Fake News Detection

Team Member:

Name	Reg No
ABDUR RAHIM S	821021104003
SiVA PRAKASH S	821021104043
VARUN RAJ M	821021104055
VIJAY RAJ B J	821021104057
SIVA G	821021104304

Table of Content

S/N	Topic	Page
1	Task And Dataset	3
2	OpenCV	4
3	TensorFlow	5
4	Convolutional Neural Network(CNN)	6
5	Preprocessing Explaination	7-9
6	Programs	10-12
7	Preprocesses the textual data Classification with Keras	13-24

Fake News Detection Using NLP

Phase-4

Task :

In this part you will continue building your project. Continue building the fake news detection model by applying NLP techniques and training a classification model.

- Text Preprocessing and Feature Extraction
- Model training and evaluation

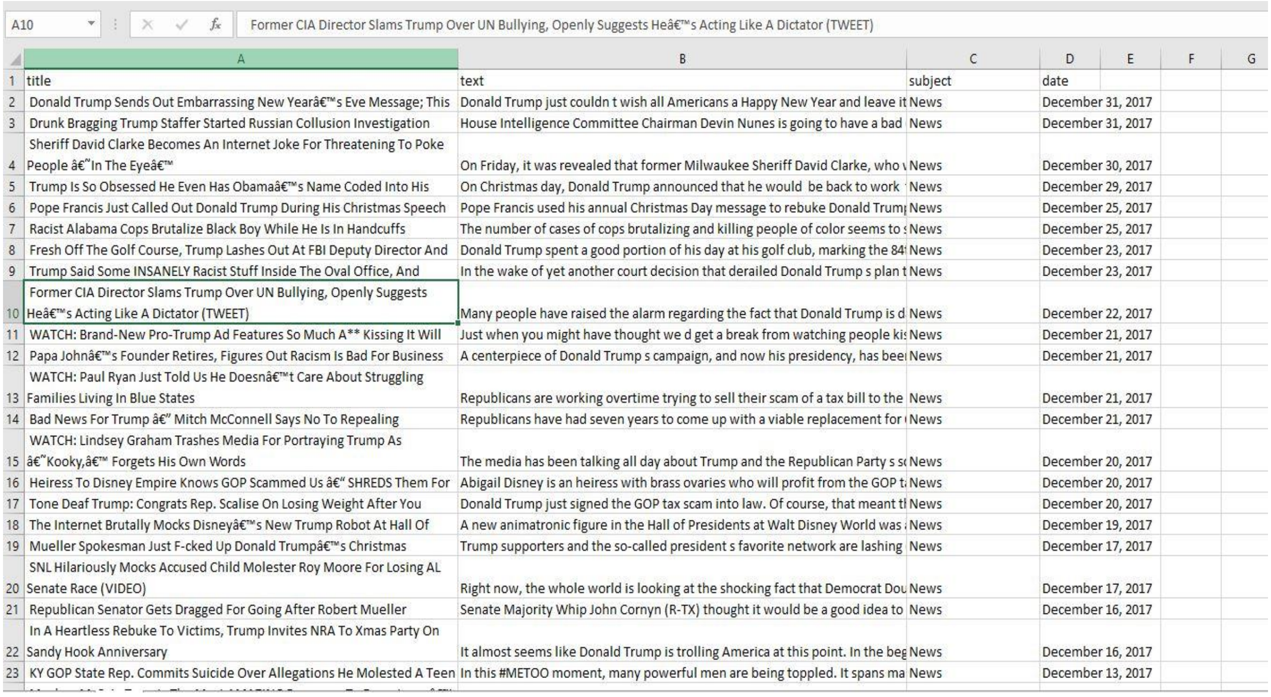
DataSet :

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

This Fake and Real News Dataset have two files :

- Fake.csv
- True.csv

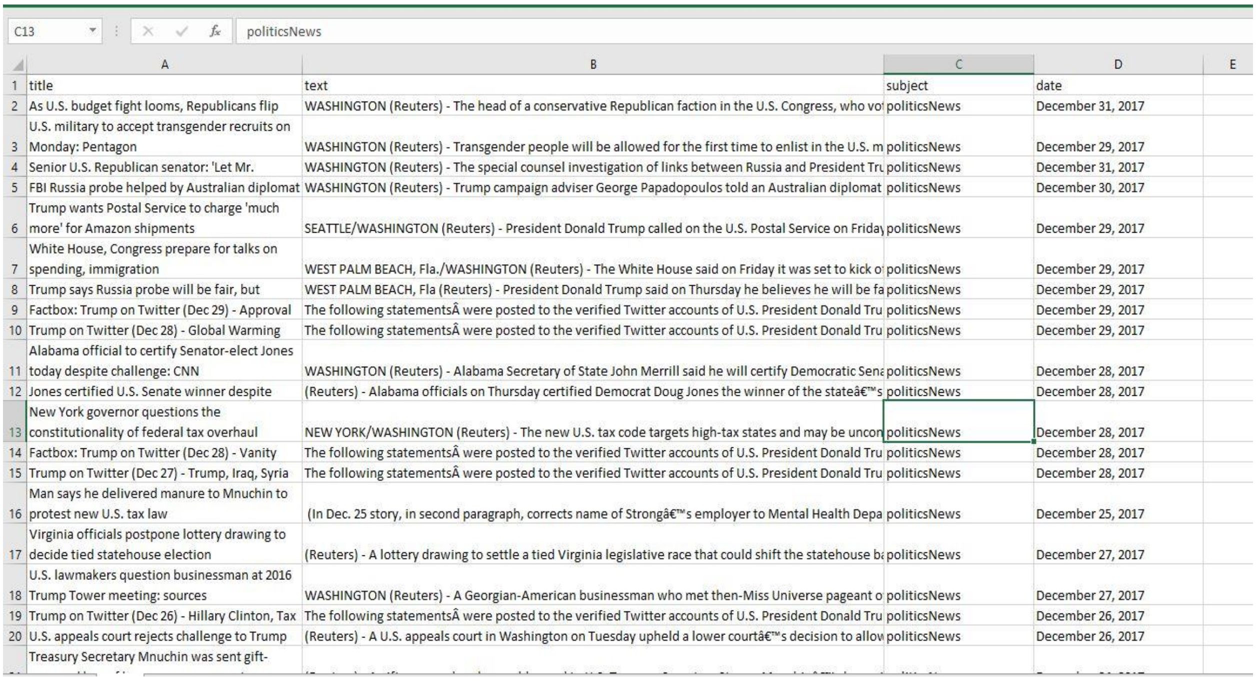
Fake.csv:



