FAKE NEWS DETECTION USING NLP IN ARTIFICIAL INTELLIGENCE

TEAM MEMBER

au820421205071: SUBANU R S

Phase-5 SUBMISSION DOCUMENT

Project: Fake News Detection

Phase 5: Project Documentation and Submission

PROBLEM STATEMENT:

Detecting fake news using Natural Language Processing (NLP) involves developing a machine learning model or system that can differentiate between genuine news articles and fake or misleading information using a kaggle dataset.

OBJECTIVES:

- To minimize false positives and false negatives.
- Enabling timely responses to misleading information.
- Allows users to easily access and verify the authenticity of news articles.
- Focus on building public trust in the system's ability to combat fake news and provide accurate information.

DESIGN THINKING:

Design thinking can be a valuable approach to creating a fake news detection system using Natural Language Processing (NLP). Here's a design thinking process tailored to this specific context:

1. **Empathize:**

   Recognize the issue with fake news: Start by learning about the subtleties and difficulties associated with fake news within the framework of NLP. This entails being aware of the characteristics, origins, and possible social repercussions of fake news.

User research: List the different parties—journalists, fact-checkers, social media companies, and the general public—that are involved in the identification of fake news. To learn about their particular demands and pain spots, conduct user studies, questionnaires, and interviews.

2. **Define:**

   Definition of the issue: Specify the objectives and particular difficulties of utilizing natural language processing (NLP) to detect fake news based on your research. Establish your goals and the essential success factors, such as scalability and accuracy.

   Think with specialists: Work together with data scientists, domain experts, and NLP experts to generate ideas for approaches and solutions for detecting false news using NLP techniques.

3. **Ideate:**

   conceive of NLP solutions: Promote original thought and the generation of concepts for NLP-based false news detection techniques, such as topic modeling, sentiment analysis, named entity identification, and text analysis.

   Assess and rank the ideas: Evaluate each idea's viability, possible impact, and suitability for the needs of the user. Give the most promising NLP methods a priority for future advancement.

4. **Test:**

   Carry out comprehensive testing: Analyze the precision and effectiveness of the NLP-based system in spotting false information. Try it with several kinds of fake news, such as pictures, videos, and text.

   Obtain user opinions: Involve prospective users to evaluate the system's usability and pinpoint areas for improvement, such as content moderators and fact-checkers.
   Refine and iterate: Based on test results and user comments, make the necessary modifications to the NLP models and system.

5. **Implement:**

   Create the ultimate resolution: Construct the NLP-based fake news detection system that is ready for production, making sure that the NLP models and algorithms are integrated.

   Connect to platforms: Make sure the system can be incorporated into news websites, social media accounts, and other online spaces where fake news could proliferate.

6. **Iterate:**

   Frequently refresh: Update the NLP models and system to reflect the most recent developments in NLP technology and the changing landscape of fake news.

   Adjust to fresh difficulties: Keep an eye out for new issues and trends in fake news and adjust the system as necessary.

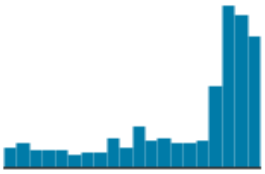
PHASES OF DEVELOPMENT:

1. Importing dataset
2. Importing libraries
3. Data preprocessing
    - Concatenation of true and fake dataset
    - Dropping unwanted columns
    - Checking for null values
    - Random shuffling of dataset
4. Model training
    - Functions to process the text
    - Splitting of data
    - Splitting data into training and testing data
5. Feature extraction
    - Text to vector
6. Classification technique
    - Logistic regression
    - Training the model
    - Prediction
    - Data visualization
    - Evaluation
    - Checking of our input
    - Checking of our input using various classification models

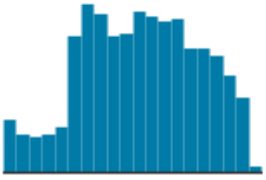
GIVEN DATASET

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 Eye'	On Friday, it was revealed that former Milwaukee Sheriff David Clarke, who was being considered	News	December 30, 2017

DATA DESCRIPTION :

Title: the title of a news article

Text: the text of the article; could be incomplete

Subject: display the field of the news

Date: publish date

CHOICE OF CLASSIFICATION ALGORITHM:

We have use the **Logistic Regression** algorithm for detecting whether news is real or fake.

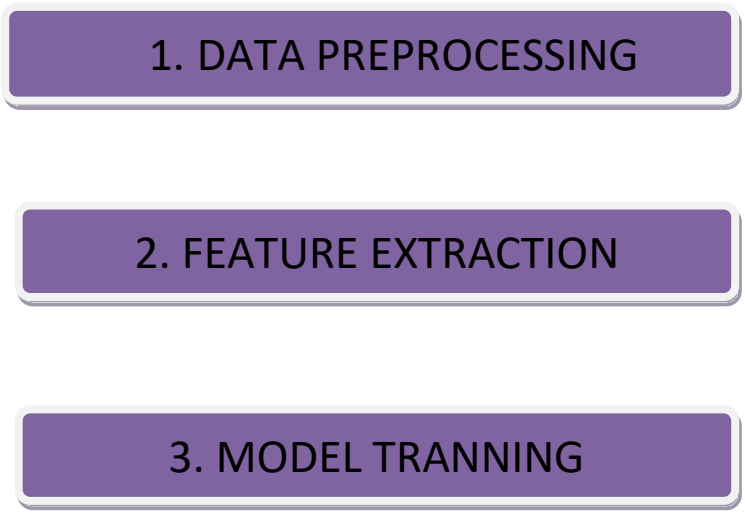
DEFENITION:

Logistic Regression:

Logistic regression is a statistical method used for binary classification tasks. Despite its name containing the term "regression," it's actually a classification algorithm rather than a regression technique.

