# Notre Dame University Bangladesh

# Artificial Intelligence Lab
## CSE4104

# Fake News Prediction

## Submitted To :

**Humayara Binte Rashid**
Lecturer
Department of CSE

## Submitted By :

**Abrity Paul Chowdhury**
ID : 0692130005101005
Batch : CSE 17

Submission Date: November 20, 2024

# Contents

## 0.1 Introduction

Fake news refers to false or misleading information presented as news, often created to influence public opinion or generate web traffic. The spread of fake news has become a significant concern, impacting societal trust and political stability. With the growing volume of online news, developing models that can accurately identify and filter out fake news is crucial to ensure the reliability of information consumed by the public.

## 0.2 Objective

To develop a machine learning model that can accurately classify news articles as real or fake, several key steps are involved. First, the news data must be pre-processed and cleaned to ensure it's suitable for analysis. Next, meaningful features are extracted from the text using natural language processing techniques. These features are then used to train and evaluate machine learning models to achieve optimal performance. Finally, different models are compared to select the most effective one for accurately identifying fake news.

## 0.3 Motivation

The rapid spread of misinformation can have a profound social impact, undermining public trust in media and even influencing political outcomes. Given the overwhelming volume of information, manually verifying news articles is impractical, highlighting the need for automated detection systems that offer a scalable and efficient solution. Advances in machine learning and natural language processing have made the development of robust fake news detection systems increasingly feasible, offering new ways to combat misinformation effectively.

## 0.4 Related works

### 0.4.1 Related works-1

**Fake news detection using discourse segment structure analysis** [5]
**Contribution :**
Here the author introduced a novel discourse-level approach for detecting fake news using deep learning, leveraging a Bidirectional GRU to analyze the hierarchical structure of content. This approach achieved an accuracy of 74.62% and an F1 score of 0.76, demonstrating its effectiveness in distinguishing fake news based on discourse analysis.
**Tools: Python, Google Colab, Word2Vec, GRU models**

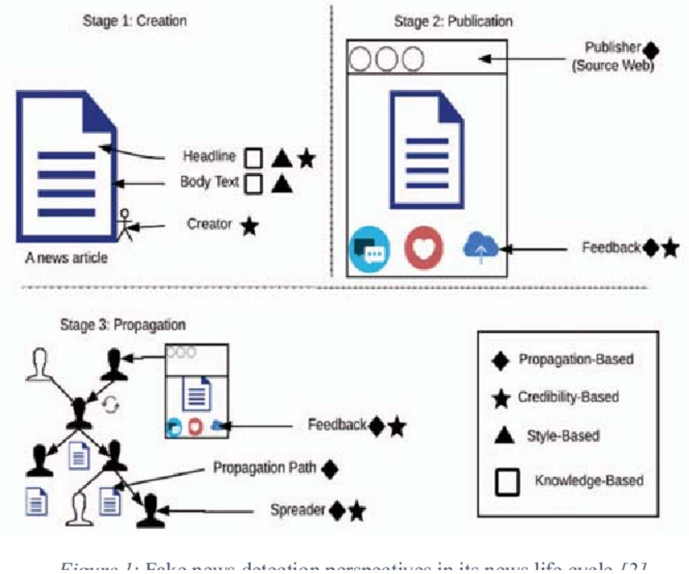


**Figure1:Stages of Approach**

**Limitations:**
The model faces limitations due to a lack of dataset diversity, which risks introducing bias into the results. Its performance could be enhanced by incorporating more advanced models or expanding data coverage to better represent various news sources. Additionally, the model may struggle to generalize effectively across different news contexts, potentially impacting its reliability in diverse real-world scenarios.

### 0.4.2 Related works-2

**Fake News Detection Using Machine Learning Approaches**[1]
**Contribution:**
They developed a fake news detection model leveraging algorithms such as XGBoost, Random Forest, and Naive Bayes. To enhance the feature extraction process, we applied NLP techniques like tokenization and TF-IDF. The model reached an accuracy of up to 75% with its performance assessed using confusion matrices.
**Tools**: Python, scikit-learn, NLTK, Anaconda, and the LIAR Dataset.