A10 | Former CIA Director Slams Trump Over UN Bullying, Openly Suggests Heâ€™s Acting Like A Dictator (TWEET)

	A	B	C	D	E	F	G
1	title	text	subject	date			
2	Donald Trump Sends Out Embarrassing New Yearâ€™s Eve Message; This	Donald Trump just couldn t wish all Americans a Happy New Year and leave it	News	December 31, 2017			
3	Drunk Bragging Trump Staffer Started Russian Collusion Investigation	House Intelligence Committee Chairman Devin Nunes is going to have a bad	News	December 31, 2017			
4	Sheriff David Clarke Becomes An Internet Joke For Threatening To Poke People â€˜In The Eyeâ€™	On Friday, it was revealed that former Milwaukee Sheriff David Clarke, who v	News	December 30, 2017			
5	Trump Is So Obsessed He Even Has Obamaâ€™s Name Coded Into His	On Christmas day, Donald Trump announced that he would be back to work	News	December 29, 2017			
6	Pope Francis Just Called Out Donald Trump During His Christmas Speech	Pope Francis used his annual Christmas Day message to rebuke Donald Trum	News	December 25, 2017			
7	Racist Alabama Cops Brutalize Black Boy While He Is In Handcuffs	The number of cases of cops brutalizing and killing people of color seems to	News	December 25, 2017			
8	Fresh Off The Golf Course, Trump Lashes Out At FBI Deputy Director And	Donald Trump spent a good portion of his day at his golf club, marking the 84	News	December 23, 2017			
9	Trump Said Some INSANELY Racist Stuff Inside The Oval Office, And	In the wake of yet another court decision that derailed Donald Trump s plan t	News	December 23, 2017			
10	Former CIA Director Slams Trump Over UN Bullying, Openly Suggests Heâ€™s Acting Like A Dictator (TWEET)	Many people have raised the alarm regarding the fact that Donald Trump is d	News	December 22, 2017			
11	WATCH: Brand-New Pro-Trump Ad Features So Much A** Kissing It Will	Just when you might have thought we d get a break from watching people kis	News	December 21, 2017			
12	Papa Johnâ€™s Founder Retires, Figures Out Racism Is Bad For Business	A centerpiece of Donald Trump s campaign, and now his presidency, has bee	News	December 21, 2017			
13	WATCH: Paul Ryan Just Told Us He Doesnâ€™t Care About Struggling Families Living In Blue States	Republicans are working overtime trying to sell their scam of a tax bill to the	News	December 21, 2017			
14	Bad News For Trump â€" Mitch McConnell Says No To Repealing	Republicans have had seven years to come up with a viable replacement for	News	December 21, 2017			
15	WATCH: Lindsey Graham Trashes Media For Portraying Trump As â€˜Kooky,â€™ Forgets His Own Words	The media has been talking all day about Trump and the Republican Party s s	News	December 20, 2017			
16	Heiress To Disney Empire Knows GOP Scammed Us â€" SHREDS Them For	Abigail Disney is an heiress with brass ovaries who will profit from the GOP t	News	December 20, 2017			
17	Tone Deaf Trump: Congrats Rep. Scalise On Losing Weight After You	Donald Trump just signed the GOP tax scam into law. Of course, that meant t	News	December 20, 2017			
18	The Internet Brutally Mocks Disneyâ€™s New Trump Robot At Hall Of	A new animatronic figure in the Hall of Presidents at Walt Disney World was	News	December 19, 2017			
19	Mueller Spokesman Just F-cked Up Donald Trumpâ€™s Christmas	Trump supporters and the so-called president s favorite network are lashing	News	December 17, 2017			
20	SNL Hilariously Mocks Accused Child Molester Roy Moore For Losing AL Senate Race (VIDEO)	Right now, the whole world is looking at the shocking fact that Democrat Dou	News	December 17, 2017			
21	Republican Senator Gets Dragged For Going After Robert Mueller	Senate Majority Whip John Cornyn (R-TX) thought it would be a good idea to	News	December 16, 2017			
22	In A Heartless Rebuke To Victims, Trump Invites NRA To Xmas Party On Sandy Hook Anniversary	It almost seems like Donald Trump is trolling America at this point. In the beg	News	December 16, 2017			
23	KY GOP State Rep. Commits Suicide Over Allegations He Molested A Teen	In this #METOO moment, many powerful men are being toppled. It spans ma	News	December 13, 2017			

True.csv :