The primary objective of logistic regression is to predict the probability that a given input belongs to a particular category or class. It's especially useful when the dependent variable is binary, meaning it has two possible outcomes (e.g., 0 or 1, yes or no, true or false).

NECESSARY STEPS TO FOLLOW:



DATA PREPROCESSING:

1. **Data collection:** Compile unprocessed data from a range of sources, including files, databases, web scraping, sensors, and surveys.

2. **Data Inspection:** Look over the information to gain a rudimentary comprehension of its composition and organization. This entails looking for outliers, duplicates, and missing values.
3. **Cleaning Data:**

   a. Managing Missing Data: Choose a plan of action for handling missing data. This could involve eliminating rows that contain missing values, impute missing values using a mean or median, or apply more advanced imputation techniques.

b. Managing Duplicates: Find and eliminate any rows or records in the dataset that are duplicates.

4. Data Transformation:

a. Feature Selection: Choose which features (columns) are relevant to the analysis or modeling task. Eliminate irrelevant or redundant features.

b. Feature Engineering: Create new features or transform existing ones to better represent the underlying patterns in the data. This can involve scaling, encoding categorical variables, or creating interactions between features.

c. Text Data Processing: Tokenize and preprocess text data, which can involve removing punctuation, stop words, and stemming or lemmatization.

5. Data Splitting:

Split the data into training, validation, and test sets for machine learning tasks. The training set is used to train the model, the validation set is used to tune hyperparameters and evaluate performance during training, and the test set is used to assess the final model's performance.

6. **Data Scaling:**

   If using algorithms sensitive to feature scales (e.g., k-NN, SVM), ensure that all features have been appropriately scaled or standardized.

7. **Data Visualization:**
   Visualize the data to gain insights, detect patterns, and identify relationships between variables. Visualization aids in exploratory data analysis (EDA).

FEATURE EXTRACTION:

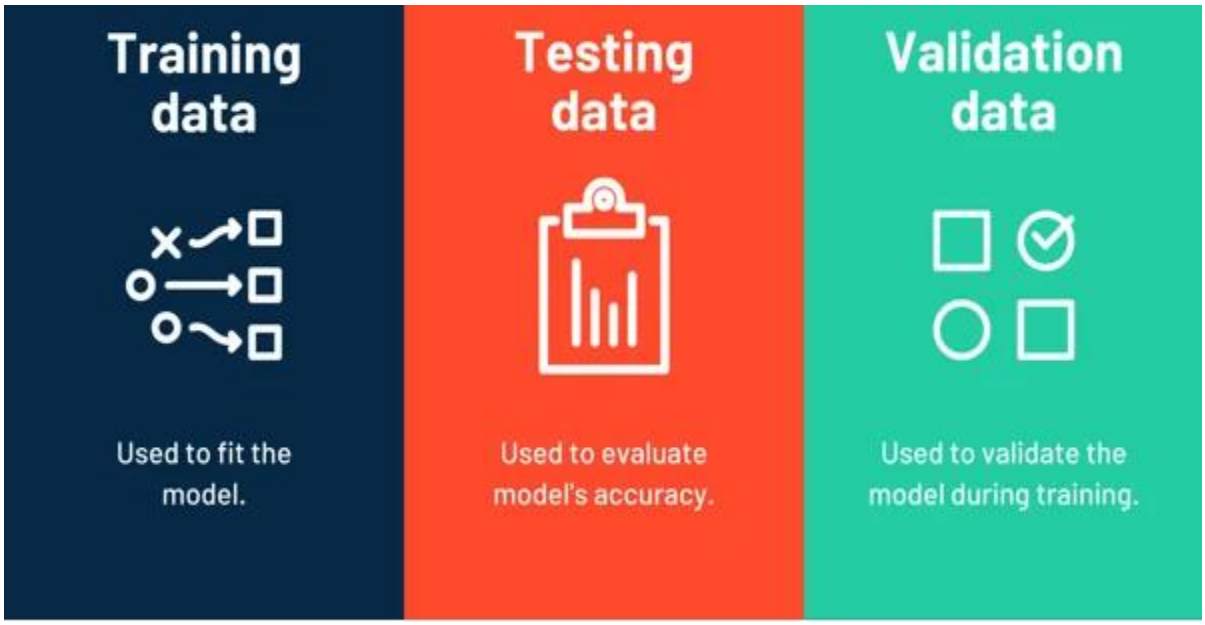
Feature extraction techniques encompass a variety of methods used to derive relevant, informative, and reduced representations from raw data. These techniques are employed across various domains such as natural language processing, image processing, signal processing, and more. Here we use word embeddings feature extraction technique.

Word Embeddings:

In natural language processing, word embeddings (e.g., Word2Vec, GloVe, FastText) convert words into dense, low-dimensional vectors that capture semantic relationships between words based on their usage in large text corpora.

