FAKE NEWS DETECTION USING NLP

PREPROCESSING AND DATA MODELING

PREPROCESSING

Preprocessing the data is a critical step in building a fake news detection model.

Common preprocessing steps include:

Text cleaning : Remove special characters, extra whitespaces, and perform lowercasing.

Tokenization : Split text into individual words (tokens).

Stopword removal : Eliminate common words that don't carry much meaning.

Vectorization : Convert text data into numerical features, typically using TF-IDF or Count Vectorization.

IMPORT PACKAGES

import numpy as np

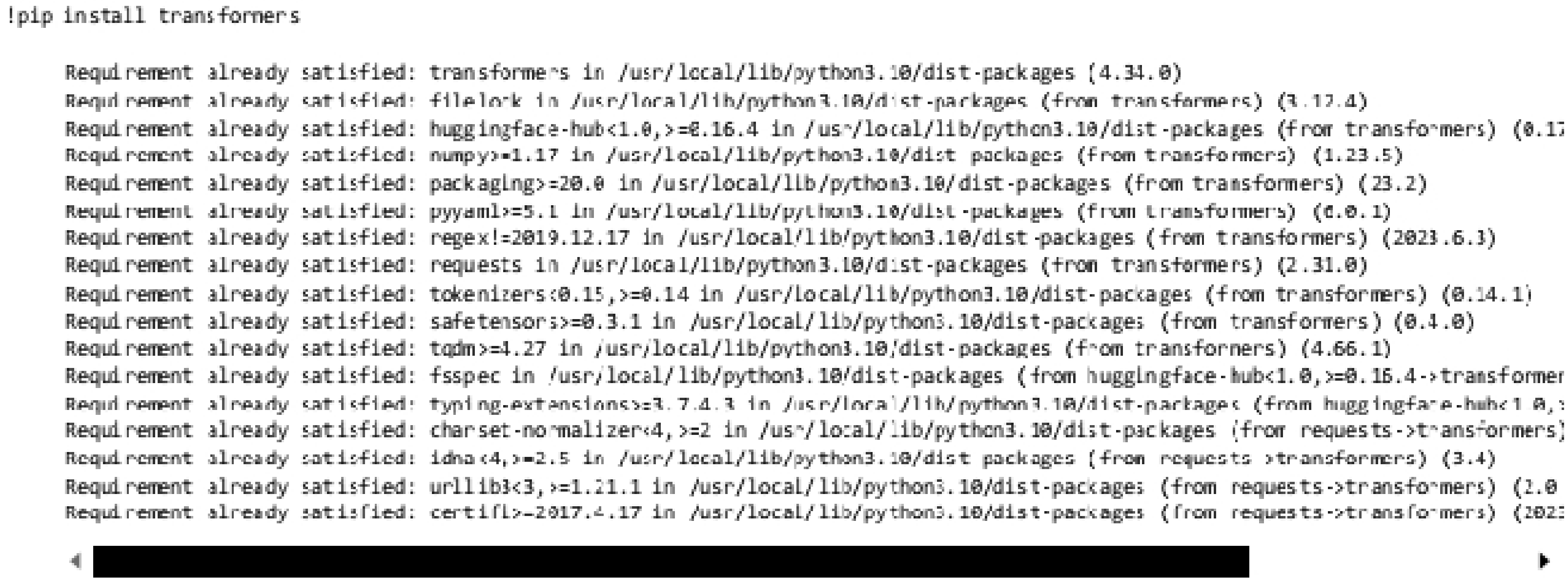
import pandas as pd

import plotly.express as px

import plotly.graph_objs as go

from plotly.subplots import make_subplot

[op]



```
!pip install transformers

     Requirement already satisfied: transformers in /usr/local/lib/python3.10/dist-packages (4.34.0)
     Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from transformers) (3.12.4)
     Requirement already satisfied: huggingface-hub<1.0,>=0.16.4 in /usr/local/lib/python3.10/dist-packages (from transformers) (0.17
     Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-packages (from transformers) (1.23.5)
     Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from transformers) (23.2)
     Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-packages (from transformers) (6.0.1)
     Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10/dist-packages (from transformers) (2023.6.3)
     Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from transformers) (2.31.0)
     Requirement already satisfied: tokenizers<0.15,>=0.14 in /usr/local/lib/python3.10/dist-packages (from transformers) (0.14.1)
     Requirement already satisfied: safetensors>=0.3.1 in /usr/local/lib/python3.10/dist-packages (from transformers) (0.4.0)
     Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.10/dist-packages (from transformers) (4.66.1)
     Requirement already satisfied: fsspec in /usr/local/lib/python3.10/dist-packages (from huggingface-hub<1.0,>=0.16.4->transformer
     Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub<1.0,:
     Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->transformers)
     Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->transformers) (3.4)
     Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests->transformers) (2.0
     Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->transformers) (2023
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.10/dist-packages (from transformers) (4.66.1)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub<1.0,>=0.16.4->t