C13 | politicsNews

	A	B	C	D	E
1	title	text	subject	date	
2	As U.S. budget fight looms, Republicans flip	WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, who vo	politicsNews	December 31, 2017	
3	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	
4	Senior U.S. Republican senator: 'Let Mr.	WASHINGTON (Reuters) - The special counsel investigation of links between Russia and President Tru	politicsNews	December 31, 2017	
5	FBI Russia probe helped by Australian diplomat	WASHINGTON (Reuters) - Trump campaign adviser George Papadopoulos told an Australian diplomat	politicsNews	December 30, 2017	
6	Trump wants Postal Service to charge 'much more' for Amazon shipments	SEATTLE/WASHINGTON (Reuters) - President Donald Trump called on the U.S. Postal Service on Friday	politicsNews	December 29, 2017	
7	White House, Congress prepare for talks on spending, immigration	WEST PALM BEACH, Fla./WASHINGTON (Reuters) - The White House said on Friday it was set to kick o	politicsNews	December 29, 2017	
8	Trump says Russia probe will be fair, but	WEST PALM BEACH, Fla (Reuters) - President Donald Trump said on Thursday he believes he will be fa	politicsNews	December 29, 2017	
9	Factbox: Trump on Twitter (Dec 29) - Approval	The following statementsÂ were posted to the verified Twitter accounts of U.S. President Donald Tru	politicsNews	December 29, 2017	
10	Trump on Twitter (Dec 28) - Global Warming	The following statementsÂ were posted to the verified Twitter accounts of U.S. President Donald Tru	politicsNews	December 29, 2017	
11	Alabama official to certify Senator-elect Jones today despite challenge: CNN	WASHINGTON (Reuters) - Alabama Secretary of State John Merrill said he will certify Democratic Sena	politicsNews	December 28, 2017	
12	Jones certified U.S. Senate winner despite	(Reuters) - Alabama officials on Thursday certified Democrat Doug Jones the winner of the stateâ€™s	politicsNews	December 28, 2017	
13	New York governor questions the constitutionality of federal tax overhaul	NEW YORK/WASHINGTON (Reuters) - The new U.S. tax code targets high-tax states and may be uncon	politicsNews	December 28, 2017	
14	Factbox: Trump on Twitter (Dec 28) - Vanity	The following statementsÂ were posted to the verified Twitter accounts of U.S. President Donald Tru	politicsNews	December 28, 2017	
15	Trump on Twitter (Dec 27) - Trump, Iraq, Syria	The following statementsÂ were posted to the verified Twitter accounts of U.S. President Donald Tru	politicsNews	December 28, 2017	
16	Man says he delivered manure to Mnuchin to protest new U.S. tax law	(In Dec. 25 story, in second paragraph, corrects name of Strongâ€™s employer to Mental Health Depa	politicsNews	December 25, 2017	
17	Virginia officials postpone lottery drawing to decide tied statehouse election	(Reuters) - A lottery drawing to settle a tied Virginia legislative race that could shift the statehouse b	politicsNews	December 27, 2017	
18	U.S. lawmakers question businessman at 2016 Trump Tower meeting: sources	WASHINGTON (Reuters) - A Georgian-American businessman who met then-Miss Universe pageant o	politicsNews	December 27, 2017	
19	Trump on Twitter (Dec 26) - Hillary Clinton, Tax	The following statementsÂ were posted to the verified Twitter accounts of U.S. President Donald Tru	politicsNews	December 26, 2017	
20	U.S. appeals court rejects challenge to Trump	(Reuters) - A U.S. appeals court in Washington on Tuesday upheld a lower courtâ€™s decision to allow	politicsNews	December 26, 2017	
	Treasury Secretary Mnuchin was sent gift-				

Object detection using yolo :

Object detection is a technique used in computer vision for the identification and localization of objects within an image or a video. Image Localization is the process of identifying the correct location of one or multiple objects using bounding boxes, which correspond to rectangular shapes around the objects. This process is sometimes confused with image classification or image recognition, which aims to predict the class of an image or an object within an image into one of the categories or classes. The authors frame the object detection problem as a regression problem instead of a classification task by spatially separating bounding boxes and associating probabilities to each of the detected images using a single convolutional neural network (CNN). By taking the **Image Processing with Keras in Python** course, you will be able to build Keras based deep neural networks for image classification tasks.

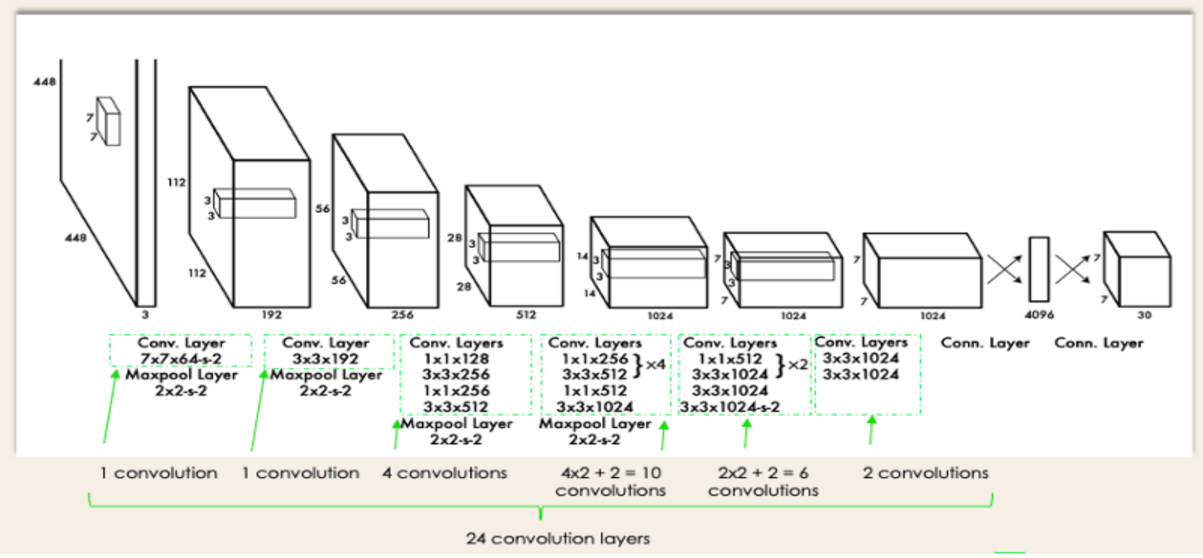
Some of the reasons why YOLO is leading the competition include its:

- Speed
- Detection accuracy
- Good generalization
- Open-source

What is YOLO?

You Only Look Once (YOLO) is a state-of-the-art, real-time object detection algorithm introduced in 2015 by Joseph Redmon**,** Santosh Divvala**,** Ross Girshick**, and** Ali Farhadi **in their famous research paper "**You Only Look Once: Unified, Real-Time Object Detection**". You Only Look Once (YOLO) proposes using an end-to-end** **neural network** **that makes predictions of bounding boxes and class probabilities all at once. It differs from the approach taken by previous object detection algorithms, which repurposed classifiers to perform detection.**

YOLO Architecture :



Natural Language Processing :

Natural language processing (NLP) is a subfield of Artificial Intelligence (AI). This is a widely used technology for personal assistants that are used in various business fields/areas. This technology works on the speech provided by the user breaks it down for proper understanding and processes it accordingly. This is a very recent and effective approach due to which it has a really high demand in today’s market. Natural Language Processing is an upcoming field where already many transitions such as compatibility with smart devices, and interactive talks with a human have been made possible. Knowledge representation, logical reasoning, and constraint satisfaction were the emphasis of AI applications in NLP.

NLP is used in a wide range of applications, including machine translation, sentiment analysis, speech recognition, chatbots, and text classification. Some common techniques used in NLP include:

Tokenization	the process of breaking text into individual words or phrases.
Part-of-speech tagging	the process of labeling each word in a sentence with its grammatical part of speech.
Named entity recognition	the process of identifying and categorizing named entities, such as people, places, and organizations, in text.
Sentiment analysis	the process of determining the sentiment of a piece of text, such as whether it is positive, negative, or neutral.
Machine translation	the process of automatically translating text from one language to another.