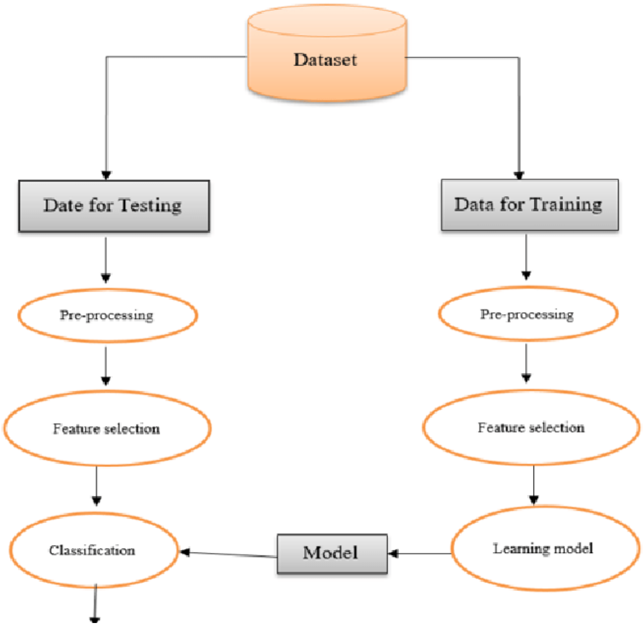


**Figure2:Working Procedure**

**Limitations :**
The model's accuracy is currently limited to 75% indicating room for improvement. Additionally, it may not generalize well beyond political news, limiting its broader applicability. The model's performance also relies heavily on effective feature extraction, which is crucial for accurate classification.

## 0.5 Why consider it AI?

This project is considered an AI endeavor because it uses machine learning techniques to automate the identification of fake and real news. It employs natural language processing (NLP) to process and analyze textual data, extracting patterns and features that are not immediately obvious to humans. The system mimics human cognitive abilities by learning from vast amounts of labeled news articles and generalizing this knowledge to predict the credibility of unseen news. By training models like Naive Bayes, Logistic Regression, or advanced neural networks, the AI can classify news articles with high accuracy. AI brings scalability to this task, enabling the evaluation of thousands of articles in seconds, which would be impossible for humans to achieve manually. The project demonstrates how AI can aid in combating misinformation, a critical challenge in today's digital age. It not only reduces human workload but also improves the efficiency and reliability of misinformation detection systems. The use of advanced AI techniques, such as deep learning and transformers, further enhances the model's ability to understand complex patterns in language. This real-world application of AI has significant societal impact, helping maintain trust in journalism and curbing the spread of false information. By automating decision-making in this domain, the project exemplifies the power of AI in addressing pressing global challenges.

## 0.6 Workflow

Here is the procedure that will be followed by the proposed approach .

