Fake News Detection

Phase-5 Document Submission

Project: Fake News Detection using NLP

Team Members:

Logeshwaran M - 110521106012 [Team leader]

Gunasegkaran S V - 110521106311

Karthi M - 110521106312

Shalini R - 110521106024

Overview:

Fake news detection with NLP and Logistic Regression involves building a model that classifies news articles or text data as either "fake" or "real" based on linguistic and textual features. The method combines the strengths of NLP for text analysis and Logistic Regression for binary classification. The main steps in this process are as follows:

Effects of Fake News:

Fake news can misinform individuals, leading them to believe and spread false information. This can have serious consequences, especially when the misinformation relates to health, politics, or other critical topics.The spread of fake news can erode public trust in traditional media sources. When people are unable to distinguish between credible journalism and fabricated content, it can lead to a general distrust of all news.Companies and individuals may be

financially harmed by fake news, as false information can affect stock prices, reputations, and consumer behavior.

Benefits of Fake News Detection:

- Fake news detection helps maintain the accuracy of information, ensuring that individuals have access to truthful and reliable sources of information.
- The ability to identify fake news contributes to the protection of public trust in the media, as individuals can have confidence in the credibility of news sources.
- Fake news detection reduces the spread of misinformation, which is essential for addressing public health crises, political decision-making, and other important matters.
- By identifying and debunking false information, fake news detection can prevent unnecessary panic and fear among the public during times of crisis.
- Fake news detection contributes to economic stability by preventing false information from affecting stock prices, consumer behaviours, and business operations.
- The focus on fake news detection encourages responsible and ethical journalism practices, promoting transparency and accountability.

Challenges:

- Obtaining a high-quality labelled dataset with accurate labels for fake and real news is often difficult.

- Fake news datasets often have imbalanced class distributions, with far fewer fake news examples than real news.
- Choosing the right features, including NLP techniques like TF-IDF or word embeddings, can be complex. Selecting irrelevant or redundant features may affect model performance.
- Determining the right evaluation metrics can be tricky. Accuracy may not be sufficient, especially in imbalanced datasets.
- Deploying a fake news detection model in real-time requires efficient preprocessing, model serving, and integration with existing systems.

Reason for choosing this algorithm:

Logistic Regression is a classification algorithm used to find the probability of event success and event failure. It is used when the dependent variable is binary(0/1, True/False, Yes/No) in nature. It supports categorising data into discrete classes by studying the relationship from a given set of labelled data. It learns a linear relationship from the given dataset and then introduces a non-linearity in the form of the Sigmoid function.

Logistic regression is also known as Binomial logistics regression. It is based on Sigmoid function where output is probability and input can be from -infinity to +infinity.

Advantages of Logistic regression:

- Logistic regression is easier to implement, interpret, and very efficient to train.

- It makes no assumptions about distributions of classes in feature space.
- It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions.
- It not only provides a measure of how appropriate a predictor(coefficient size)is, but also its direction of association (positive or negative).
- It is very fast at classifying unknown records.
- Good accuracy for many simple data sets and it performs well when the dataset is linearly separable.
- It can interpret model coefficients as indicators of feature importance.
- Logistic regression is less inclined to over-fitting but it can overfit in high dimensional datasets.One may consider Regularization (L1 and L2) techniques to avoid over-fitting in these scenarios.

Steps:

Step1:

Data Collection and Preprocessing:

- Gather a labelled dataset of news articles or text data. Ensure that the dataset is correctly labelled as "fake" or "real."
- Preprocess the text data by removing stopwords, punctuation, and applying techniques like tokenization and stemming/lemmatization.

Step-2:

Feature Extraction:

- Convert the preprocessed text data into numerical features that can be used by the Logistic Regression model. Common techniques include TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings like Word2Vec or GloVe to represent words as numerical vectors.
- Consider additional features such as sentiment scores, readability metrics, and linguistic features.

Step-3:

Data Splitting:

- Split the dataset into a training set and a testing set. Common splits are 70/30 or 80/20, but this can vary depending on the dataset size.

Step-4:

Model Selection:

- Choose a Logistic Regression model for binary classification. Logistic Regression is interpretable and works well for problems with a linear decision boundary, which is often sufficient for basic fake news detection.
- Apply a Logistic Regression model to classify the news articles into "fake" or "real" based on the extracted features.

Step-5:

Model Training:

- Train the Logistic Regression model on the training data. The model will learn the relationship between the features and the "fake" or "real" label.

Step-6:

Model Evaluation:

- Evaluate the model's performance on the testing set using common evaluation metrics like accuracy, precision, recall, F1-score, and the receiver operating characteristic (ROC) curve. Analyse the confusion matrix to understand false positives and false negatives.

Step-7:

Prediction:

- The project provides insights into the model's effectiveness in detecting fake news, with a focus on achieving high accuracy and reducing false positives.

Step-8:

Real Time Detection:

- Discuss the deployment of the trained model in a production environment for real-time fake news detection.
- A Fakenews can be detected by getting a news as an input and Converting into numerical form and make prediction using trained model.