MODEL TRANNING:

Model training is a fundamental process in machine learning and deep learning where a machine learning model learns to make predictions or decisions based on input data. During the training process, a model is exposed to a dataset containing examples (input data) and their corresponding correct or target outputs (labels or responses). The model adjusts its internal parameters or weights to make its predictions as accurate as possible.

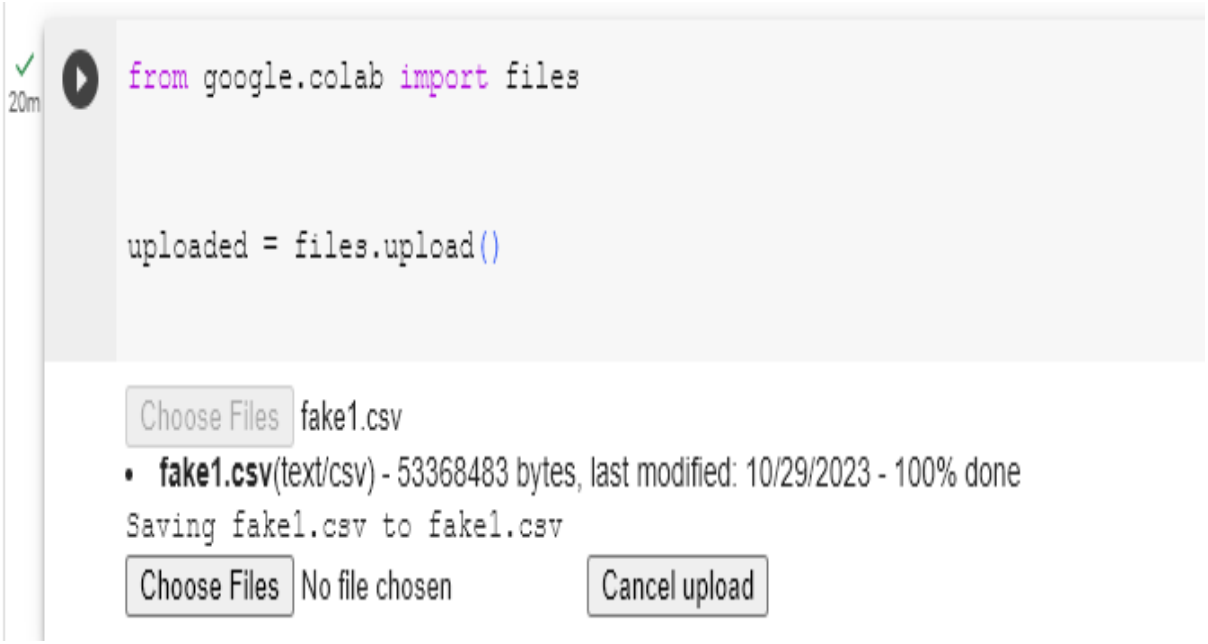


1. **Data Preparation:** The first step in model training involves preparing the training data. This includes collecting, cleaning, and preprocessing the data to ensure it's in a suitable format for the model. The dataset is typically split into two parts: a training set for training the model and a validation set to assess the model's performance during training and tune hyperparameters.

2. **Training Loop:** The model is trained in a series of iterations or epochs. During each epoch, the model processes the training data, makes predictions, computes the loss, and adjusts the weights using the optimization algorithm.

3. **Validation:** At the end of each epoch or in regular intervals, the model is evaluated on the validation set to assess its generalization performance. This helps prevent overfitting, a common issue in machine learning where a model becomes too specialized on the training data and performs poorly on new, unseen data.

UPLOADING DATASET



IMPORTING LIBRARIES

```
from google.colab import files


uploaded = files.upload()
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix,ConfusionMatrixDisplay
import re
import string
```

TRUE DATASET:

```
true_data=pd.read_csv('true1.csv',encoding='ISO-8859-1')    #reading true news dataset
fake_data = pd.read_csv('fake1.csv',encoding='ISO-8859-1')  #reading fake news dataset
print(true_data.head())
```

```
                                               title  \
0  As U.S. budget fight looms, Republicans flip t...
1  U.S. military to accept transgender recruits o...
2  Senior U.S. Republican senator: 'Let Mr. Muell...
3  FBI Russia probe helped by Australian diplomat...
4  Trump wants Postal Service to charge 'much mor...

                                                text       subject       date
0  WASHINGTON (Reuters) - The head of a conservat...  politicsNews  31-Dec-17
1  WASHINGTON (Reuters) - Transgender people will...  politicsNews  29-Dec-17
2  WASHINGTON (Reuters) - The special counsel inv...  politicsNews  31-Dec-17
3  WASHINGTON (Reuters) - Trump campaign adviser ...  politicsNews  30-Dec-17
4  SEATTLE/WASHINGTON (Reuters) - President Donal...  politicsNews  29-Dec-17
```

FAKE DATASET:

```
print(fake_data.head())
```

```
                                               title  \
0  As U.S. budget fight looms, Republicans flip t...
1  U.S. military to accept transgender recruits o...
2  Senior U.S. Republican senator: 'Let Mr. Muell...
3  FBI Russia probe helped by Australian diplomat...
4  Trump wants Postal Service to charge 'much mor...

                                                text       subject       date
0  WASHINGTON (Reuters) - The head of a conservat...  politicsNews  31-Dec-17
1  WASHINGTON (Reuters) - Transgender people will...  politicsNews  29-Dec-17
2  WASHINGTON (Reuters) - The special counsel inv...  politicsNews  31-Dec-17
3  WASHINGTON (Reuters) - Trump campaign adviser ...  politicsNews  30-Dec-17
4  SEATTLE/WASHINGTON (Reuters) - President Donal...  politicsNews  29-Dec-17
```