Common Natural Language Processing (NLP) Task:

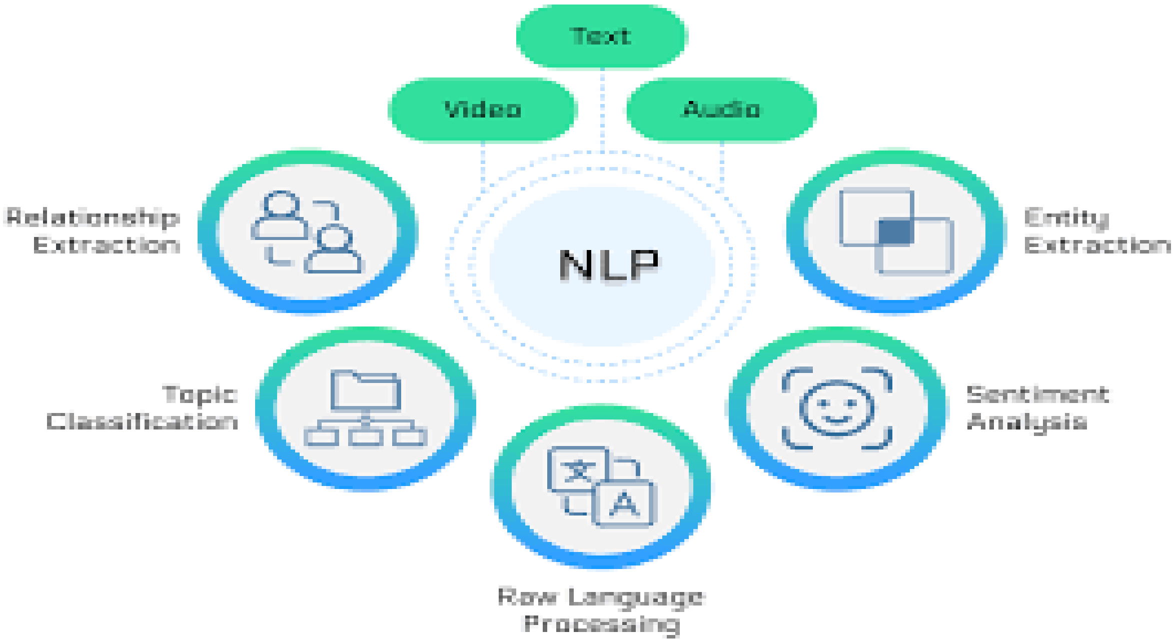
- **Text and speech processing: This includes Speech recognition, text-&-speech processing, encoding(i.e converting speech or text to machine-readable language), etc.**
- **Text classification: This includes Sentiment Analysis in which the machine can analyze the qualities, emotions, and sarcasm from text and also classify it accordingly.**
- **Language generation: This includes tasks such as machine translation, summary writing, essay writing, etc. which aim to produce coherent and fluent text.**
- **Language interaction: This includes tasks such as dialogue systems, voice assistants, and chatbots, which aim to enable natural communication between humans and computers.**

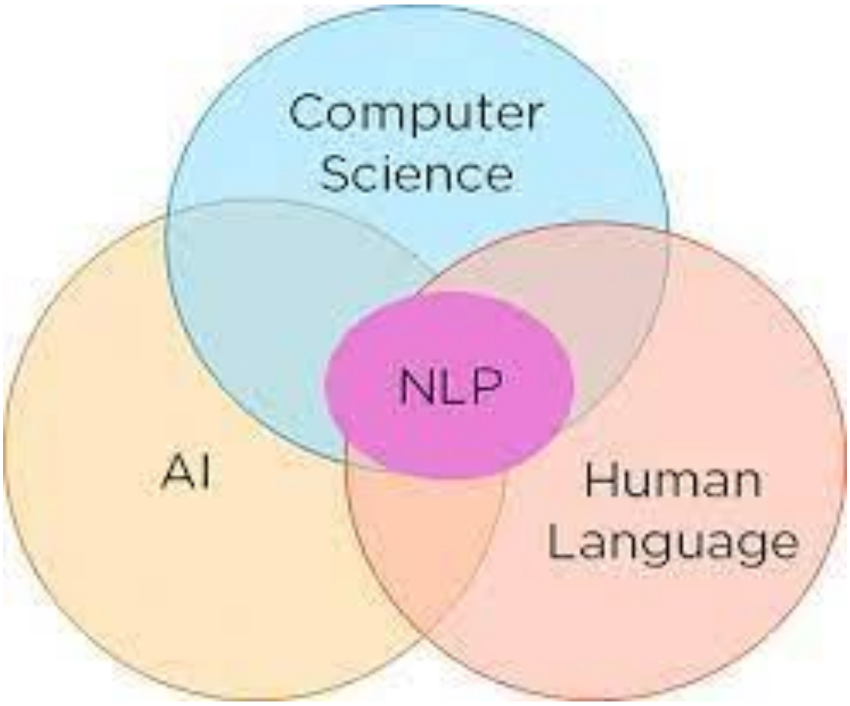
The field is divided into three different parts:

1. **Speech Recognition** — The translation of spoken language into text.
2. **Natural Language Understanding (NLU)** — The computer's ability to understand what we say.
3. **Natural Language Generation (NLG)** — The generation of natural language by a computer.

Applications of Natural Language Processing (NLP):

- **Spam Filters**
- **Algorithmic Trading**
- **Questions Answering**
- **Summarizing Information**





What is Recurrent Neural Network (RNN)?

Recurrent Neural Network(RNN) is a type of Neural Network where the output from the previous step is fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other, but in cases when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words. Thus RNN came into existence, which solved this issue with the help of a Hidden Layer.

Applications of Recurrent Neural Network

1. Language Modelling and Generating Text
2. Speech Recognition
3. Machine Translation
4. Image Recognition, Face detection
5. Time series Forecasting
6.

