    if normalize:
        cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
        print("Normalized confusion matrix")
    else:
        print('Confusion matrix, without normalization')

    thresh = cm.max() / 2.
    for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
        plt.text(j, i, cm[i, j],
                 horizontalalignment="center",
                 color="white" if cm[i, j] > thresh else "black")

    plt.tight_layout()
    plt.ylabel('True label')
    plt.xlabel('Predicted label')
```

In [31]:

```
 # Split the data
X_train,X_test,y_train,y_test = train_test_split(data['text'], data.target, test_s
```

In [32]:

```
X_train.head()
```

Out[32]:

```
36335    reuters raucous republican party debate thursd...
12384    ever since donald trump rise heartless people ...
24419    new york reuters donald trump’s political fort...
24740    episode 149 sunday wire show resumes sunday au...
27039    washington reuters us senate committee wednesd...
Name: text, dtype: object
```

In [33]:

```
y_train.head()
```

Out[33]:

```
36335    true
12384    fake
24419    true
24740    fake
27039    true
Name: target, dtype: object
```

In [34]:

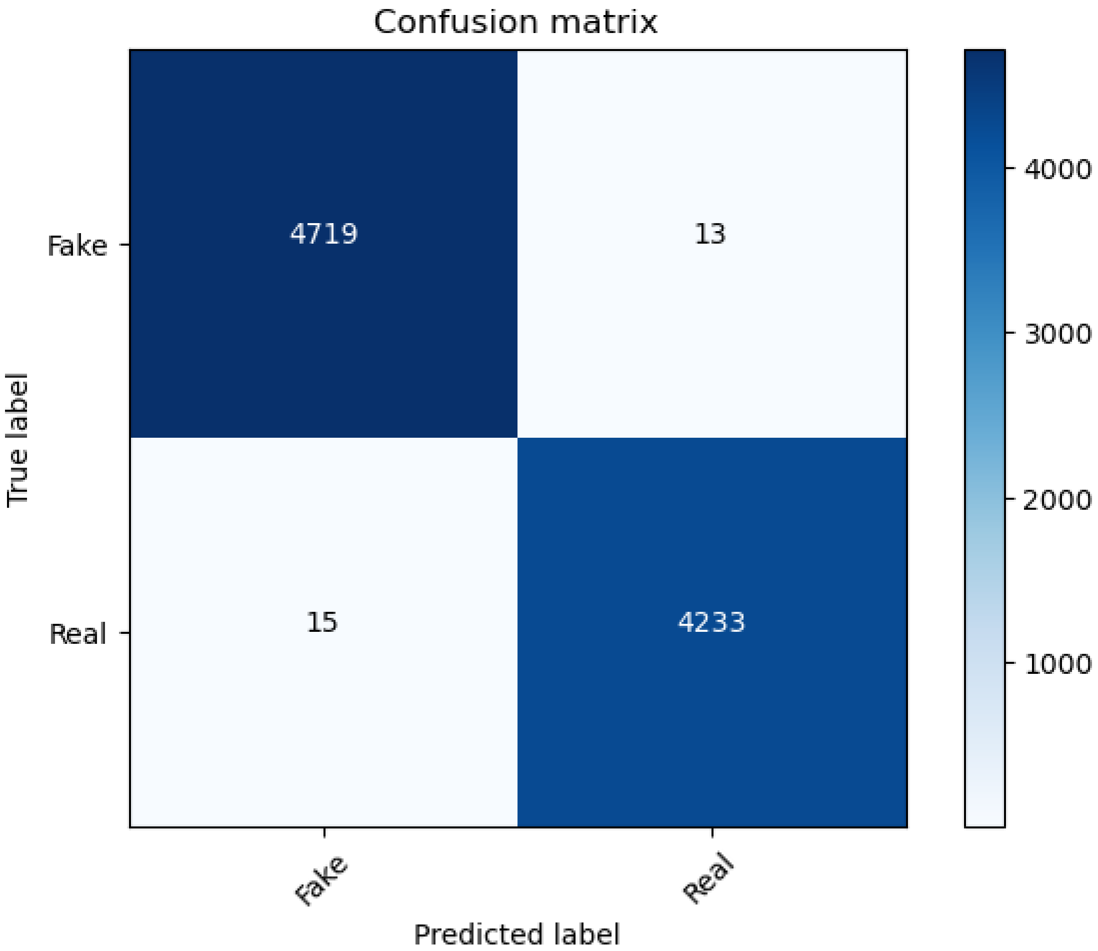
```
from sklearn.tree import DecisionTreeClassifier

# Vectorizing and applying TF-IDF
pipe = Pipeline([('vect', CountVectorizer()),
                 ('tfidf', TfidfTransformer()),
                 ('model', DecisionTreeClassifier(criterion= 'entropy',
                                            max_depth = 20,
                                            splitter='best',
                                            random_state=42))])
# Fitting the model
model = pipe.fit(X_train, y_train)

# Accuracy
prediction = model.predict(X_test)
print("accuracy: {}%".format(round(accuracy_score(y_test, prediction)*100,2)))
```

```
accuracy: 99.69%
```

In [37]:

```
cm = metrics.confusion_matrix(y_test, prediction)
plot_confusion_matrix(cm, classes=['Fake', 'Real'])
```

```
Confusion matrix, without normalization
```


Confusion matrix
True label
Fake
Real
4719
13
15
4233
Predicted label
Fake
Real
4000
3000
2000
1000

```
In [ ]:
```