FAKE NEWS DETECTION USING NLP IN ARTIFICIAL INTELLIGENCE

TEAM MEMBER

au820421205071: SUBANU R S

Phase-4 SUBMISSION DOCUMENT

Project: Fake News Detection

Phase 4: ***Development Part 2***

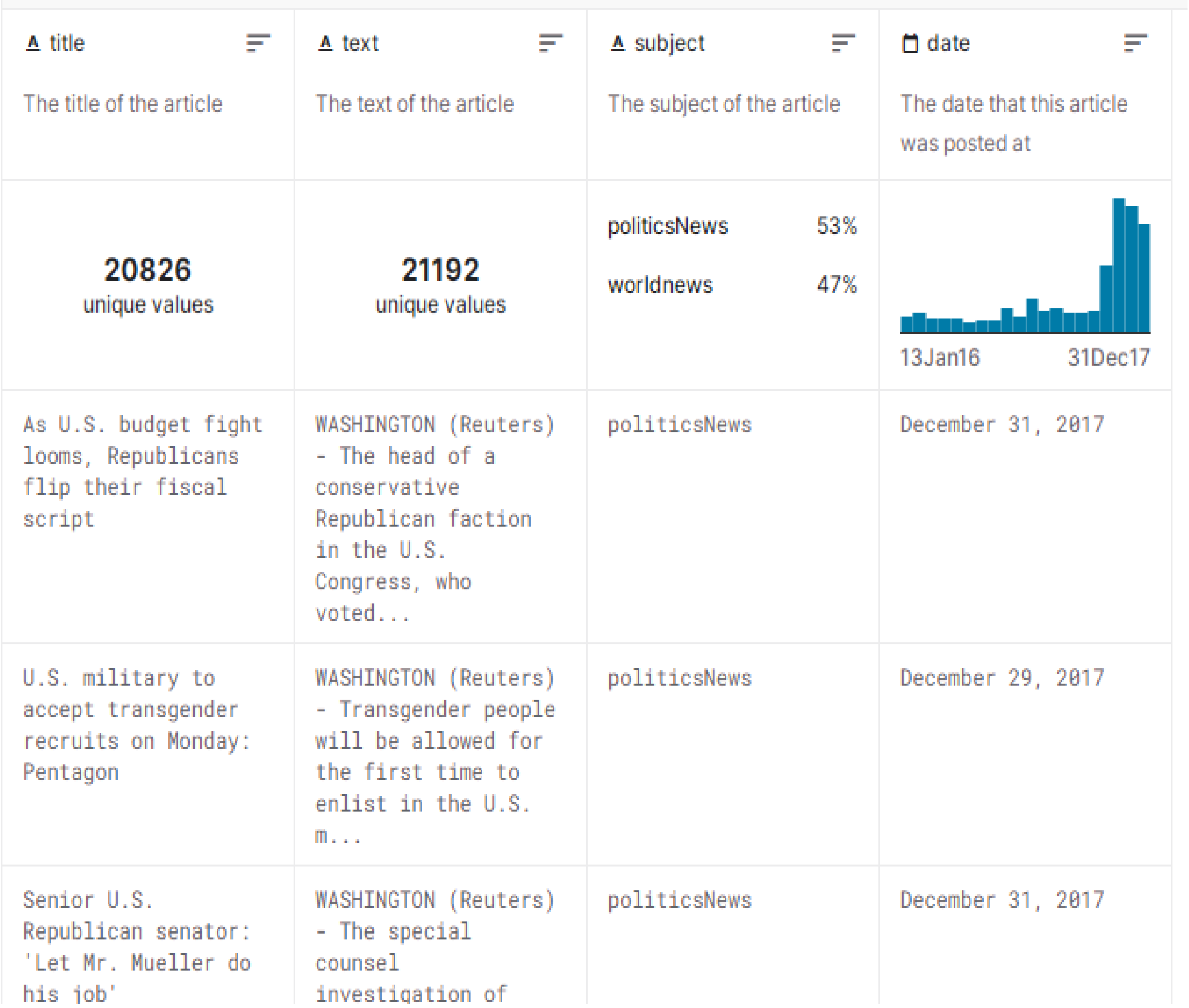
PROBLEM STATEMENT

Continue using NLP techniques to train a classification model in order to develop the false news detection model. Training and assessing models for text preprocessing and feature extraction

GIVEN DATASET

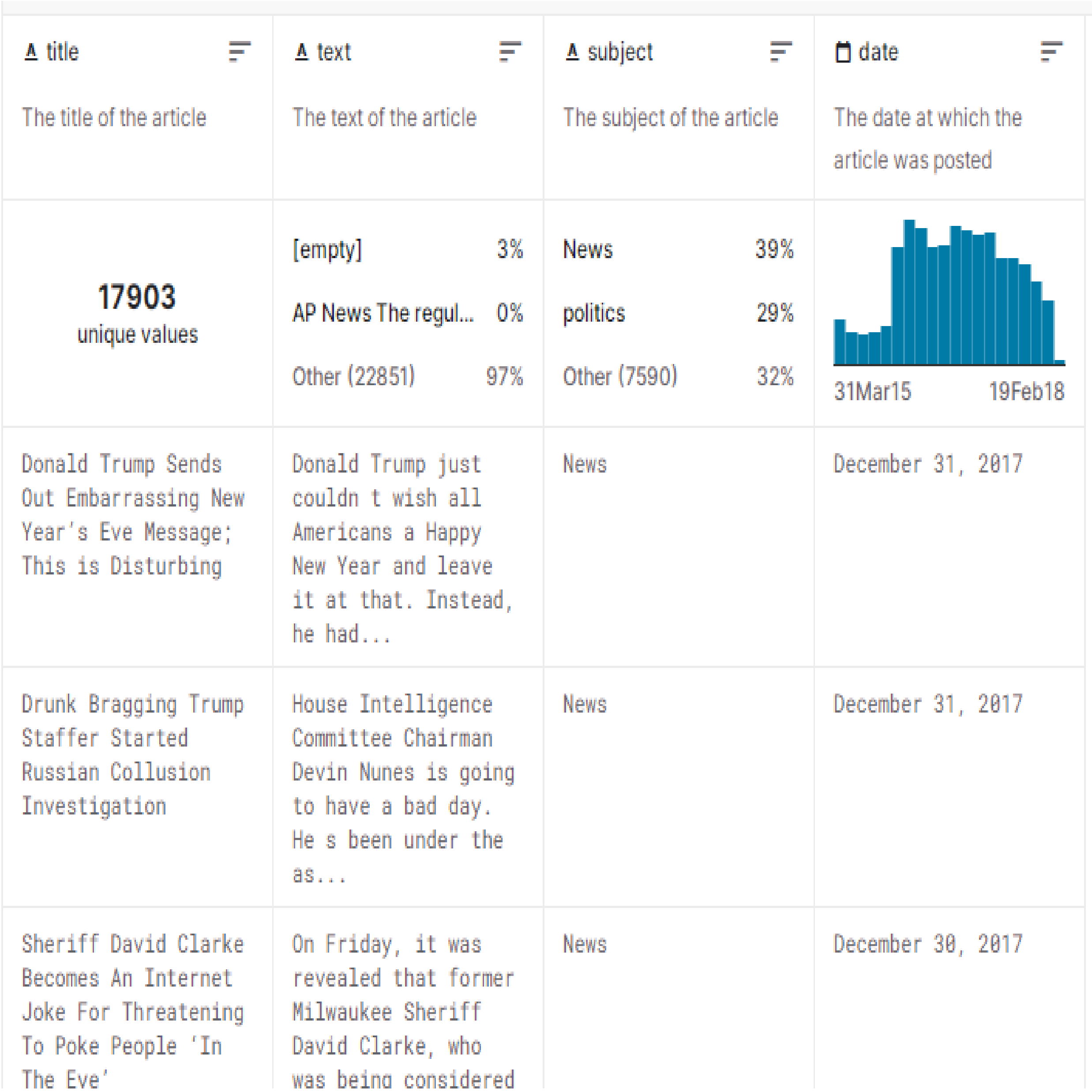
Dataset Link: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