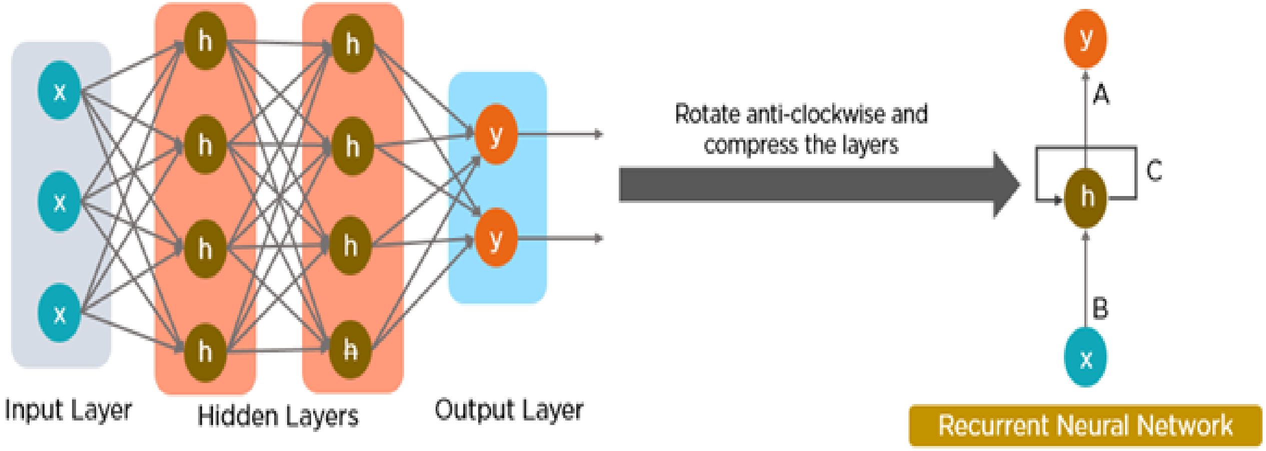
Types Of RNN

There are four types of RNNs based on the number of inputs and outputs in the network.

1. One to One
2. One to Many
3. Many to One
4. Many to Many

Using RNN models and sequence datasets, you may tackle a variety of problems, including :

- Speech recognition
- Generation of music
- Automated Translations
- Analysis of video action
- Sequence study of the genome and DNA



Text Preprocessing:

1. **Text Cleaning:** Clean the text data by removing any special characters, punctuation, and HTML tags if your data comes from web sources. Use libraries like BeautifulSoup or regular expressions for this task.
2. **Tokenization:** Split the text into individual words or tokens. You can use libraries like NLTK or spaCy for tokenization.
3. **Stop Word Removal:** Remove common stop words (e.g., "and," "the," "in") from the text as they don't contribute much to distinguishing between fake and real news.
4. **Lowercasing:** Convert all the text to lowercase to ensure uniformity.
5. **Stemming or Lemmatization:** Reduce words to their root forms. This step is optional, but it can help reduce feature dimensionality and improve model performance. NLTK and spaCy provide tools for stemming and lemmatization.
6. **Vectorization:** Transform the preprocessed text into numerical vectors that machine learning models can understand. You can use techniques like Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), or Word Embeddings (e.g., Word2Vec, GloVe) for this purpose.

Feature Extraction:

Feature Engineering: Consider adding new features, such as the length of the text, the number of unique words, and the presence of certain keywords, to improve the model's performance.
Select Features: Based on the nature of your data and problem, you may need to select the most relevant features using techniques like mutual information, chi-squared, or feature importance from tree-based models.

Model Training:

Choose a Classification Algorithm: Select a machine learning algorithm for the classification task. Common choices include:

- Logistic Regression
- Naïve Bayes
- Random Forest
- Support Vector Machines (SVM)
- Deep Learning (e.g., LSTM or BERT)

Split Data: Split your dataset into training, validation, and test sets. The training set is used to train the model, the validation set is used for hyperparameter tuning, and the test set is used for final evaluation.

Model Training: Train the chosen classification model on the training data. Ensure that you apply appropriate hyperparameter tuning to optimize the model's performance. Cross-validation is often used to fine-tune hyperparameters.

Model Evaluation:

Metrics: Evaluate the model's performance using appropriate metrics, such as accuracy, precision, recall, F1-score, and ROC AUC. Given the imbalanced nature of fake news detection, you may want to pay special attention to precision and recall.

Confusion Matrix: Visualize the confusion matrix to understand how well your model is distinguishing between fake and real news.

Cross-Validation: Perform k-fold cross-validation to assess the model's generalization performance.

Tune and Optimize: If the model performance is not satisfactory, consider adjusting hyperparameters, trying different algorithms, or collecting more data.