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub<1.0,>
Collecting huggingface-hub<1.0,>=0.16.4 (from transformers)
  Downloading huggingface_hub-0.17.3-py3-none-any.whl (295 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 295.0/295.0 kB 31.9 MB/s eta 0:00:00
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->transformers)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->transformers) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests->transformers) (2.0
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->transformers) (202
Installing collected packages: safetensors, huggingface-hub, tokenizers, transformers
Successfully installed huggingface-hub-0.17.3 safetensors-0.4.0 tokenizers-0.14.1 transformers-4.34.0
```

```
!pip install transformers
```

```
Requirement already satisfied: transformers in /usr/local/lib/python3.10/dist-packages (4.34.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from transformers) (3.12.4)
Requirement already satisfied: huggingface-hub<1.0,>=0.16.4 in /usr/local/lib/python3.10/dist-packages (from transformers) (0.17
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-packages (from transformers) (1.23.5)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from transformers) (23.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-packages (from transformers) (6.0.1)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10/dist-packages (from transformers) (2023.6.3)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from transformers) (2.31.0)
Requirement already satisfied: tokenizers<0.15,>=0.14 in /usr/local/lib/python3.10/dist-packages (from transformers) (0.14.1)
Requirement already satisfied: safetensors>=0.3.1 in /usr/local/lib/python3.10/dist-packages (from transformers) (0.4.0)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.10/dist-packages (from transformers) (4.66.1)
Requirement already satisfied: fsspec in /usr/local/lib/python3.10/dist-packages (from huggingface-hub<1.0,>=0.16.4->transformer
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub<1.0,>
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->transformers)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->transformers) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests->transformers) (2.0
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->transformers) (2023
```

SPLIT THE DATA INTO TRAINING AND TESTING SETS:

Divide the data into training and testing sets for model development and evaluation.