real.csv



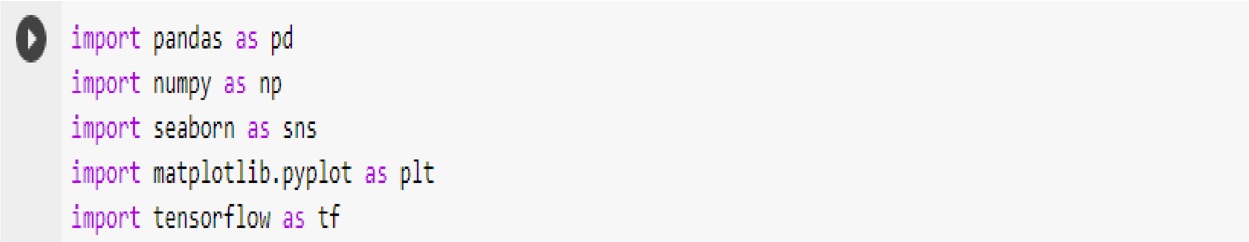
title	text	subject	date
The title of the article	The text of the article	The subject of the article	The date that this article was posted at
20826 unique values	21192 unique values	politicsNews 53% worldnews 47%	13Jan16 31Dec17
As U.S. budget fight looms, Republicans flip their fiscal script	WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, who voted...	politicsNews	December 31, 2017
U.S. military to accept transgender recruits on Monday: Pentagon	WASHINGTON (Reuters) - Transgender people will be allowed for the first time to enlist in the U.S. m...	politicsNews	December 29, 2017
Senior U.S. Republican senator: 'Let Mr. Mueller do his job'	WASHINGTON (Reuters) - The special counsel investigation of	politicsNews	December 31, 2017

fake.csv

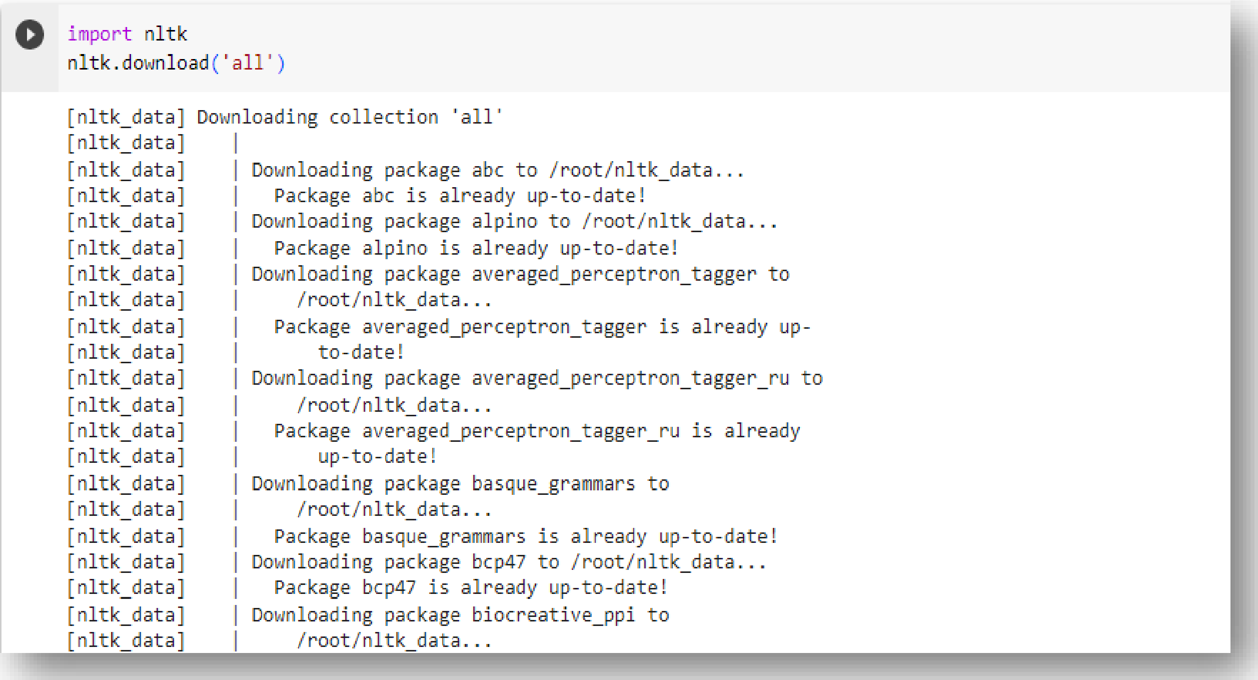


title	text	subject	date
The title of the article	The text of the article	The subject of the article	The date at which the article was posted
17903 unique values	[empty] 3% AP News The regul... 0% Other (22851) 97%	News 39% politics 29% Other (7590) 32%	31Mar15 19Feb18
Donald Trump Sends Out Embarrassing New Year’s Eve Message; This is Disturbing	Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he had...	News	December 31, 2017
Drunk Bragging Trump Staffer Started Russian Collusion Investigation	House Intelligence Committee Chairman Devin Nunes is going to have a bad day. He s been under the as...	News	December 31, 2017
Sheriff David Clarke Becomes An Internet Joke For Threatening To Poke People ‘In The Eve’	On Friday, it was revealed that former Milwaukee Sheriff David Clarke, who was being considered	News	December 30, 2017

IMPORTING LIBRARIES



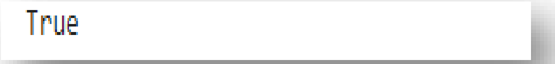
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import tensorflow as tf
```



