```
import nltk
nltk.download('punkt')
import tensorflow as tf
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud, STOPWORDS
import nltk
import re
from nltk.stem import PorterStemmer, WordNetLemmatizer
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize, sent_tokenize
import gensim
from gensim.utils import simple_preprocess
from gensim.parsing.preprocessing import STOPWORDS
# import keras
from tensorflow.keras.preprocessing.text import one_hot, Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Flatten, Embedding, Input, LSTM, Conv1D, MaxPool1D, Bidirectional
from tensorflow.keras.models import Model
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
```

```
fake = pd.read_csv'data/Fake.csv')
```

```
true = pd.read_csv('data/True.csv')
```

```
true
```

	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
...	...	...	...	...
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017

```
fake
```

	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 ...	House Intelligence Committee ...	News	December 31,
4	Pope Francis Just Called Out Donald ...	Pope Francis used his annual ...	News	December 25,

```
true.isnull().sum()
```

```
title      0
text       0
subject    0
date       0
dtype: int64
```

```
fake.isnull().sum()
```

```
title      0
text       0
subject    0
date       0
dtype: int64
```

```
true.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21417 entries, 0 to 21416
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   title    21417 non-null  object
 1   text     21417 non-null  object
 2   subject  21417 non-null  object
 3   date     21417 non-null  object
dtypes: object(4)
memory usage: 669.4+ KB
```

```
fake.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23481 entries, 0 to 23480
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   title    23481 non-null  object
 1   text     23481 non-null  object
 2   subject  23481 non-null  object
 3   date     23481 non-null  object
dtypes: object(4)
memory usage: 733.9+ KB
```

```
true['isfake'] = 0
true.head()
```

	title	text	subject	date	isfake
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	0
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	0
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	0
3	FBI Russia probe helped by ...	WASHINGTON (Reuters) - Trump ...	...	December 30,	0

```
fake['isfake'] = 1
fake.head()
```

```
df = pd.concat([true, fake]).reset_index(drop = True)
df
```

	title	text	subject	date	isfake
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	0
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	0
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	0
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	0
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	0
...	...	...	...	...	...
44893	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	1
44894	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016	1

```
# drpo unnecessary Date column # RUN ONLY ONCE
df.drop(columns = ['date'], inplace = True)


# combine title and text together
df['original'] = df['title'] + ' ' + df['text']
df.head()
```

	title	text	subject	isfake	original
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	0	As U.S. budget fight looms, Republicans flip t...
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	0	U.S. military to accept transgender recruits o...
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	0	Senior U.S. Republican senator: 'Let Mr. Muell...
3	FBI Russia probe helped	WASHINGTON (Reuters) - Trump	.... ..	0	FBI Russia probe helped

```
# download stopwords
nltk.download("stopwords")
```

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

```
# Obtain additional stopwords from nltk
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
stop_words.extend(['from', 'subject', 're', 'edu', 'use']) # add additional stop words to remove
# Remove stopwords and remove words with 2 or less characters using gensim
def preprocess(text):
    result = []
    for token in gensim.utils.simple_preprocess(text):
        if token not in gensim.parsing.preprocessing.STOPWORDS and len(token) > 3 and token not in stop_words:
            result.append(token)
    return result


# Apply the function to the dataframe
df['clean'] = df['original'].apply(preprocess)


# Obtain the total words present in the dataset
list_of_words = []
for i in df.clean:
    for j in i:
        list_of_words.append(j)
```

```
len(list_of_words)
```

```
9276947
```

```
# Obtain the total number of unique words (using set())
total_words = len(list(set(list_of_words)))
total_words
```

```
108704
```