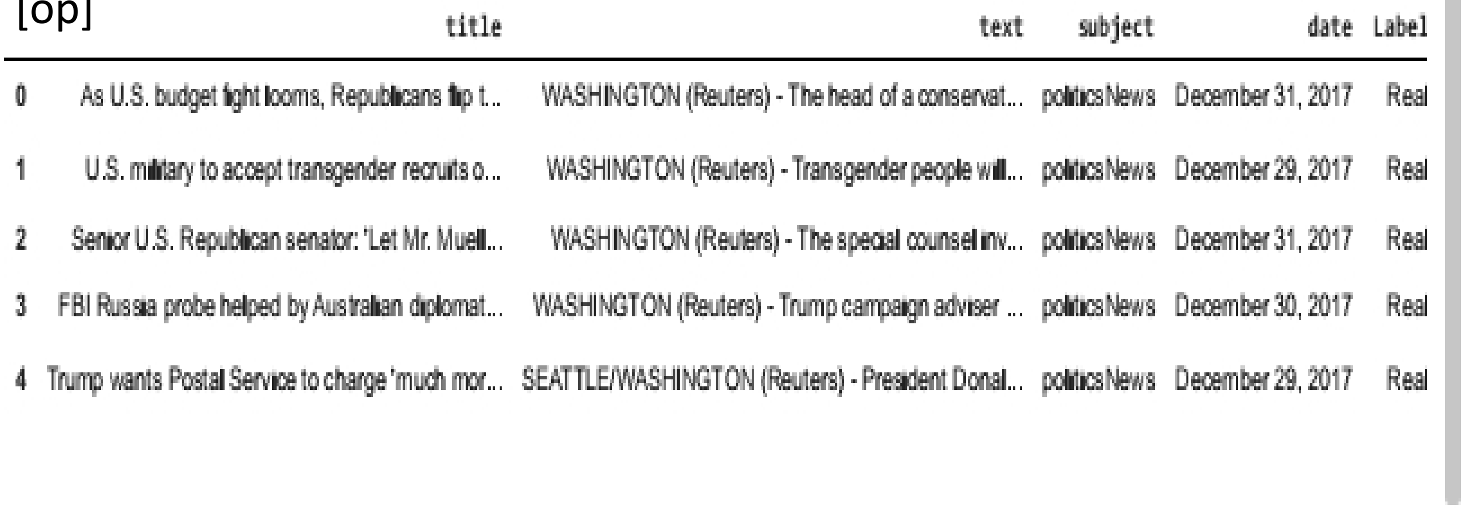
```
import nltk from nltk.corpus
import stopwords
import tensorflow as tf
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.callbacks import ModelCheckpoint
from sklearn.model_selection import train_test_split
```

```
!pip install nltk
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages
(3.8.1) Requirement already satisfied: click in /usr/local/lib/python3.10/dist-
packages (from nltk) (8.1.7) Requirement already satisfied: joblib in
/usr/local/lib/python3.10/dist-packages (from nltk) (1.3.2) Requirement
```

```
fake_news = pd.read_csv('/content/Fake.csv')

real_news = pd.read_csv('/content/True.csv')

fake_news.head(3)

real = real_news.copy()

fake = fake_news.copy()
real['Label'] = 'Real‘

fake['Label'] = 'Fake'
news = pd.concat([real, fake], axis=0,

ignore  index=True) news.reset  index()

new
```

[op]



	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	Real
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	Real
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	Real
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	Real
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	Real

```
Samples available: 44898

#features of dataset: 5
```

```
news_ds = news.sample(1000).drop(['title', 'date', 'subject'], axis=1)

news_ds.head(3)
```

```
 [op]                                                     text          label
   14423   RAMALLAH/WASHINGTON (Reuters) - Palestinian of...     Real
   18027   ZURICH (Reuters) - Switzerland on Sunday calle...     Real
   39100   Is there a more corrupt and power hungry group...     Fake
```

```
CLASS_NAMES = ['Fake', 'Real']

class_mapper = { 'Fake':0, 'Real':1 }

news_ds['Label'] =

news_ds['Label'].map(class_mapper)

news_ds.head(10)
```