```
import nltk
nltk.download('all')
```

```
[nltk_data] Downloading collection 'all'
[nltk_data]    |
[nltk_data]    | Downloading package abc to /root/nltk_data...
[nltk_data]    |   Package abc is already up-to-date!
[nltk_data]    | Downloading package alpino to /root/nltk_data...
[nltk_data]    |   Package alpino is already up-to-date!
[nltk_data]    | Downloading package averaged_perceptron_tagger to
[nltk_data]    |     /root/nltk_data...
[nltk_data]    |   Package averaged_perceptron_tagger is already up-
[nltk_data]    |       to-date!
[nltk_data]    | Downloading package averaged_perceptron_tagger_ru to
[nltk_data]    |     /root/nltk_data...
[nltk_data]    |   Package averaged_perceptron_tagger_ru is already
[nltk_data]    |       up-to-date!
[nltk_data]    | Downloading package basque_grammars to
[nltk_data]    |     /root/nltk_data...
[nltk_data]    |   Package basque_grammars is already up-to-date!
[nltk_data]    | Downloading package bcp47 to /root/nltk_data...
[nltk_data]    |   Package bcp47 is already up-to-date!
[nltk_data]    | Downloading package biocreative_ppi to
[nltk_data]    |     /root/nltk_data...
```

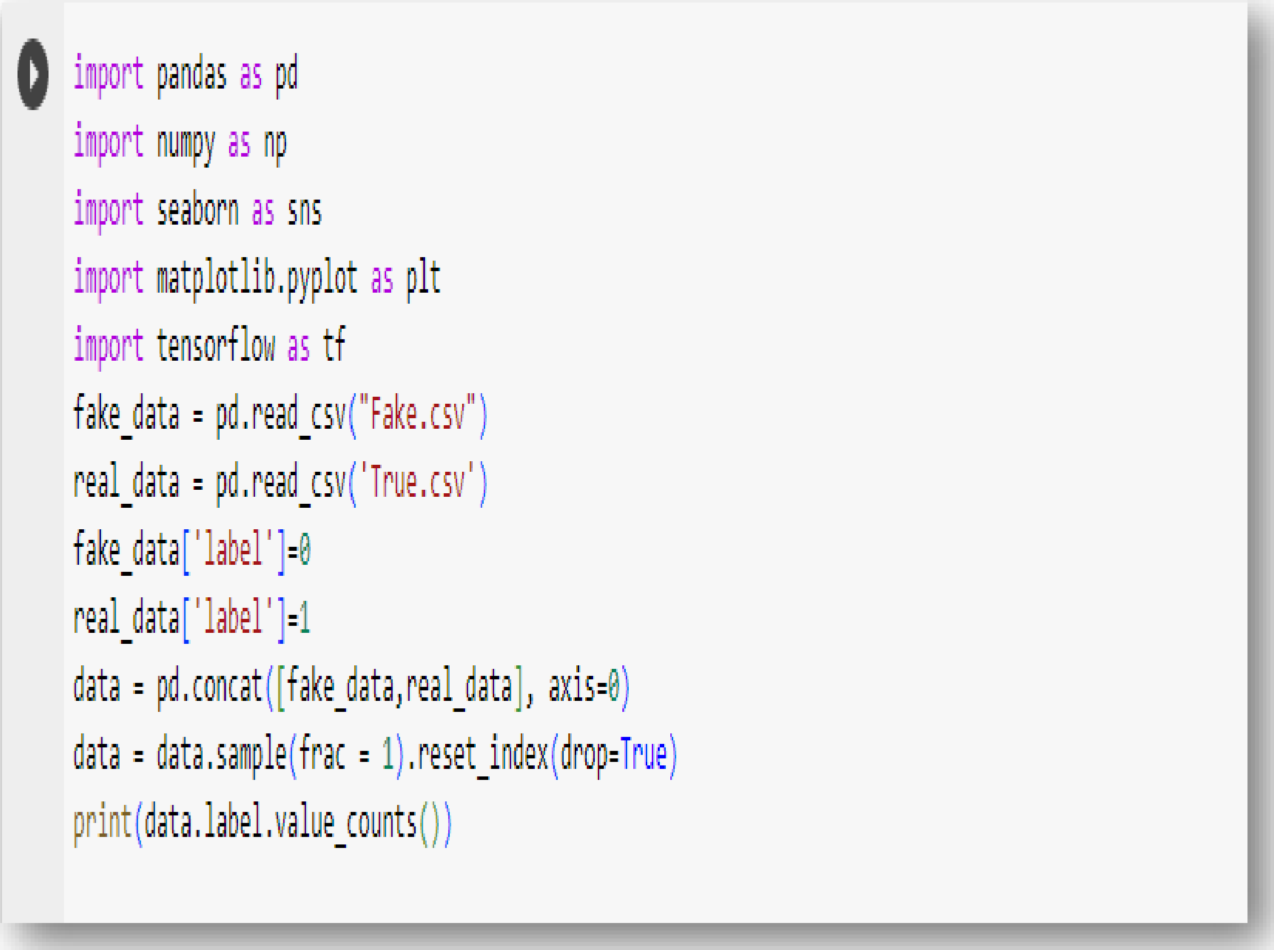
OUTPUT:



```
True
```

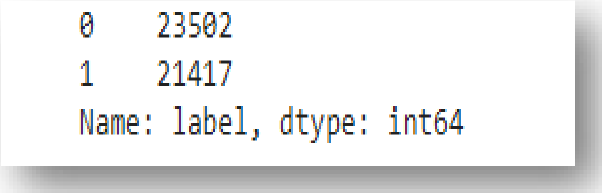
LABELING:

PROGRAM:



```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import tensorflow as tf
fake_data = pd.read_csv("Fake.csv")
real_data = pd.read_csv('True.csv')
fake_data['label']=0
real_data['label']=1
data = pd.concat([fake_data,real_data], axis=0)
data = data.sample(frac = 1).reset_index(drop=True)
print(data.label.value_counts())
```

OUTPUT:

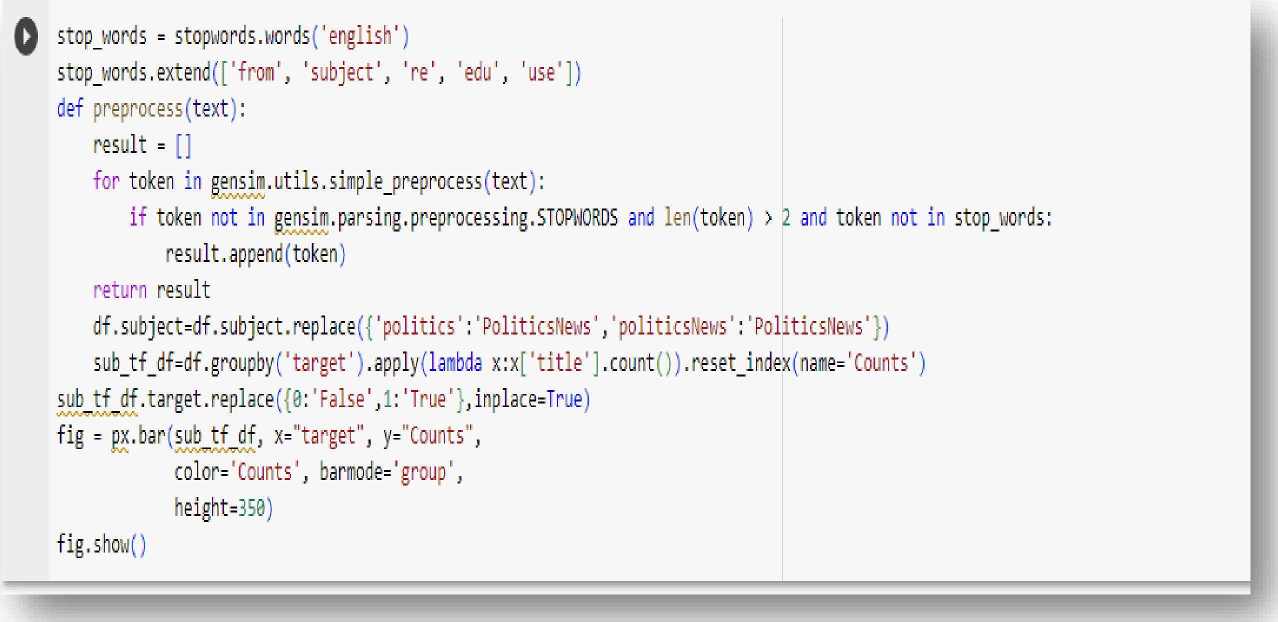