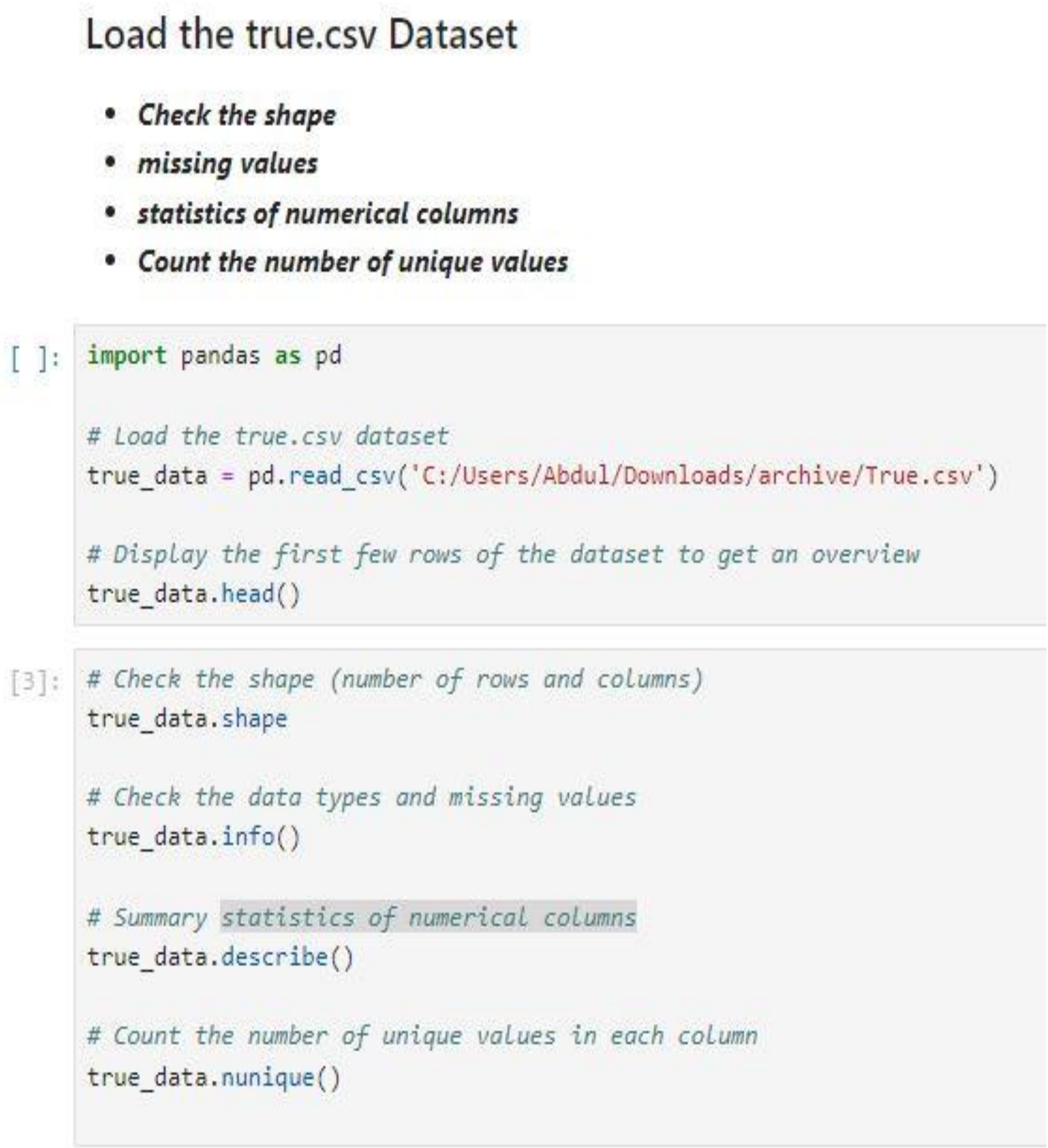
Ensemble Methods: Experiment with ensemble methods like bagging (e.g., Random Forest) or boosting (e.g., AdaBoost, XGBoost) to potentially improve your model's performance.

Bias and Fairness: Be aware of and mitigate potential biases in your dataset and model, as well as ensuring fairness in the model's predictions.

Deployment: Once you're satisfied with your model's performance, deploy it in a production environment for real-time or batch processing.

Execution : (***Phase4.ipynb***)

Load the true.csv Dataset :



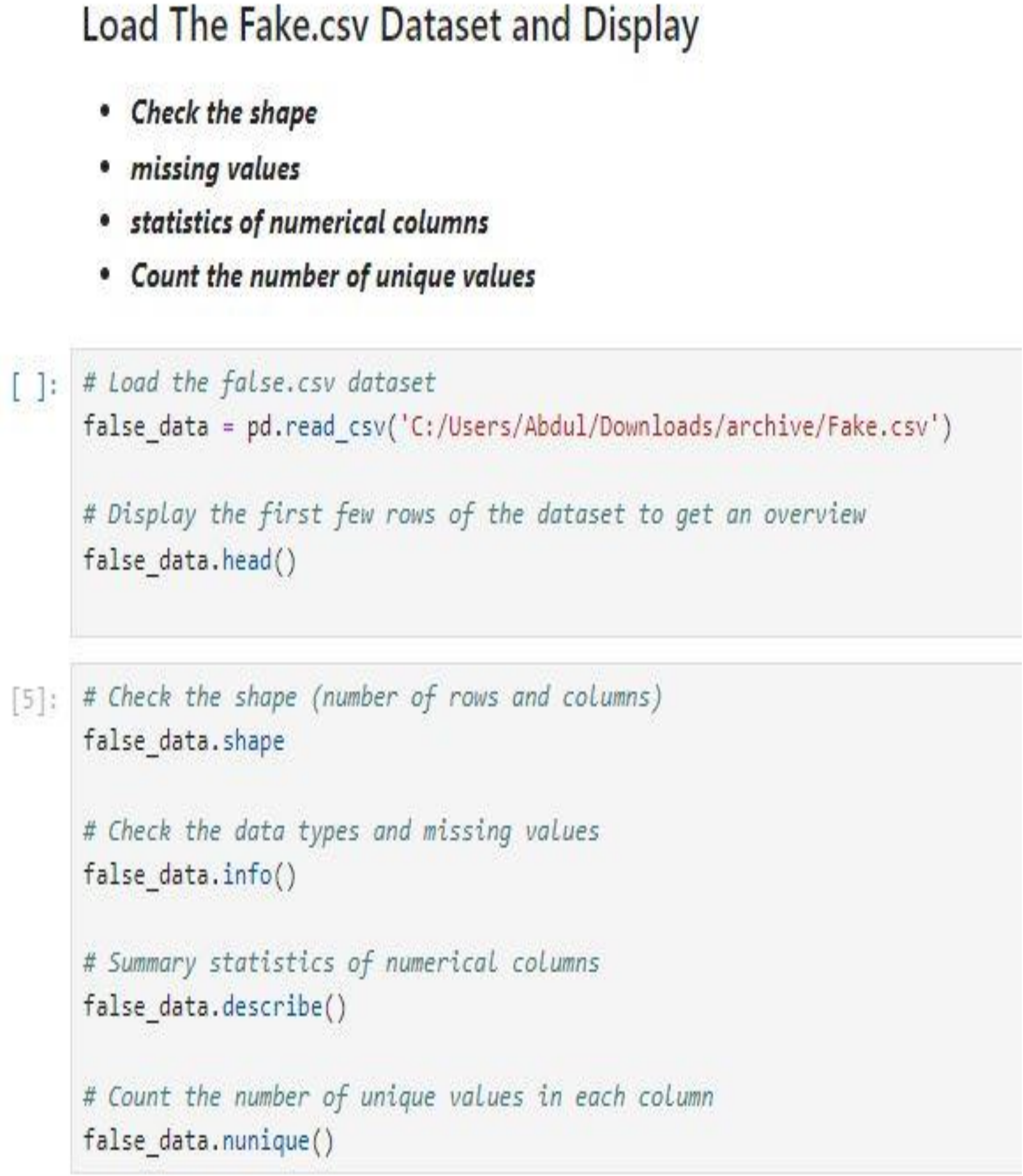
Load the true.csv Dataset

- ***Check the shape***
- ***missing values***
- ***statistics of numerical columns***
- ***Count the number of unique values***

```
[ ]: import pandas as pd

     # Load the true.csv dataset
     true_data = pd.read_csv('C:/Users/Abdul/Downloads/archive/True.csv')

     # Display the first few rows of the dataset to get an overview
     true_data.head()
```

```
[3]: # Check the shape (number of rows and columns)
     true_data.shape

     # Check the data types and missing values
     true_data.info()

     # Summary statistics of numerical columns
     true_data.describe()

     # Count the number of unique values in each column
     true_data.nunique()
```