Datasets:

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

Importing Necessary Libraries:

```
#  importing necessary module
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score,precision_score,recall_score,f1_score,roc_auc_score
import pickle as pk
```

Text Preprocessing and Training model:

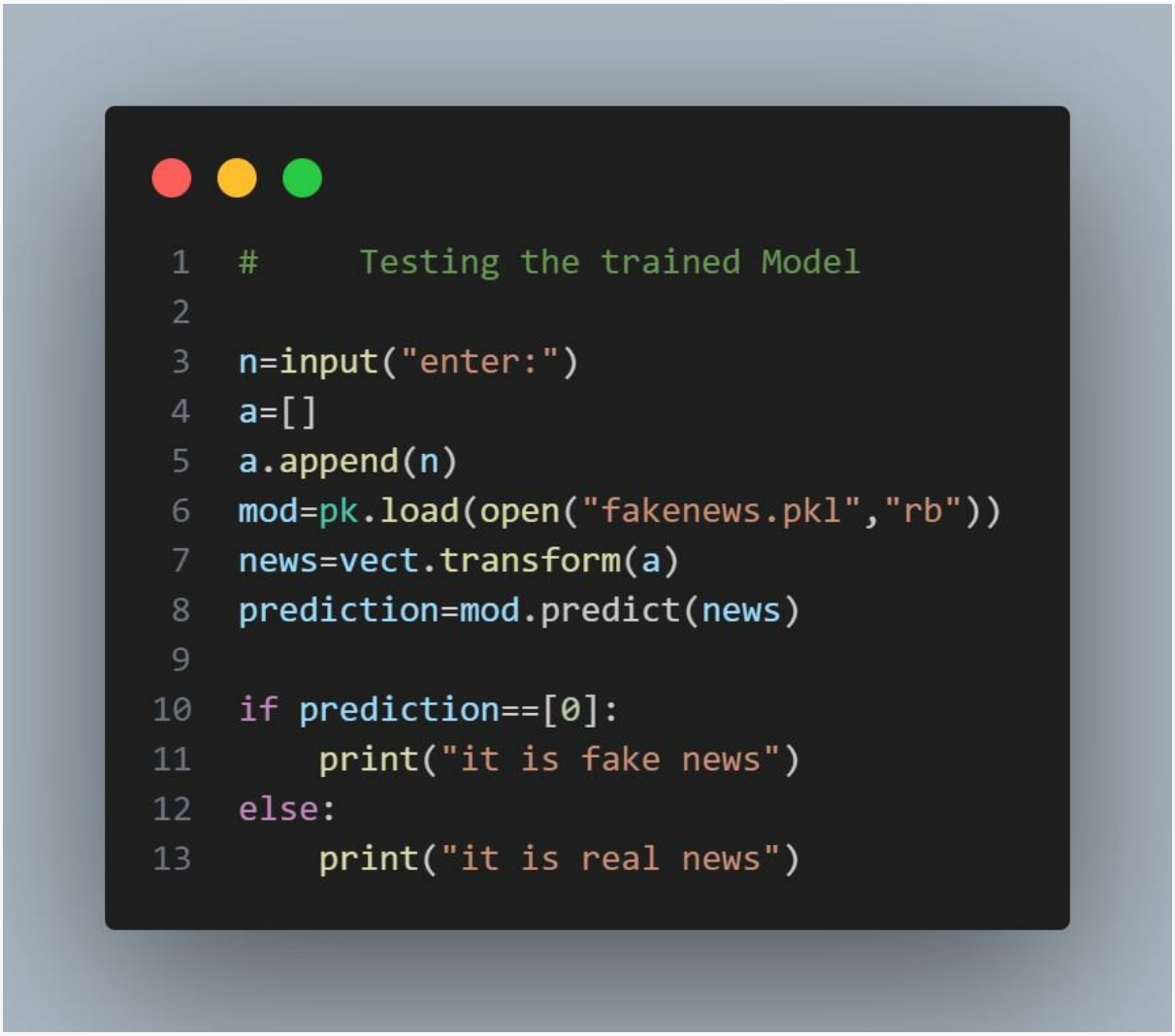
Text Preprocessing can be done by using

TfidfVectorizer and Training Model by using **LogisticRegression Algorithm.**

```
x_train,x_test,y_train,y_test=train_test_split(data['text'],data['class'],test_size=0.2,random_state=0)
print(x_train.shape)
print(y_train.shape)
print(x_test.shape)
print(y_test.shape)

# converting a text into numerical form
vect=TfidfVectorizer()
x_train=vect.fit_transform(x_train)
x_test=vect.transform(x_test)

# Training Model
clas2=LogisticRegression()
clas2.fit(x_train,y_train)

y_pred2=clas2.predict(x_test)
accuracy1=accuracy_score(y_test,y_pred2)

precision=precision_score(y_test,y_pred2)
recall=recall_score(y_test,y_pred2)
f1=f1_score(y_test,y_pred2)
rou=roc_auc_score(y_test,y_pred2)

```

Prediction:

Predicting fake news using trained model.



```
#     Testing the trained Model

n=input("enter:")
a=[]
a.append(n)
mod=pk.load(open("fakenews.pkl","rb"))
news=vect.transform(a)
prediction=mod.predict(news)

if prediction==[0]:
    print("it is fake news")
else:
    print("it is real news")
```

Accuracy Details:

Accuracy score: 0.988641425389755

precision score: 0.9856807511737089

recall score: 0.9903301886792453

f1_score : 0.988 rou_auc_score:

0.9887304951835044

Conclusion:

In conclusion, using the Logistic Regression algorithm for fake news detection has proven to be a valuable and effective approach in the ongoing battle against misinformation and disinformation. This method leverages the power of binary classification to distinguish between real and fake news articles based on the features and patterns within the data.