DATA PREPROCESSING

Adding data attribute to dataset

True data

```
true_data['class']=1
fake_data['class']=0
print(true_data.head())
```

Output

```
                                                title  \
0  As U.S. budget fight looms, Republicans flip t...
1  U.S. military to accept transgender recruits o...
2  Senior U.S. Republican senator: 'Let Mr. Muell...
3  FBI Russia probe helped by Australian diplomat...
4  Trump wants Postal Service to charge 'much mor...

                                                text       subject       date  \
0  WASHINGTON (Reuters) - The head of a conservat...  politicsNews  31-Dec-17
1  WASHINGTON (Reuters) - Transgender people will...  politicsNews  29-Dec-17
2  WASHINGTON (Reuters) - The special counsel inv...  politicsNews  31-Dec-17
3  WASHINGTON (Reuters) - Trump campaign adviser ...  politicsNews  30-Dec-17
4  SEATTLE/WASHINGTON (Reuters) - President Donal...  politicsNews  29-Dec-17

   class
0      1
1      1
2      1
3      1
4      1
```

Fake data

```
true_data['class']=1
fake_data['class']=0
#print(true_data.head())
print(fake_data.head())
```

Output

```
                                               title  \
0  As U.S. budget fight looms, Republicans flip t...
1  U.S. military to accept transgender recruits o...
2  Senior U.S. Republican senator: 'Let Mr. Muell...
3  FBI Russia probe helped by Australian diplomat...
4  Trump wants Postal Service to charge 'much mor...

                                                text       subject  \
0  WASHINGTON (Reuters) - The head of a conservat...  politicsNews
1  WASHINGTON (Reuters) - Transgender people will...  politicsNews
2  WASHINGTON (Reuters) - The special counsel inv...  politicsNews
3  WASHINGTON (Reuters) - Trump campaign adviser ...  politicsNews
4  SEATTLE/WASHINGTON (Reuters) - President Donal...  politicsNews

        date  \
0  31-Dec-17
1  29-Dec-17
2  31-Dec-17
3  30-Dec-17
4  29-Dec-17

   class
0      0
1      0
2      0
3      0
4      0
```

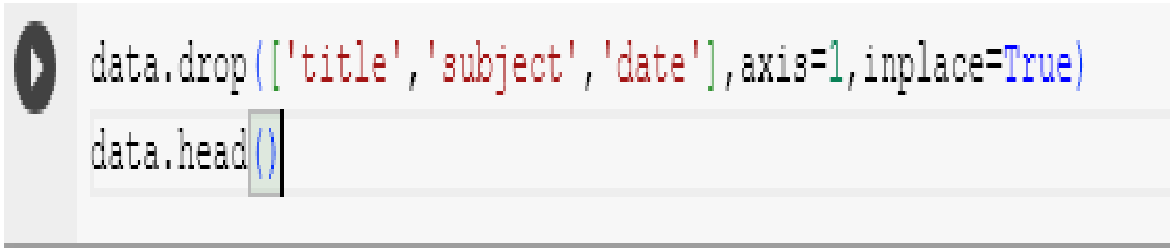
Concatenation of true and fake dataset

```
data=pd.concat([fake_data,true_data],axis=0)
print(data.head())
```

Output

	title	text	subject	date	class
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	31-Dec-17	0
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	29-Dec-17	0
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	31-Dec-17	0
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	30-Dec-17	0
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	29-Dec-17	0

Dropping unwanted columns



Output

	text	class
0	WASHINGTON (Reuters) - The head of a conservat...	1
1	WASHINGTON (Reuters) - Transgender people will...	1
2	WASHINGTON (Reuters) - The special counsel inv...	1
3	WASHINGTON (Reuters) - Trump campaign adviser ...	1
4	SEATTLE/WASHINGTON (Reuters) - President Donal...	1

Checking for null values

```
print(data.isnull().sum())
```

```
text     0
class    0
dtype: int64
```

Random shuffling of dataset

```
data=data.sample(frac=1)
print(data.head())
```

Output

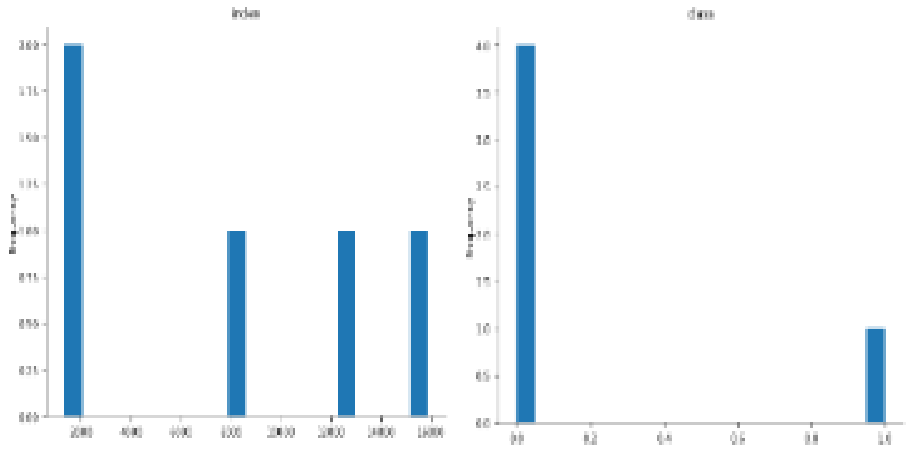
	text	class
8099	UNITED NATIONS (Reuters) - U.S. President Bara...	0
1613	WASHINGTON (Reuters) - Republican Senator Susa...	1
15858	ANKARA (Reuters) - Turkey is planning 22 new d...	0
12932	DUBLIN (Reuters) - British Prime Minister Ther...	0
1333	MEXICO CITY (Reuters) - The fourth round of ta...	0

```
data.reset_index(inplace = True)
data.head()
```