Load The Fake.csv Dataset and Display :



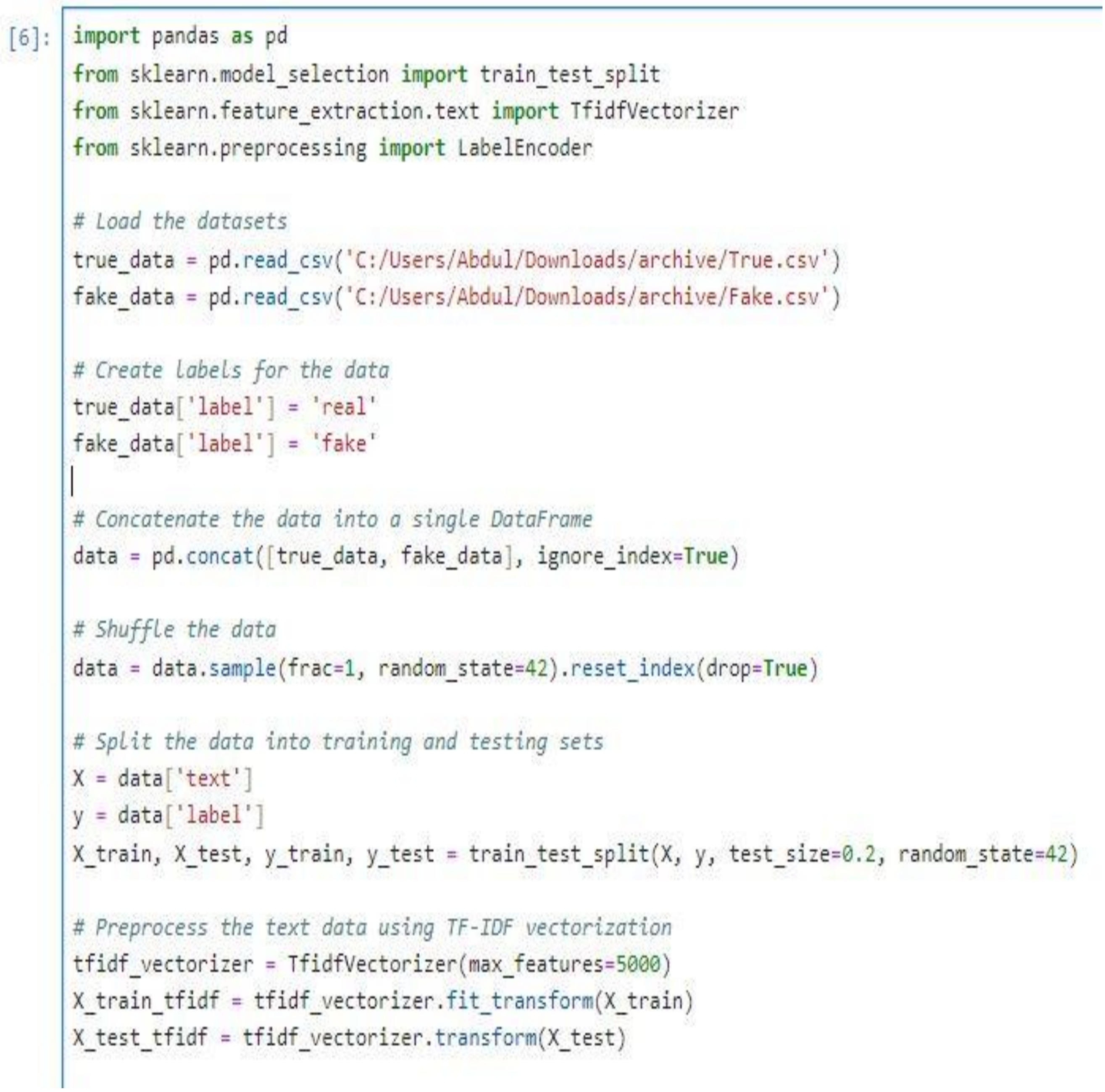
Load The Fake.csv Dataset and Display

- *Check the shape*
- *missing values*
- *statistics of numerical columns*
- *Count the number of unique values*

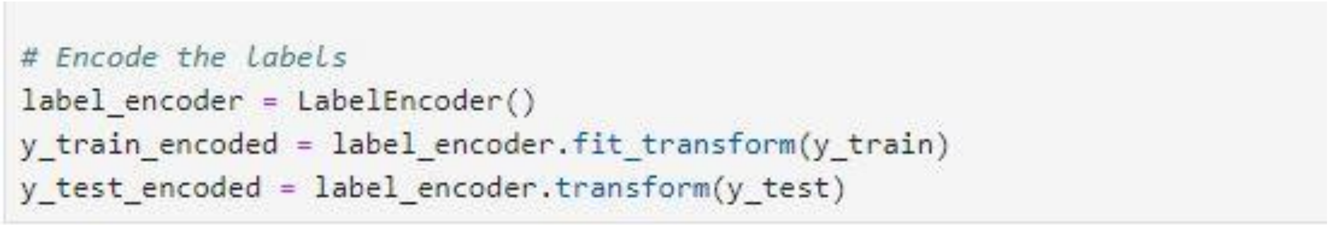
```
[ ]: # Load the false.csv dataset
     false_data = pd.read_csv('C:/Users/Abdul/Downloads/archive/Fake.csv')

     # Display the first few rows of the dataset to get an overview
     false_data.head()
```

```
[5]: # Check the shape (number of rows and columns)
     false_data.shape

     # Check the data types and missing values
     false_data.info()

     # Summary statistics of numerical columns
     false_data.describe()

     # Count the number of unique values in each column
     false_data.nunique()
```

Data Preprocessing :

