Phase - 3 : Fake News Detection Using NLP

Aim of the Project:

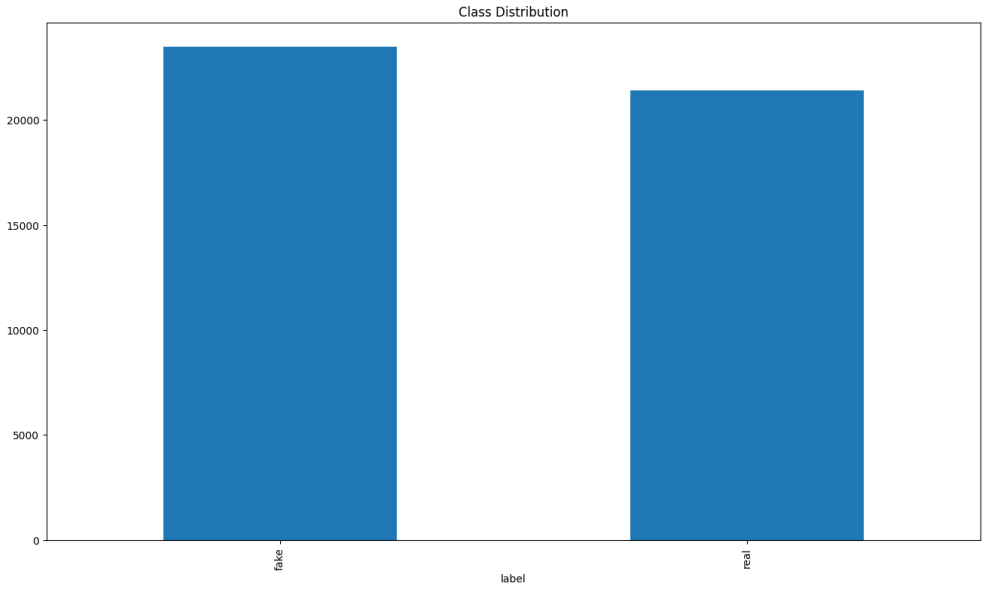
The aim of this project is to develop a Fake News Detection model using Natural Language Processing (NLP) techniques. The primary objective is to distinguish between genuine and fake news articles based on the content of their titles and text. This project involves leveraging NLP techniques for text data preprocessing and classification, enabling accurate identification of fake news articles.

Program:

```
import pandas as pd
import re
import matplotlib.pyplot as plt
import nltk
from sklearn.model_selection import train_test_split
nltk.download('stopwords')
true_data = pd.read_csv('fake-and-real-news-dataset/True.csv')
fake_data = pd.read_csv('fake-and-real-news-dataset/Fake.csv')
true_data['label'] = 'real'
fake_data['label'] = 'fake'
data = pd.concat([true_data, fake_data])
data['text'] = data['text'].str.lower()
data['text'] = data['text'].apply(lambda x: re.sub(r'[^\w\s]', '', x))
from nltk.corpus import stopwords
stop_words = set(stopwords.words('english'))
data['text'] = data['text'].apply(lambda x: ' '.join([word for word in x.split() if word not in
stop_words]))
data['text'] = data['text'].apply(lambda x: x.split())
X = data['text']
y = data['label']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
data['label'].value_counts().plot(kind='bar')
```

```
plt.title('Class Distribution')
plt.show()
data.to_csv('preprocessed_data.csv', index=False)
```

Output:



Clarifications:

- In this phase, we download the "stopwords" resource using NLTK to perform data preprocessing, which includes converting text to lowercase, removing punctuation and numbers, removing stopwords, and tokenizing the text.

- The code assigns labels ("fake" or "real") based on the "subject" column, creating a binary classification scheme for the dataset.

- The dataset is split into training and testing sets for model development and evaluation.

- The class distribution is visualized to understand the balance between "fake" and "real" articles in the dataset.

- The preprocessed data is saved to 'preprocessed_data.csv' for use in the next project phases.

Conclusion:

In Phase 3, we successfully loaded and preprocessed the dataset, making it ready for the development of our Fake News Detection model. By assigning labels and cleaning the textual data, we've laid the foundation for the model to learn and make accurate predictions. The next phase will involve model selection, training, and evaluation to achieve our project's primary objective of distinguishing between fake and real news articles.