```
0    23502
1    21417
Name: label, dtype: int64
```

DATA CLEANUP

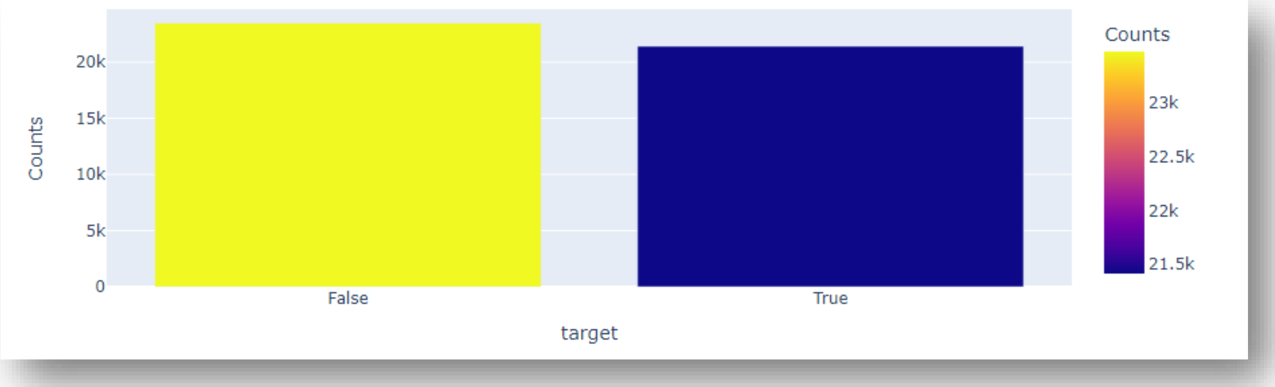
Cleaning up data in fake news detection using natural language processing (NLP) is a critical step in building an effective model for identifying fake news.

PROGRAM:



```
stop_words = stopwords.words('english')
stop_words.extend(['from', 'subject', 're', 'edu', 'use'])
def preprocess(text):
    result = []
    for token in gensim.utils.simple_preprocess(text):
        if token not in gensim.parsing.preprocessing.STOPWORDS and len(token) > 2 and token not in stop_words:
            result.append(token)
    return result
    df.subject=df.subject.replace({'politics':'PoliticsNews','politicsNews':'PoliticsNews'})
    sub_tf_df=df.groupby('target').apply(lambda x:x['title'].count()).reset_index(name='Counts')
sub_tf_df.target.replace({0:'False',1:'True'},inplace=True)
fig = px.bar(sub_tf_df, x="target", y="Counts",
             color='Counts', barmode='group',
             height=350)
fig.show()
```

OUTPUT:



TEXTPREPROCESSING

Detecting fake news using Natural Language Processing (NLP) techniques is a critical task, requiring text preprocessing to clean and transform raw text into a format suitable for analysis. Key steps include:

1. **Lowercasing:**

   Convert all text to lowercase to ensure uniformity and avoid treating words with different cases as different entities.

2. **Tokenization:**

   It is a crucial process that breaks text into individual words or tokens, enabling further processing like removing stop words and punctuation.

3. **Stop Words Removal:**

   It removes common words from a text, reducing noise and focusing analysis on essential content.(eg., 'and', 'the', 'is', etc.)

4. **Lemmatization and stemming:**

   These are techniques that reduce words to their base or root form, ensuring that different forms of the same word are treated the same.

5. **Vectorization:**

   It is the process of converting preprocessed text data into numerical format for machine learning models, using techniques like bag-of-words, TF-IDF, and word embeddings like Word2Vec or GloVe.

6. **Removing URLs and Email Addresses:**

   Fake news articles often contain irrelevant URLs or email addresses, which can be removed to enhance the focus on the textual content.

7. **Feature engineering:**

   It involves extracting relevant text features like n-grams, TF-IDF, and word embeddings to capture contextual information and word relationships.

MODEL TRAINING

1. **DATA COLLECTION AND PREPARATION:**

The study involves collecting and preparing a diverse dataset of labeled news articles, including both genuine and fake news, and preprocessing it for NLP analysis.

2. **DATA SPLITTING:**

   Data splitting involves dividing preprocessed data into training and testing sets to enable the model to learn patterns from training data and assess its performance on unseen data.

3. **HYPERPARAMETER TUNING:**

The model's performance can be enhanced by fine-tuning its hyperparameters using techniques like grid search or random search to find the optimal combination.

4. **VALIDATION ON TEST SET:**

   The final model must be validated on a testing dataset to ensure its generalizability and robustness, and any unsatisfactory results should be addressed**.**

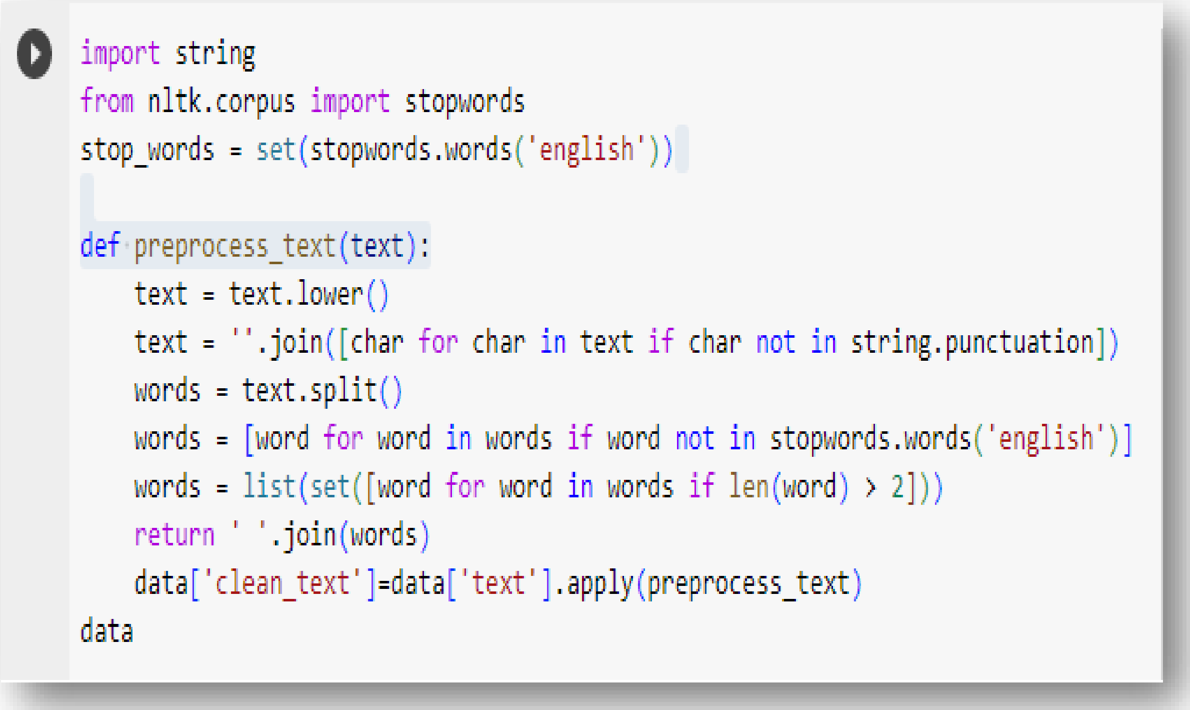
5. **CONTINUOUS IMPROVEMENT:**

   The model's performance is continuously monitored and improved through regular updates, incorporating new data to enhance accuracy and reliability over time.

6. **MODEL DEPLOYMENT:**

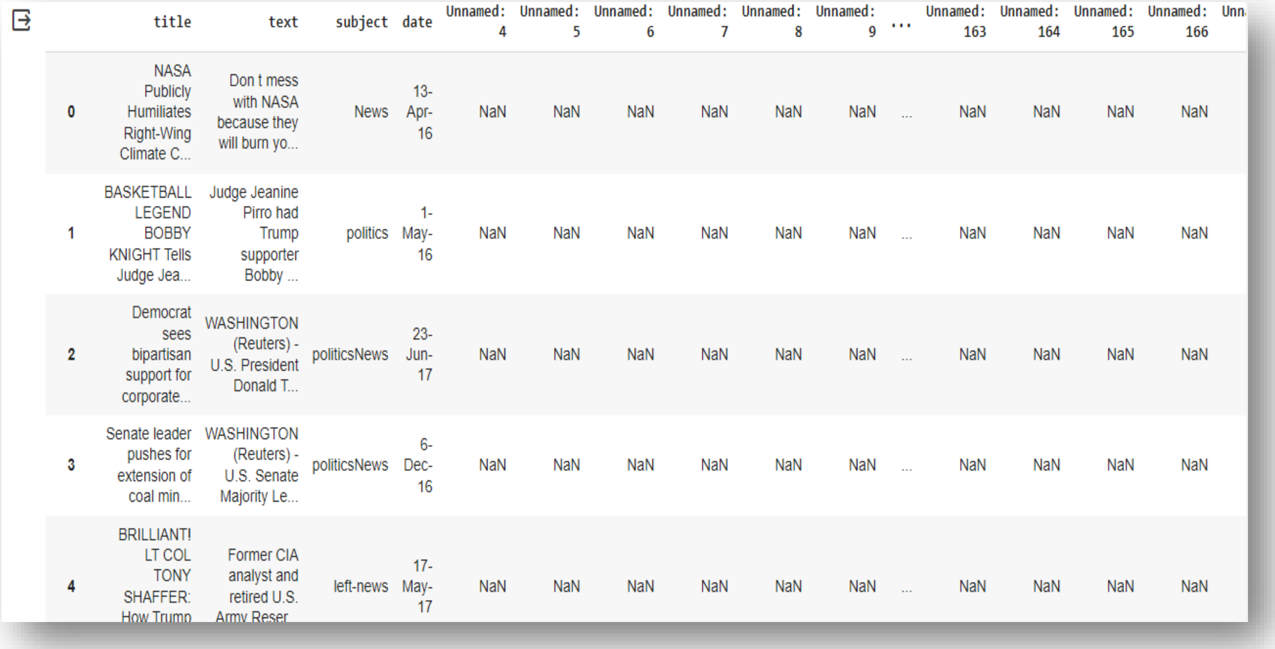
   The trained model is deployed in the production environment for real-time fake news detection, and its performance is monitored and retrained periodically to adapt to data trends.

PROGRAM:



```
import string
from nltk.corpus import stopwords
stop_words = set(stopwords.words('english'))

def preprocess_text(text):
    text = text.lower()
    text = ''.join([char for char in text if char not in string.punctuation])
    words = text.split()
    words = [word for word in words if word not in stopwords.words('english')]
    words = list(set([word for word in words if len(word) > 2]))
    return ' '.join(words)
    data['clean_text']=data['text'].apply(preprocess_text)
data
```

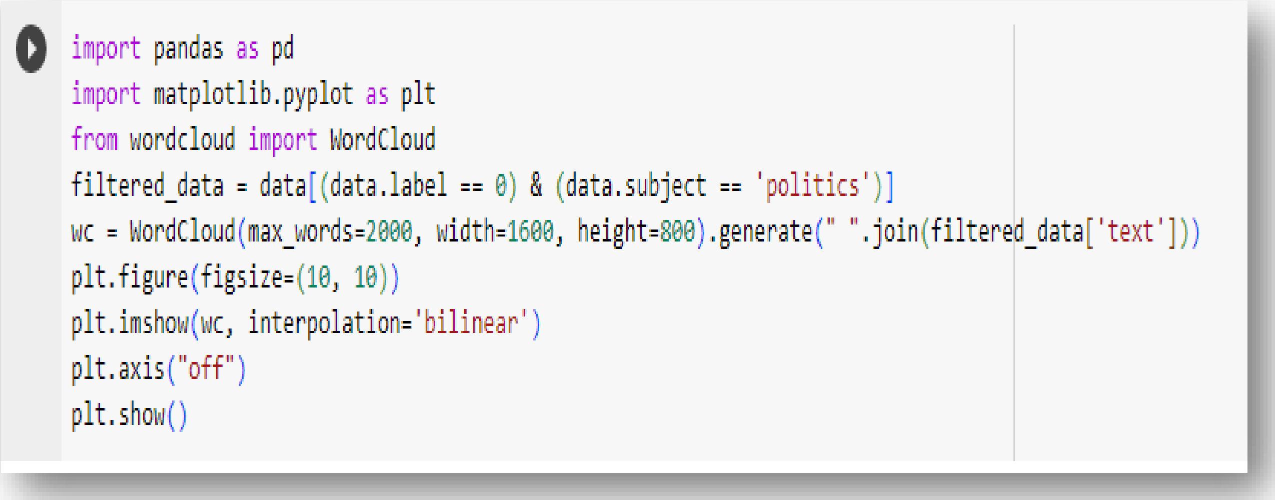
OUTPUT:



	title	text	subject	date	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	...	Unnamed: 163	Unnamed: 164	Unnamed: 165	Unnamed: 166	Unn
0	NASA Publicly Humiliates Right-Wing Climate C...	Don t mess with NASA because they will burn yo...	News	13-Apr-16	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	
1	BASKETBALL LEGEND BOBBY KNIGHT Tells Judge Jea...	Judge Jeanine Pirro had Trump supporter Bobby ...	politics	1-May-16	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	