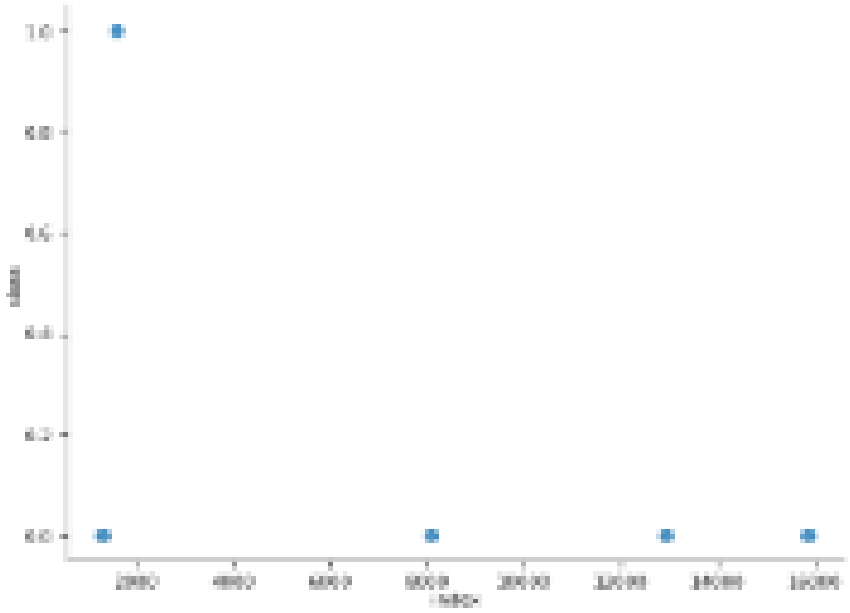
Output

	text	class
15865	If we didn t know better, we d almost believe ...	0
15469	It s not just Trump who s exposing the truth a...	0
12744	HANOI (Reuters) - Vietnamese police on Friday ...	1
6398	WASHINGTON (Reuters) - U.S. President-elect Do...	1
13980	The globalists aren t happy which is a signal ...	0