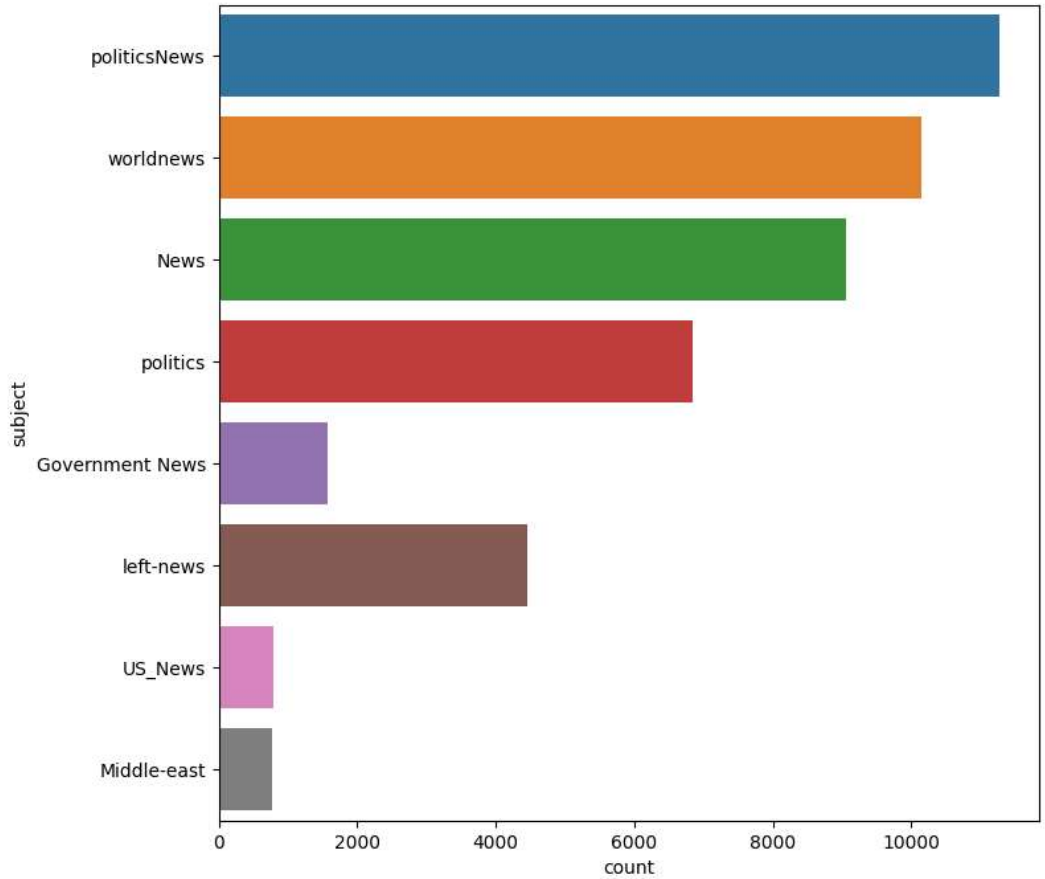
```
# join the words into a string (words seperated by " ")
df['clean_joined'] = df['clean'].apply(lambda x: " ".join(x))
```

```
df.head()
```

	title	text	subject	isfake	original	clean	clean_joined
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	0	As U.S. budget fight looms, Republicans flip t...	[budget, fight, looms, republicans, flip, fisc...	budget fight looms republicans flip fiscal scr...
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	0	U.S. military to accept transgender recruits o...	[military, accept, transgender, recruits, mond...	military accept transgender recruits monday pe...
2	Senior U.S. Republican	WASHINGTON (Reuters) -	politicsNews	0	Senior U.S. Republican	[senior, republican, senator	senior republican

```
# plot the number of samples in 'subject'
plt.figure(figsize = (8, 8))
sns.countplot(y = "subject", data = df)
```