2	Democrat sees bipartisan support for corporate...	WASHINGTON (Reuters) - U.S. President Donald T...	politicsNews	23-Jun-17	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	
3	Senate leader pushes for extension of coal min...	WASHINGTON (Reuters) - U.S. Senate Majority Le...	politicsNews	6-Dec-16	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	
4	BRILLIANT! LT COL TONY SHAFFER: How Trump	Former CIA analyst and retired U.S. Army Reser	left-news	17-May-17	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	

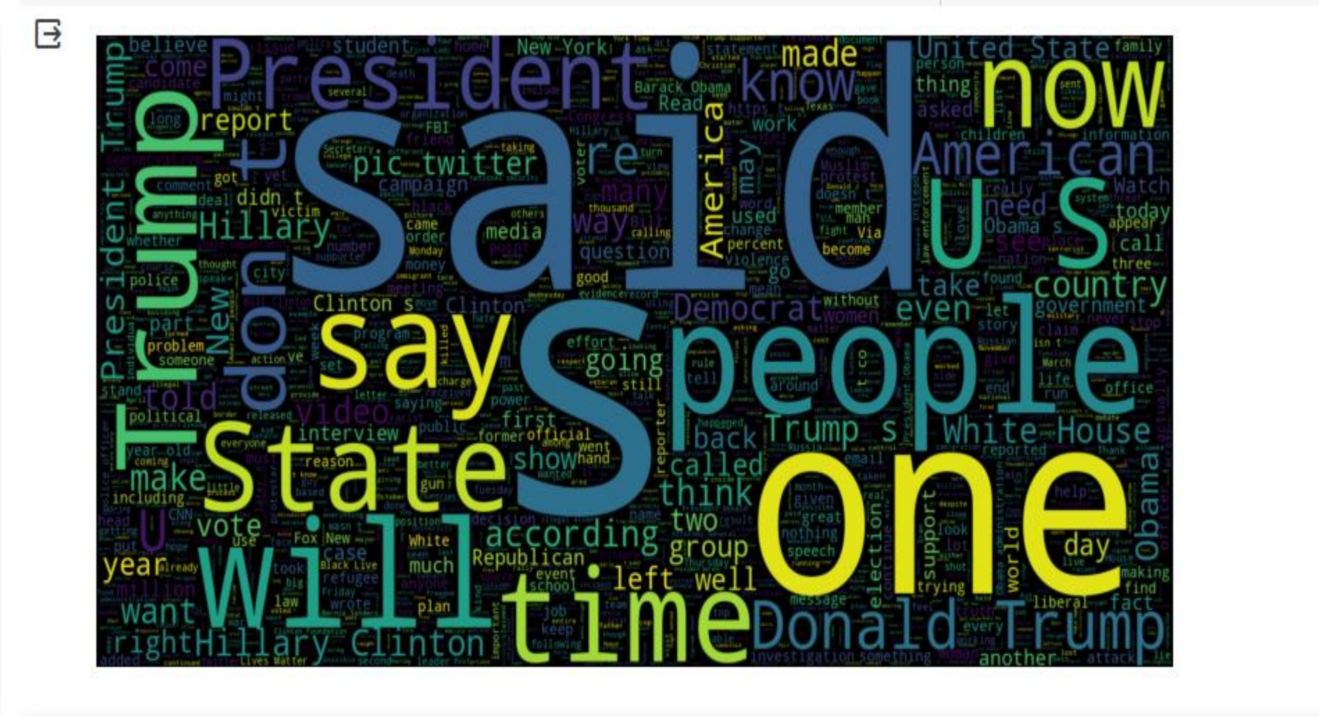
WORDCLOUD FOR REAL NEWS

PROGRAM:



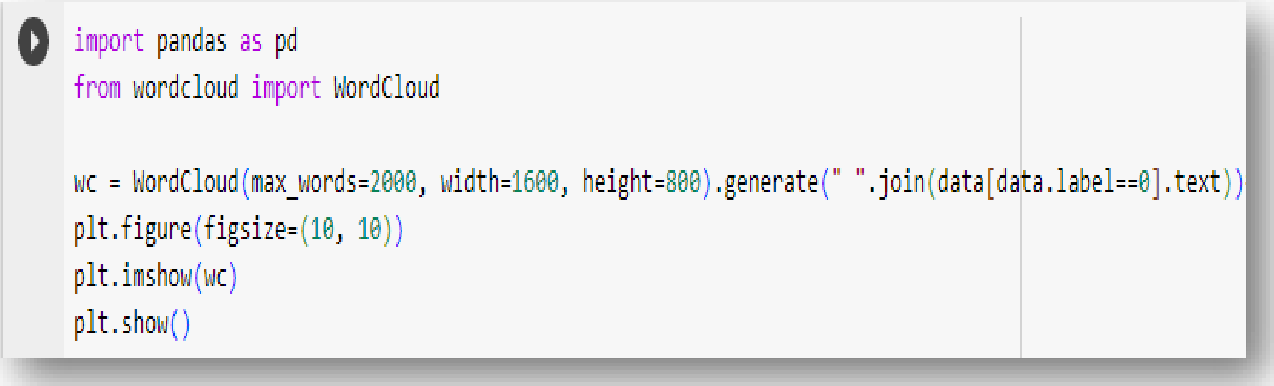
```
import pandas as pd
import matplotlib.pyplot as plt
from wordcloud import WordCloud
filtered_data = data[(data.label == 0) & (data.subject == 'politics')]
wc = WordCloud(max_words=2000, width=1600, height=800).generate(" ".join(filtered_data['text']))
plt.figure(figsize=(10, 10))
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.show()
```

OUTPUT:



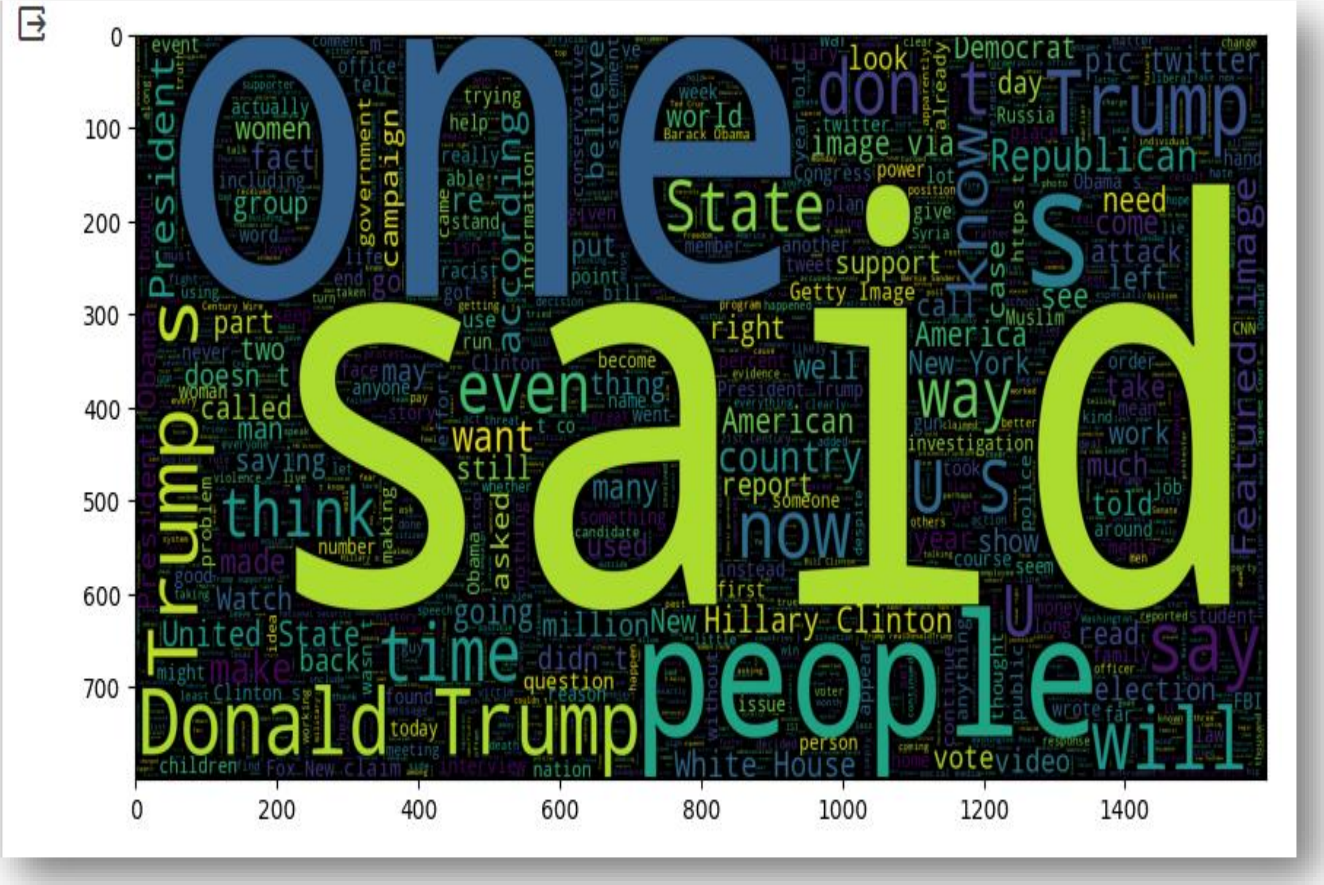
WORLDCLOUD FOR FAKENEWS

PROGRAM:



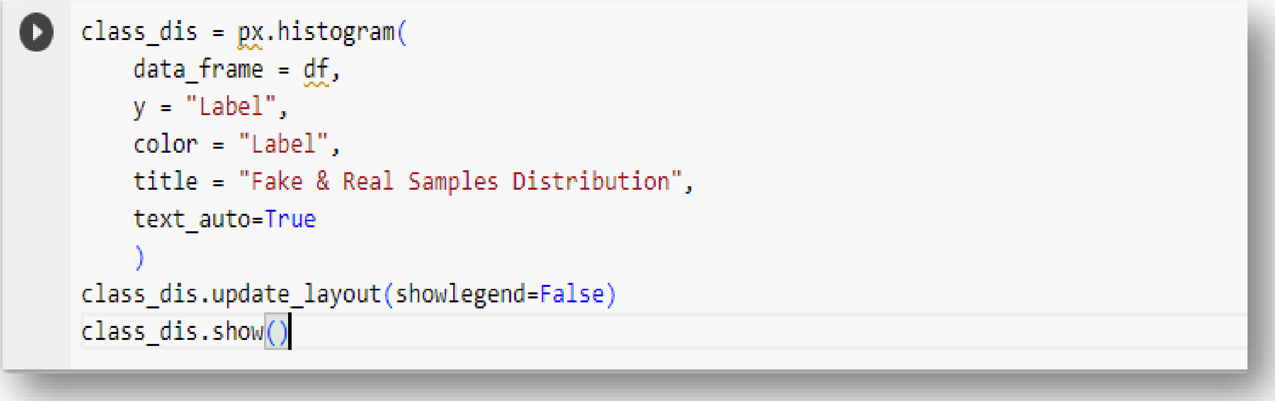
```
import pandas as pd
from wordcloud import WordCloud