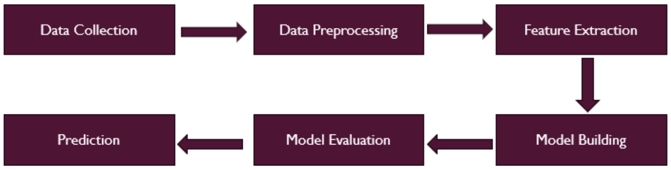


**Figure3:Workflow of proposed project**

## 0.7 Tools

**Python Libraries** [4]
The project utilizes several Python libraries to handle various tasks. **Pandas**is used for data manipulation and pre-processing, while **NumPy** supports numerical operations and array handling. **Scikit-learn** is employed for model building, data splitting, and performance evaluation. **NLTK** facilitates natural language processing tasks, including tokenization and stopword removal. Additionally, the **re** library is used for regular expressions to clean and process text efficiently.
**Machine Learning Models** [3]
The model uses several algorithms for classification tasks.**Logistic Regression** serves as a common baseline for binary classification problems. The **Naive Bayes Classifier** is often employed for text classification tasks due to its simplicity and efficiency. **Support Vector Machines (SVM)** are utilized for better performance on certain datasets, especially when the data is not linearly separable. **Random Forest** is used for ensemble-based predictions, providing greater robustness and accuracy by combining multiple decision trees.
**Text Processing Tools** [2]
The model utilizes several techniques for text pre-processing and feature extraction. **CountVectorizer** is used to convert text into a matrix of token counts, representing the frequency of each word in the document. **TfidfVectorizer** transforms the text into Term Frequency-Inverse Document Frequency

(TF-IDF) features, which help weigh words based on their importance across different documents. Additionally, **NLTK** is employed for text processing and cleaning tasks, such as removing stop words and performing stemming to reduce words to their base form.

## 0.8 Methodology

**The proposed diagram**

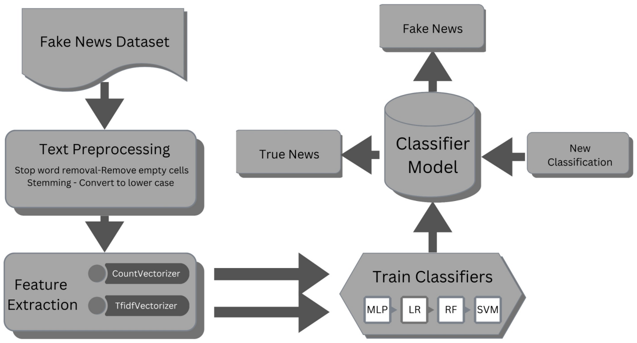


**Figure4:Diagram of proposed process of fake news prediction**

**The updated diagram**

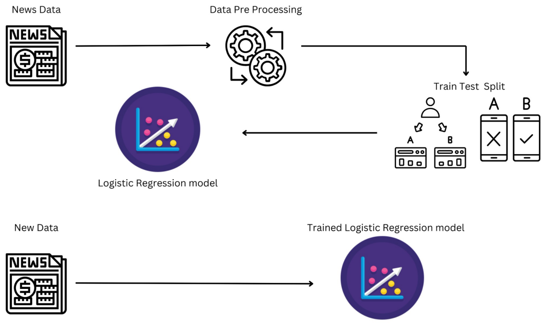
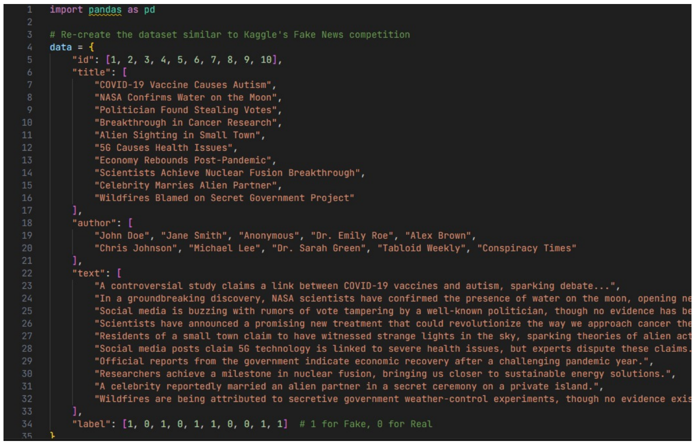