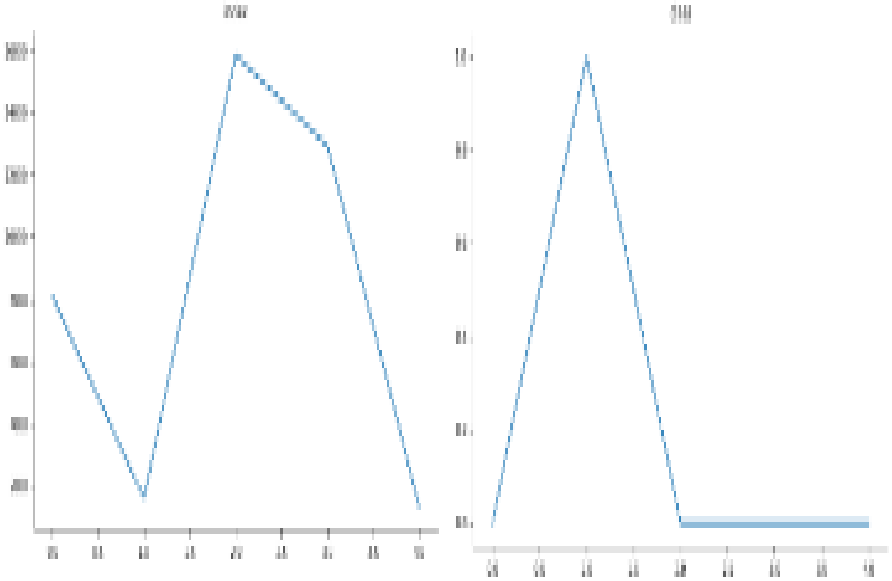
Distributions



2-d distributions



Values



```
data.drop(["index"], axis = 1, inplace = True)

data.head()
```

Output

index	level_0	text	class
0	0	UNITED NATIONS (Reuters) - U.S. President Barack Obama met Chinese Premier Li Keqiang on Monday and urged China to accelerate efforts to address the problem of industrial excess capacity, the White House said. In a meeting on the sidelines of the United Nations General Assembly in New York, Obama also urged Beijing to establish a level playing field to allow all firms to compete fairly in China, a White House statement said.	0
1	1	WASHINGTON (Reuters) - Republican Senator Susan Collins has said she is □??leaning against□?□ the latest Republican healthcare bill, the Associated Press reported on Friday. The Maine senator indicated she had major concerns with the proposal, AP said in a note on Twitter.	1
2	2	ANKARA (Reuters) - Turkey is planning 22 new defense projects worth a total of $5 billion and spanning the development of air defense missile systems, ammunition and arms systems, the defense industry executive board said on Thursday. It also said the board had discussed the modernization and mass production of tanks, combat and general purpose helicopters, as well as the production of new and additional unmanned aerial vehicle (UAV) drones.	0
3	3	DUBLIN (Reuters) - British Prime Minister Theresa May told her Irish counterpart Leo Varadkar that she would propose suggestions to Brexit negotiators over the next 24 hours to try to break an impasse on the issue of the Irish border, Varadkar said on Wednesday. A tentative deal on the border, which is required if Brexit talks are to move to the next phase, was agreed with Dublin s blessing on Monday after negotiators guaranteed regulatory alignment on both sides of the border that Ireland has with the British province of Northern Ireland. But the Northern Irish party that props up May s minority government rejected the agreement, saying it could not allow any divergence in regulations between Northern Ireland and other parts of the UK, putting up a fresh obstacle a week before EU leaders meet to decide whether to open trade talks. We discussed the idea certainly but we didn t discuss any particular words or combination of words or language but I certainly indicated a willingness to consider any proposals that the UK side have, Varadkar told a news conference after speaking to May by telephone earlier on Wednesday. Having consulted with people in London, she wants to come back to us with some text tonight or tomorrow. I expressed my willingness to consider that because I want us to move to phase two if that is possible next week. Varadkar said he had a very good call with May but that he reiterated the firm Irish position on issue of the border and that any new language proposed by London must be consistent with the text May had originally agreed to on Monday. The leader of Northern Ireland s Democratic Unionist Party Arlene Foster also spoke to May earlier on Wednesday and a spokesman for the party said there was still work to be done on any border deal. Brussels says Britain must present its offer this week or it will be too late for a decision. Varadkar was speaking at a news conference with Dutch Prime Minister Mark Rutte who assured Dublin that a satisfactory deal on the Irish border was essential if EU leaders are to declare that sufficient progress has been made in phase one of the talks when they meet on Dec. 15. Rutte also agreed with Varadkar that if an agreement cannot be reached next week, then they will have to move into January It is the desire and ambition and wish of this government that we should move onto the phase two talks but if it isn t possible to move to phase two next week because of the problems that have arisen, well then we can pick it up in the new year, Varadkar earlier told parliament.	0
4	4	MEXICO CITY (Reuters) - The fourth round of talks to renegotiate the North American Free Trade agreement have been prolonged until Oct. 17, two sources in Mexico said on Tuesday, as negotiators gathering in Washington were expected to start tackling difficult issues. The round of talks due to begin on Wednesday is expected to include discussions about including quotas for U.S. content in autos, a major bone of contention for Mexico, Canada and many companies. Previously, the talks were due to end on Oct. 15. The news was first reported by Bloomberg earlier on Tuesday.	0

Show 25 per page

MODEL TRAINING

Functions to process the text

```
def wordopt(text):
    text = text.lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub("\\W"," ",text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    return text
data['text']=data['text'].apply(wordopt)
```

Splitting of data

```
x=data['text']
y=data['class']
```

```
X_train,X_test,y_train,y_test=train_test_split(x,y,test_size=0.20)
```

FEATURE EXTRACTION

Text to vector

```
tfv=TfidfVectorizer()
x=tfv.fit_transform(x)
```

LOGISTIC REGERSSION

Training the model

```
from sklearn.linear_model import LogisticRegression
lr_model=LogisticRegression()
lr_model.fit(X_train,y_train)
```

Output

```
LogisticRegression()
```

Prediction

```
y_pred_lr=lr_model.predict(X_test)
y_pred_lr
```

Output

```
array([1, 1, 0, ..., 1, 0, 0], dtype=int64)
```

```
accuracy_score(y_pred_lr,y_test)
```

Output

```
0.9870824053452116
```

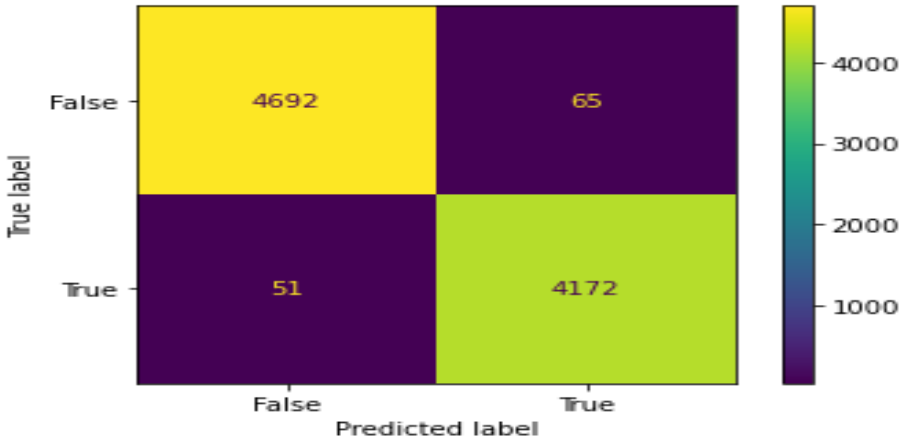
Data visualization

```
cm = confusion_matrix(y_test, y_pred_lr)

cm_display = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=[False, True])

cm_display.plot()
plt.show()
```