wc = WordCloud(max_words=2000, width=1600, height=800).generate(" ".join(data[data.label==0].text))
plt.figure(figsize=(10, 10))
plt.imshow(wc)
plt.show()
```

OUTPUT:



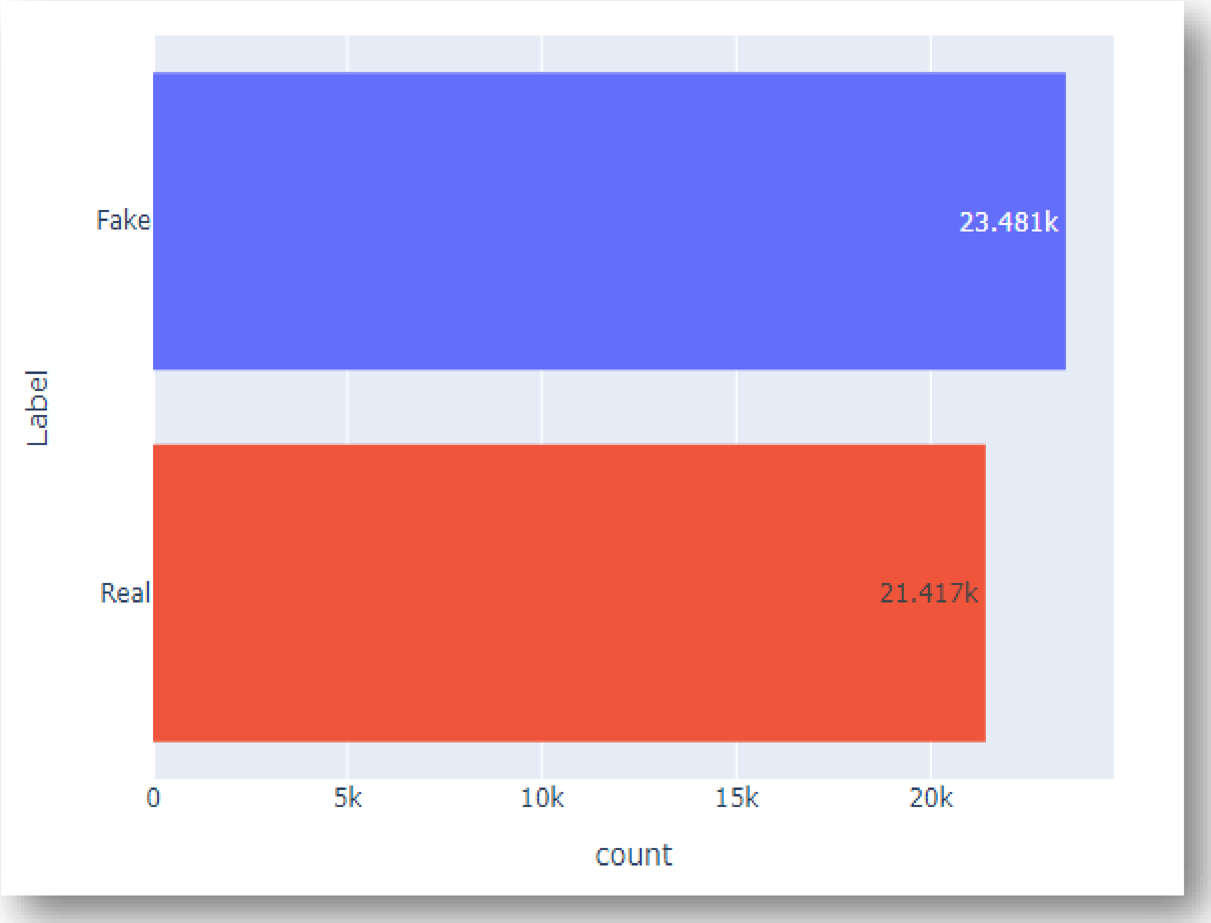
DATA VISUALIZATION

PROGRAM