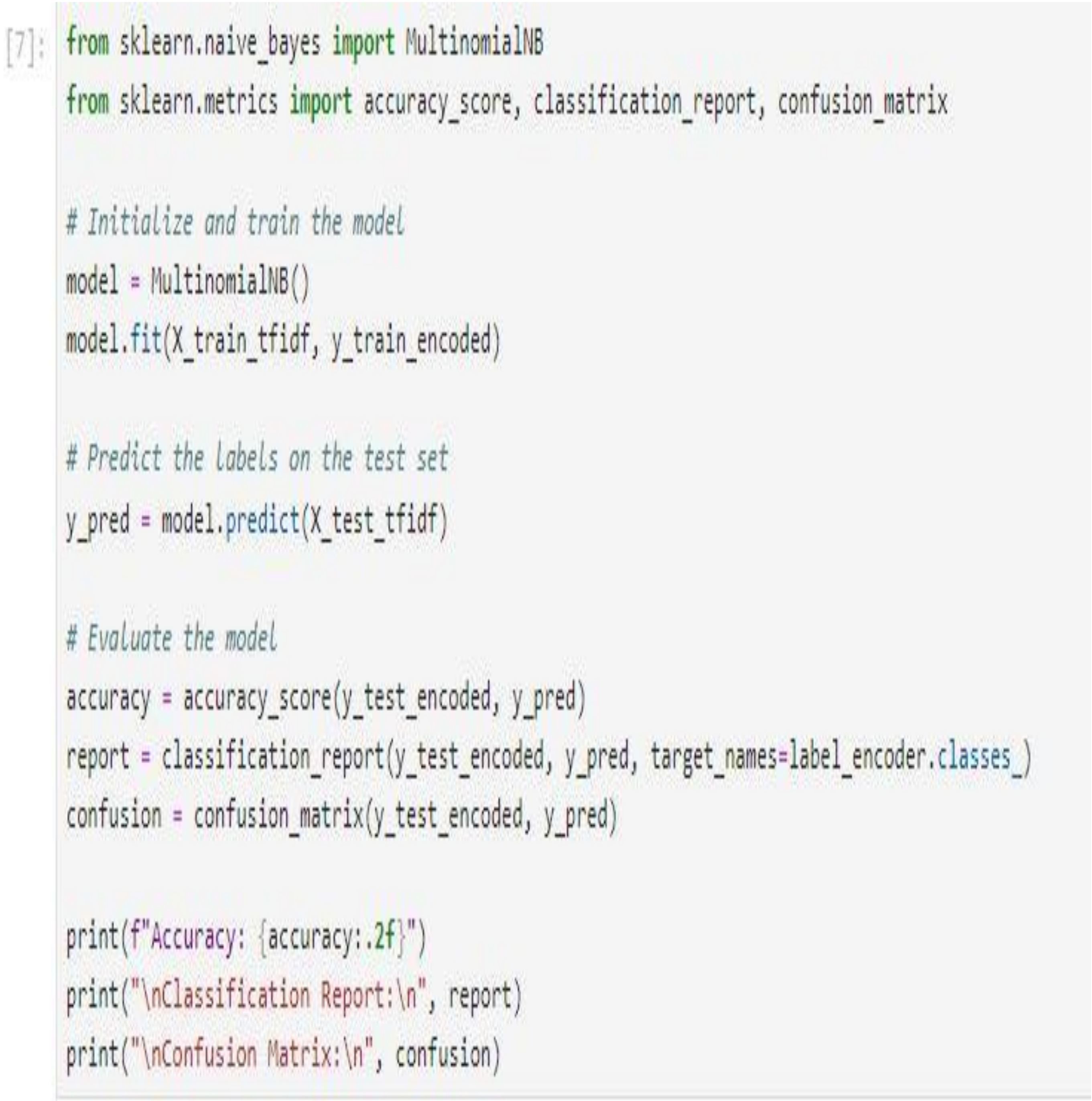


```
[6]: import pandas as pd
     from sklearn.model_selection import train_test_split
     from sklearn.feature_extraction.text import TfidfVectorizer
     from sklearn.preprocessing import LabelEncoder

     # Load the datasets
     true_data = pd.read_csv('C:/Users/Abdul/Downloads/archive/True.csv')
     fake_data = pd.read_csv('C:/Users/Abdul/Downloads/archive/Fake.csv')

     # Create labels for the data
     true_data['label'] = 'real'
     fake_data['label'] = 'fake'

     # Concatenate the data into a single DataFrame
     data = pd.concat([true_data, fake_data], ignore_index=True)

     # Shuffle the data
     data = data.sample(frac=1, random_state=42).reset_index(drop=True)

     # Split the data into training and testing sets
     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)

     # Preprocess the text data using TF-IDF vectorization
     tfidf_vectorizer = TfidfVectorizer(max_features=5000)
     X_train_tfidf = tfidf_vectorizer.fit_transform(X_train)
     X_test_tfidf = tfidf_vectorizer.transform(X_test)
```



```
# Encode the labels
label_encoder = LabelEncoder()
y_train_encoded = label_encoder.fit_transform(y_train)
y_test_encoded = label_encoder.transform(y_test)
```

Text Classification , Model Training , Model Evaluation :



```
[7]: from sklearn.naive_bayes import MultinomialNB
     from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

     # Initialize and train the model
     model = MultinomialNB()
     model.fit(X_train_tfidf, y_train_encoded)

     # Predict the labels on the test set
     y_pred = model.predict(X_test_tfidf)

     # Evaluate the model
     accuracy = accuracy_score(y_test_encoded, y_pred)
     report = classification_report(y_test_encoded, y_pred, target_names=label_encoder.classes_)
     confusion = confusion_matrix(y_test_encoded, y_pred)

     print(f"Accuracy: {accuracy:.2f}")
     print("\nClassification Report:\n", report)
     print("\nConfusion Matrix:\n", confusion)
```

Conclusion :

In this part of your project, you've loaded and preprocessed the fake news dataset. Preprocessing is a critical step in building a reliable fake news detection model. After preprocessing, you can proceed to feature engineering, model selection, training, and evaluation. Your choice of machine learning or deep learning algorithms will depend on the dataset size and complexity. Make sure to continuously monitor and fine-tune your model to improve its performance. Lastly, be prepared to handle real-world scenarios where fake news can evolve and adapt, so regular updates and retraining might be necessary for a robust solution.