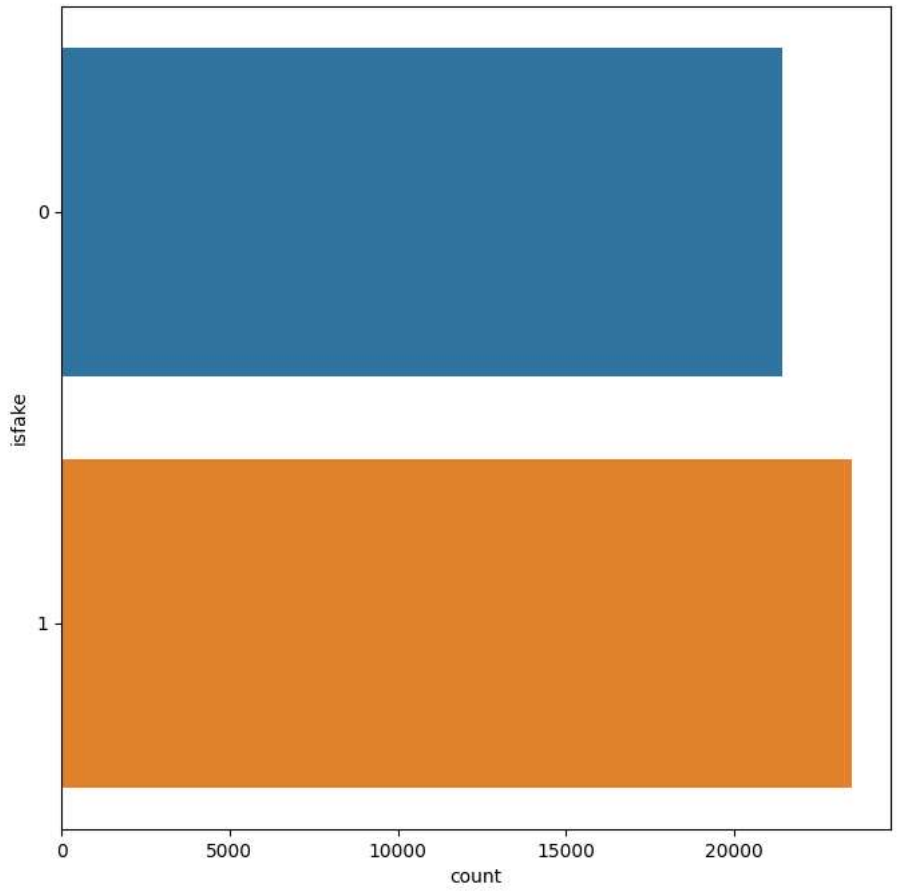
```
<Axes: xlabel='count', ylabel='subject'>
```




```
# plot the number of samples in 'isfake'
plt.figure(figsize = (8, 8))
```

```
sns.countplot(y = "isfake", data = df)
```

```
<Axes: xlabel='count', ylabel='isfake'>
```




```
# length of maximum document will be needed to create word embeddings
maxlen = -1
for doc in df.clean_joined:
    tokens = nltk.word_tokenize(doc)
    if(maxlen<len(tokens)):
        maxlen = len(tokens)
print("The maximum number of words in any document is =", maxlen)
```

```
The maximum number of words in any document is = 4405
```

```
# visualize the distribution of number of words in a text
import plotly.express as px #interactive visualizations
fig = px.histogram(x = [len(nltk.word_tokenize(x)) for x in df.clean_joined], nbins = 100)
fig.show()
```

```
# split data into test and train
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(df.clean_joined, df.isfake, test_size = 0.2)


from nltk import word_tokenize


# Tokenizer to tokenize the words and create sequences of tokenized words
tokenizer = Tokenizer(num_words = total_words)
tokenizer.fit_on_texts(x_train)
train_sequences = tokenizer.texts_to_sequences(x_train)
test_sequences = tokenizer.texts_to_sequences(x_test)


len(train_sequences)
```

```
35918
```

```
len(test_sequences)
```

```
8980
```

```
padded_train = pad_sequences(train_sequences,maxlen = 40, padding = 'post', truncating = 'post')
padded_test = pad_sequences(test_sequences,maxlen = 40, truncating = 'post')


for i,doc in enumerate(padded_train[:2]):
    print("The padded encoding for document",i+1," is : ",doc)
```

```
The padded encoding for document 1  is :  [  71    1  732  370 2574  249 4836  734    9  493    3 2865    3   10
    1  309 1133  370 2574  249 4000 1045  637   77 7951    5   21  152
 4836    2  443  576    0    0    0    0    0    0    0    0]
The padded encoding for document 2  is :  [11561  9243   451  1804     1    27  6922  7076  2725  2164  3311  3789
    80     2   616  2239  6178  2943  3150    27     3     1  6663  6793
  1722   240  3141   673  1266   485  1023  2239   293  1301   438 10567
  5552    37  4686  1804]
```

```
#butilding the sequential neural network with LSTM
model = Sequential()
model.add(Embedding(total_words, output_dim = 128))
model.add(Bidirectional(LSTM(128))) # no of neurons
model.add(Dense(128, activation = 'relu'))
model.add(Dense(1,activation= 'sigmoid'))
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['acc'])
model.summary()
```

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, None, 128)         13914112

 bidirectional (Bidirectiona  (None, 256)              263168
 l)

 dense (Dense)               (None, 128)               32896

 dense_1 (Dense)             (None, 1)                 129

=================================================================
Total params: 14,210,305
Trainable params: 14,210,305
Non-trainable params: 0
_________________________________________________________________
```

```
y_train = np.asarray(y_train)
```

```
# training the model
model.fit(padded_train, y_train, batch_size = 64, validation_split = 0.1, epochs = 2)
```

```
Epoch 1/2
506/506 [==============================] - 294s 571ms/step - loss: 0.0421 - acc: 0.9821 - val_loss: 0.0155 - val_acc: 0.9992
Epoch 2/2
506/506 [==============================] - 261s 517ms/step - loss: 0.0032 - acc: 0.9993 - val_loss: 0.0108 - val_acc: 0.9983
<keras.callbacks.History at 0x7f792ba8fa90>
```

```
#making prediction
pred = model.predict(padded_test)
```

```
281/281 [==============================] - 13s 42ms/step
```

```
# if the predicted value is >0.5 it is real else it is fake
prediction = []
for i in range(len(pred)):
    if pred[i].item() > 0.5:
        prediction.append(1)
    else:
        prediction.append(0)
```

```
# getting the accuracy of the model
from sklearn.metrics import accuracy_score
accuracy = accuracy_score(list(y_test), prediction)
print("Model Accuracy : ", accuracy)
```

```
Model Accuracy :  0.9978841870824053
```

```
# confusion matrix of the model
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(list(y_test), prediction)
plt.figure(figsize = (4,4))
sns.heatmap(cm, annot = True, cmap = 'Blues')
```

```
<Axes: >
```