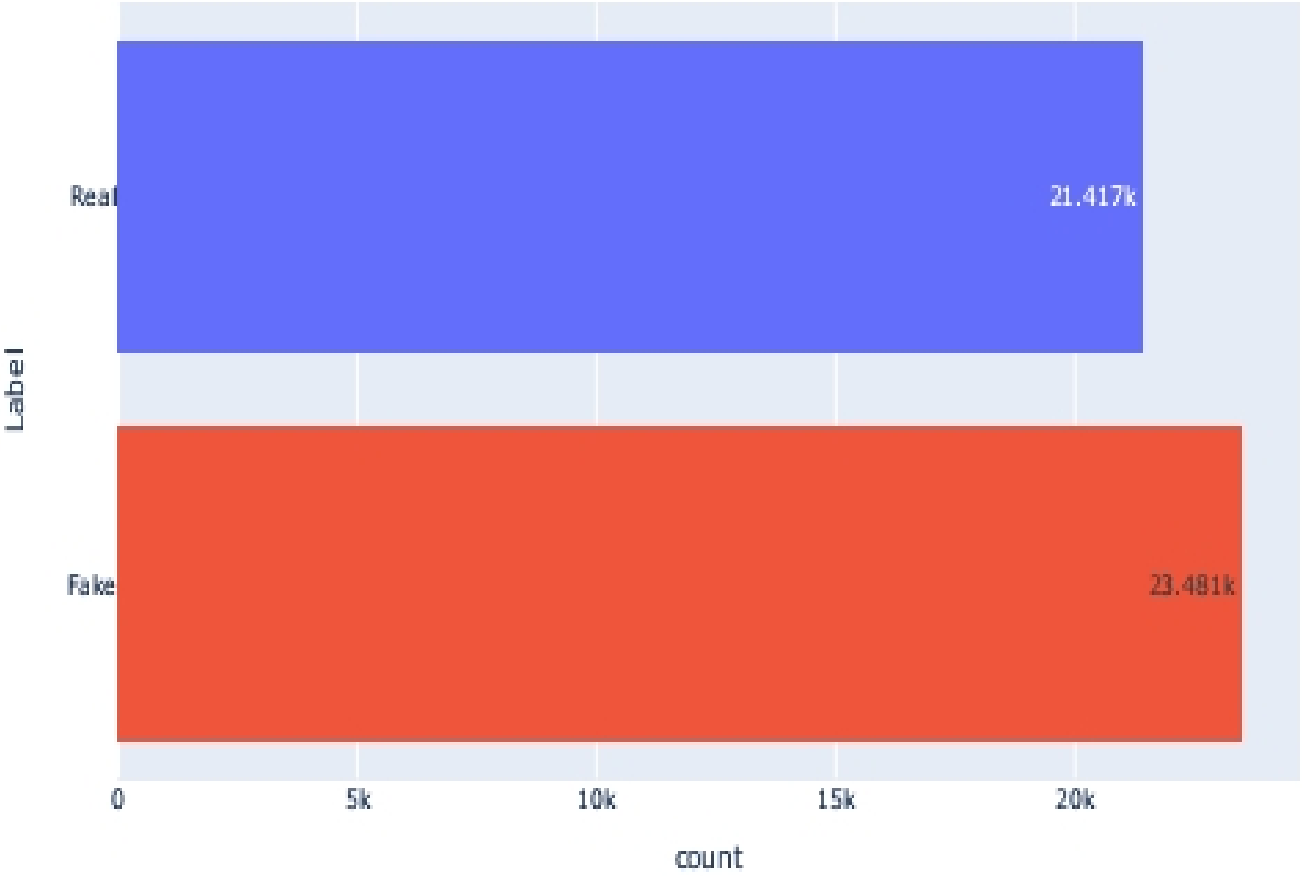
```
[op]
                                                    Text   label1
14423  RAMALLAH/WASHINGTON (Reuters) - Palestinian of...      NaN
18027  ZURICH (Reuters) - Switzerland on Sunday calle...      NaN
39100  Is there a more corrupt and power hungry group...      NaN
22351  Leaked text messages between the daughters of ...      NaN
11652  BEIJING/TAIPEI (Reuters) - A Beijing court on ...      NaN
29114  When Donald Trump isn t bragging about the siz...      NaN
14935  BERLIN (Reuters) - Environmental policy domina...      NaN
15391  HONG KONG (Reuters) - Some activists in Hong K...      NaN
7089   (Reuters) - A federal judge on Tuesday blocked...      NaN
9568   NEW YORK (Reuters) - Puerto Rico’s debt crisis...      NaN
```

MODEL THE DATA

Data Modeling in software engineering is the process of simplifying the diagram or data model of a software system by applying certain formal techniques. It involves expressing data and information through text and symbols

```
class_dist = px.histogram(data_frame=news,
                          y='Label',
                          color='Label‘,
                          title='Fake vs Real news Original dataset',
                          text_auto=True)
class_dist.update_layout(showlegend=False)
 class_dist.show()
```