Output



Evaluation

```
print(classification_report(y_pred_lr,y_test))
```

Output

	precision	recall	f1-score	support
0	0.99	0.99	0.99	4743
1	0.99	0.98	0.99	4237
accuracy			0.99	8980
macro avg	0.99	0.99	0.99	8980
weighted avg	0.99	0.99	0.99	8980

CHECKING OF OUR INPUT

```
def checking_our_value():
    x=X_train[100]
    print('Logistic Regression',output(lr_model.predict(x)))
def output(n):
   if n==1:
       return 'True news'
   else:
         return 'Fake news'
checking_our_value()
```

Output

```
Logistic Regression True news
```

CHECKING OF INPUT USING VARIOUS CLASSIFICATION TECHNIQUE

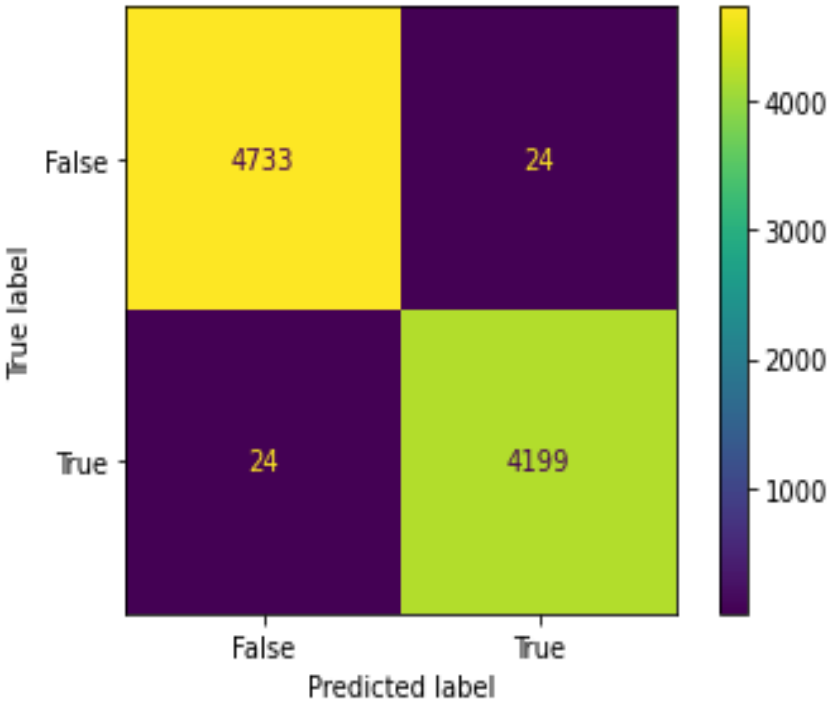
DECISION TREE CLASSIFICATION

```
from sklearn.tree import DecisionTreeClassifier
dtc_model = DecisionTreeClassifier()
dtc_model.fit(X_train, y_train)
accuracy_score(y_pred_dtc,y_test)
cm = confusion_matrix(y_test, y_pred_dtc)
cm_display = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=[False, True])
cm_display.plot()
plt.show()
print(classification_report(y_pred_dtc,y_test))
```

Output

```
DecisionTreeClassifier()

0.9946547884187082
```



	precision	recall	f1-score	support
0	0.99	0.99	0.99	4757
1	0.99	0.99	0.99	4223
accuracy			0.99	8980
macro avg	0.99	0.99	0.99	8980
weighted avg	0.99	0.99	0.99	8980

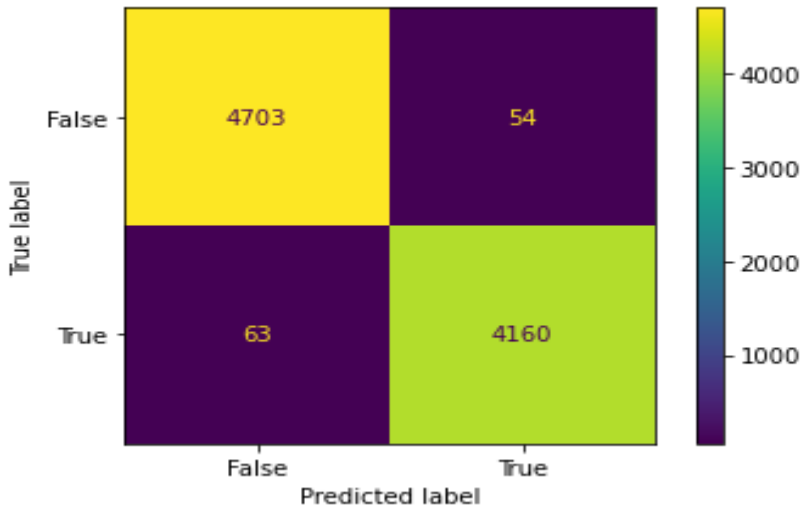
RANDOM FOREST CLASSIFICATION

```
from sklearn.ensemble import RandomForestClassifier
rfc_model= RandomForestClassifier(n_estimators=100,criterion='entropy')
rfc_model.fit(X_train, y_train)
accuracy_score(y_test,y_pred_rfc)
cm = confusion_matrix(y_test, y_pred_rfc)
cm_display = ConfusionMatrixDisplay(confusion_matrix=cm,display_labels=[False, True])
cm_display.plot()
plt.show()
print(classification_report(y_pred_rfc,y_test))
```