```
class_dis = px.histogram(
    data_frame = df,
    y = "Label",
    color = "Label",
    title = "Fake & Real Samples Distribution",
    text_auto=True
    )
class_dis.update_layout(showlegend=False)
class_dis.show()
```

OUTPUT:



FEATURE EXTRACTION

Feature extraction is a crucial step in constructing machine learning models for fake news detection using Natural Language Processing (NLP), capturing essential information to distinguish genuine and fake news.

EVALULATION

Evaluation of fake news detection using NLP is crucial for assessing model performance and effectiveness in distinguishing genuine and misleading information using various metrics and techniques.

1. ACCURACY:

The model's overall accuracy, which represents the ratio of correctly classified articles to the total dataset, may not be sufficient in an imbalanced dataset.

2. PRECISION AND RECALL:

The study calculates precision, indicating the proportion of accurately identified fake news articles, and recall, indicating the proportion of truly fake articles in the dataset.

3. CONFUSION MATRIX:

The confusion matrix provides a comprehensive view of a model's performance and helps identify types of errors made by it.

4. CROSS VALIDATION:

Cross-validation techniques like k-fold cross-validation ensure consistent model performance across different dataset subsets, reducing overfitting risk and ensuring consistent performance across different datasets.

5. BIAS AND FAIRNESS ANALYSIS:

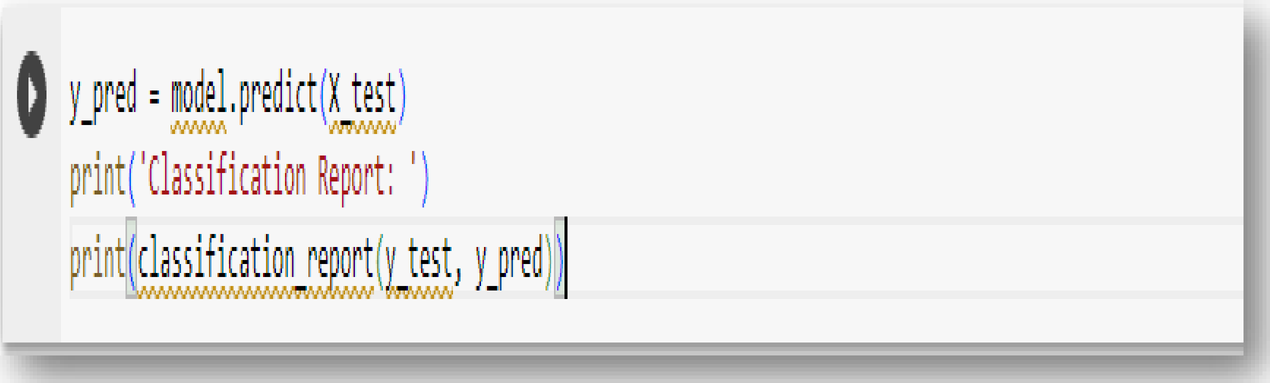
The analysis of the model's bias and fairness is crucial to ensure it does not exhibit any biases towards specific groups or topics, and its predictions are fair and unbiased.

SPECIFICITY AND SENSITIVITY

Calculate specificity (true negative rate) and sensitivity (true positive rate) to measure the proportion of genuine and fake news articles in a dataset.

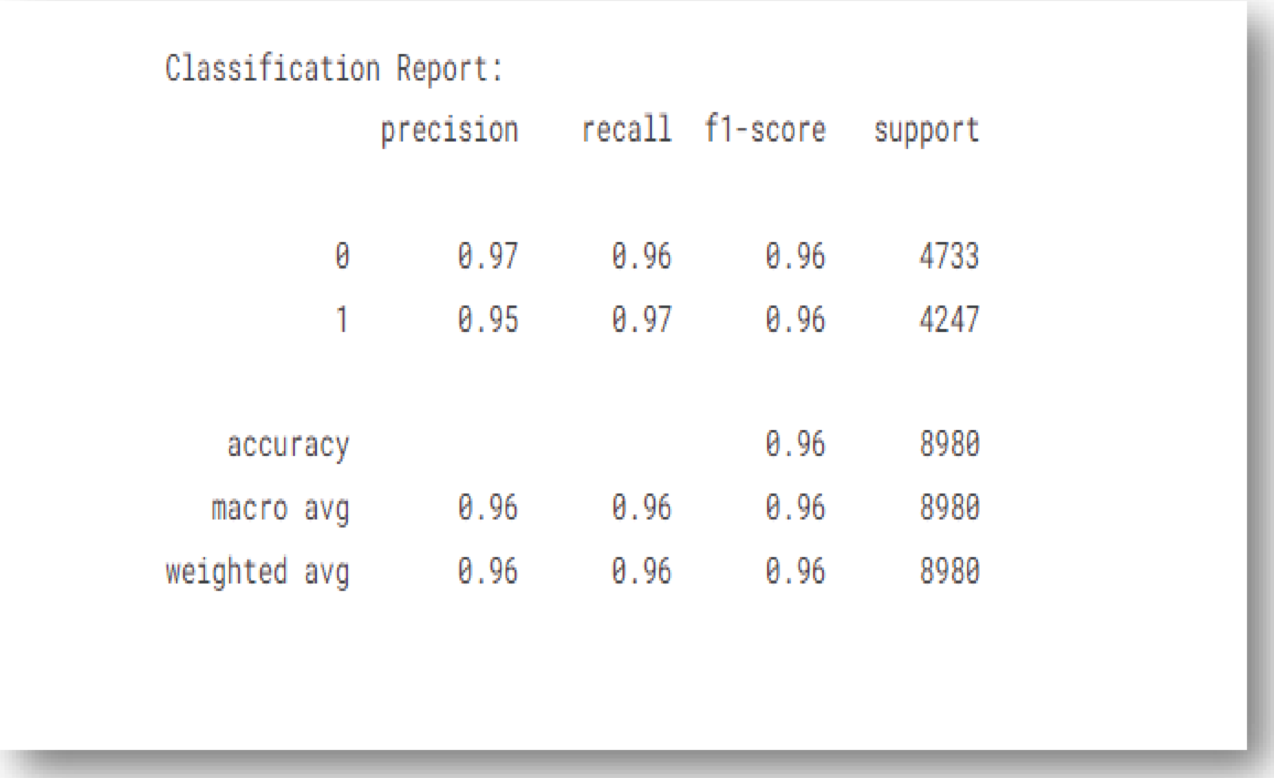
CLASSIFICATION REPORT:

PROGRAM:



```
y_pred = model.predict(X_test)
print('Classification Report: ')
print(classification_report(y_test, y_pred))
```

OUTPUT:



```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.96      0.96      4733
           1       0.95      0.97      0.96      4247

    accuracy                           0.96      8980
   macro avg       0.96      0.96      0.96      8980
weighted avg       0.96      0.96      0.96      8980
```

CONCLUSION

The use of Natural Language Processing (NLP) in detecting fake news is a promising method to combat misinformation. It uses text preprocessing techniques and feature extraction methods to capture linguistic nuances and contextual cues, requiring regular evaluation and refinement for accuracy.