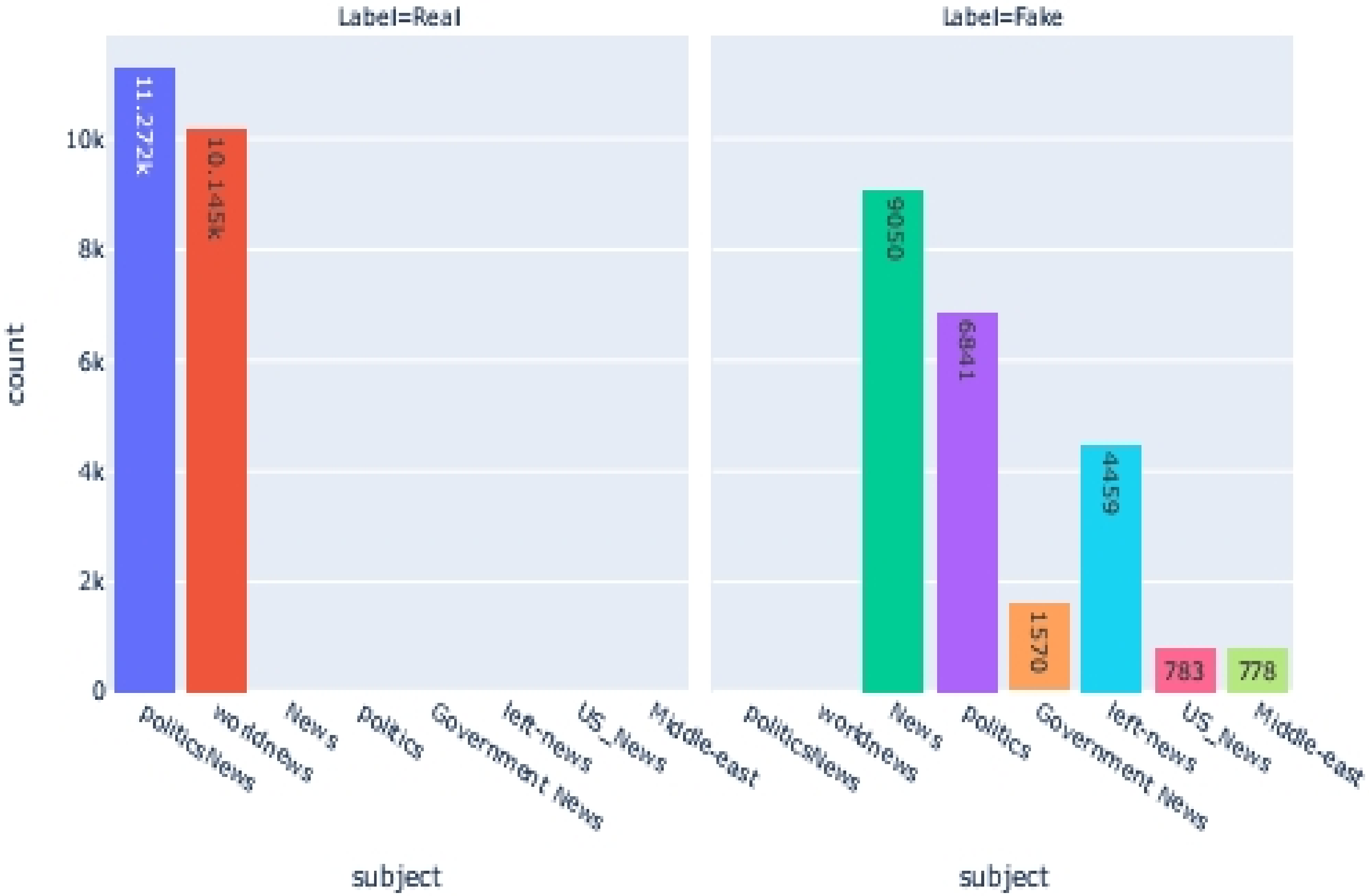
[op]



```
from google.colab import drive

drive.mount('/content/drive')

subject_dist = px.histogram(data_frame=news,

                                    x='subject',

                                color='subject',

                                title='Fake vs Real news Subject Distribution',

text_auto=True,

facet_col='Label')

subject_dist.update_layout(showlegend=False)

subject_dist.show()
```

[op]

FAKE VS REAL NEWS SUBJECT DISTRIBUTION


Label=Real
Label=Fake
count
2k
4k
6k
8k
10k
11.272k
10.145k
politicsNews
worldnews
News
politics
Government News
left-news
US_News
Middle-east
politicsNews
worldnews
News
politics
Government News
left-news
US_News
Middle-east
subject
subject