Output

```
RandomForestClassifier(criterion='entropy')
```

```
0.9869710467706013
```

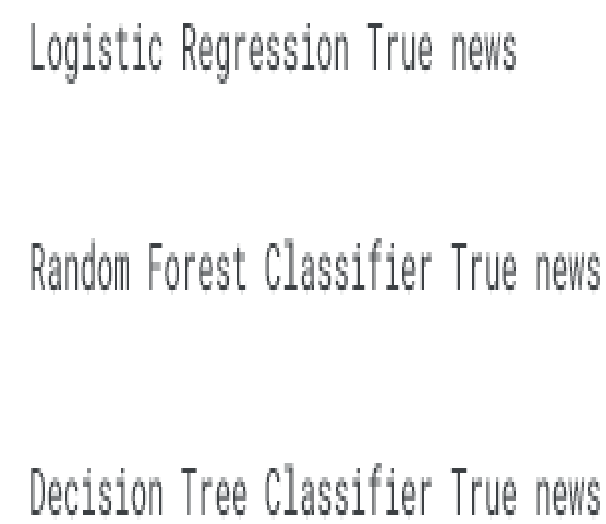


	precision	recall	f1-score	support
0	0.99	0.99	0.99	4766
1	0.99	0.99	0.99	4214
accuracy			0.99	8980
macro avg	0.99	0.99	0.99	8980
weighted avg	0.99	0.99	0.99	8980

If you want a straightforward and interpretable model for binary classification, Logistic Regression is a good choice.

```
def checking_our_value():
    x=X_train[100]
    print('Logistic Regression',output(lr_model.predict(x)))
    print('Random Forest Classifier',output(rfc_model.predict(x)))
    print('Decision Tree Classifier',output(dtc_model.predict(x)))
def output(n):
   if n==1:
      return 'True news'
   else:
         return 'Fake news'
checking_our_value()
```

Output



Full code:

Code:

```
import numpy as np
import pandas as pd
 #for visualization of the data
import matplotlib.pyplot as plt
import seaborn as sns
```

```
#to split train and test data set
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
 #for feature scaling
# for checking accuracy, precision, f1score, confusion matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix,ConfusionMatrixDisplay
# regular expression
import re
# string manipulation
import string
#reading true news dataset
true_data=pd.read_csv
 #reading fake news dataset
fake_data = pd.read_csv
true_data.head()
fake_data.head()
#labeling
true_data['class']=1
fake_data['class']=0
true_data.head()
```

```
fake_data.head()
true_data.shape
fake_data.shape
#Concatenation of true and fake dataset
data=pd.concat([true_data,fake_data],axis=0)
data.head()
#Dropping unwanted columns
data.drop(['title','subject','date'],axis=1,inplace=True)
data.head()
Checking for null values
data.isnull().sum()
Random shuffling of dataset
data=data.sample(frac=1)
data.head()
data.reset_index(inplace = True)
data.head()
data.drop(["index"], axis = 1, inplace = True)
data.head()
Functions to process the text
def wordopt(text):
    text = text.lower()
    text = re.sub('\[.*?\]', '', text)
```

```
    text = re.sub("\\W"," ",text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    return text
data['text']=data['text'].apply(wordopt)
```

Splitting of data

```
x=data['text']
y=data['class']
```

Text to vector

```
tfv=TfidfVectorizer()
x=tfv.fit_transform(x)
X_train,X_test,y_train,y_test=train_test_split(x,y,test_size=0.20)
```

Logistic regression

```
from sklearn.linear_model import LogisticRegression
lr_model=LogisticRegression()
lr_model.fit(X_train,y_train)
y_pred_lr=lr_model.predict(X_test)
y_pred_lr
accuracy_score(y_pred_lr,y_test)
```

```
cm = confusion_matrix(y_test, y_pred_lr)

cm_display = ConfusionMatrixDisplay(confusion_matrix=cm,
display_labels=[False, True])

cm_display.plot()

plt.show()

print(classification_report(y_pred_lr,y_test))
```

Decision Tree Classifier

```
from sklearn.tree import DecisionTreeClassifier

dtc_model = DecisionTreeClassifier()

dtc_model.fit(X_train, y_train)

y_pred_dtc=dtc_model.predict(X_test)

y_pred_dtc

accuracy_score(y_pred_dtc,y_test)

cm = confusion_matrix(y_test, y_pred_dtc)

cm_display = ConfusionMatrixDisplay(confusion_matrix=cm,
display_labels=[False, True])

cm_display.plot()

plt.show()

print(classification_report(y_pred_dtc,y_test))
```

Random Forest Classifier

```
from sklearn.ensemble import RandomForestClassifier

rfc_model=RandomForestClassifier(n_estimators=100,criterion='entrop
y')
```

```
rfc_model.fit(X_train, y_train)

y_pred_rfc=rfc_model.predict(X_test)

y_pred_rfc

accuracy_score(y_test,y_pred_rfc)

cm = confusion_matrix(y_test, y_pred_rfc)

cm_display=ConfusionMatrixDisplay(confusion_matrix=cm,display_labe
ls=[False, True])


cm_display.plot()

plt.show()
```

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.