**Figure5:Diagram of updated process of fake news prediction**

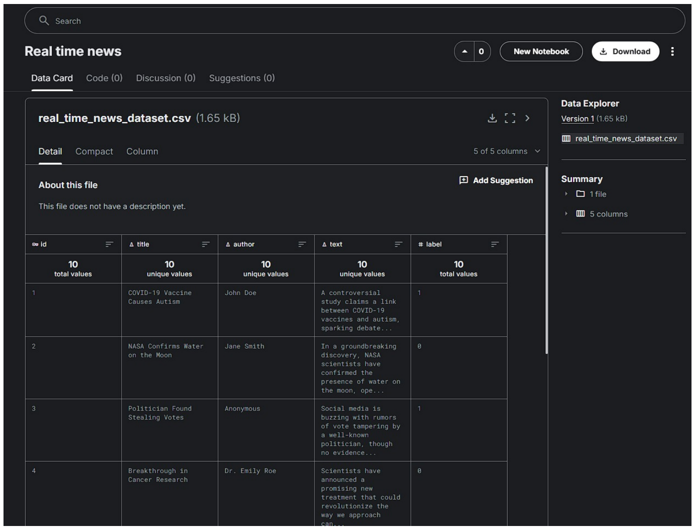
## 0.9 Code & Result

### Dataset [6]



```python
import pandas as pd

# Re-create the dataset similar to Kaggle's Fake News competition
data = {
    "id": [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
    "title": [
        "COVID-19 Vaccine Causes Autism",
        "NASA Confirms Water on the Moon",
        "Politician Found Stealing Votes",
        "Breakthrough in Cancer Research",
        "Alien Sighting in Small Town",
        "5G Causes Health Issues",
        "Economy Rebounds Post-Pandemic",
        "Scientists Achieve Nuclear Fusion Breakthrough",
        "Celebrity Marries Alien Partner",
        "Wildfires Blamed on Secret Government Project"
    ],
    "author": [
        "John Doe", "Jane Smith", "Anonymous", "Dr. Emily Roe", "Alex Brown",
        "Chris Johnson", "Michael Lee", "Dr. Sarah Green", "Tabloid Weekly", "Conspiracy Times"
    ],
    "text": [
        "A controversial study claims a link between COVID-19 vaccines and autism, sparking debate...",
        "In a groundbreaking discovery, NASA scientists have confirmed the presence of water on the moon, opening ne
        "Social media is buzzing with rumors of vote tampering by a well-known politician, though no evidence has be
        "Scientists have announced a promising new treatment that could revolutionize the way we approach cancer the
        "Residents of a small town claim to have witnessed strange lights in the sky, sparking theories of alien act
        "Social media posts claim 5G technology is linked to severe health issues, but experts dispute these claims.
        "Official reports from the government indicate economic recovery after a challenging pandemic year.",
        "Researchers achieve a milestone in nuclear fusion, bringing us closer to sustainable energy solutions.",
        "A celebrity reportedly married an alien partner in a secret ceremony on a private island.",
        "Wildfires are being attributed to secretive government weather-control experiments, though no evidence exis
    ],
    "label": [1, 0, 1, 0, 1, 1, 0, 0, 1, 1]  # 1 for Fake, 0 for Real
}
```

Figure6:Code for Dataset



Real time news

real_time_news_dataset.csv (1.65 kB)

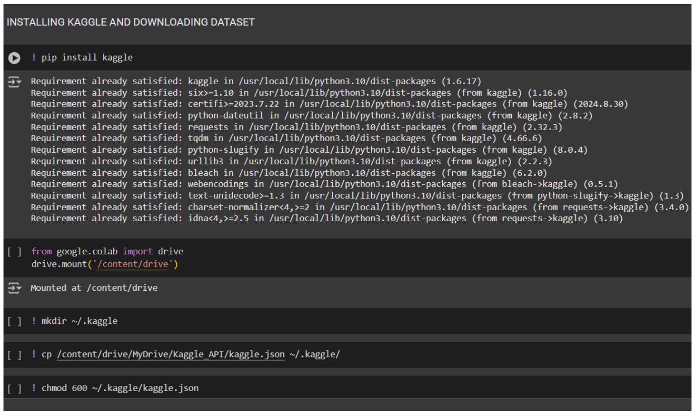
About this file

This file does not have a description yet.

| id | title | author | text | label |
|---|---|---|---|---|
| 10<br>total values | 10<br>unique values | 10<br>unique values | 10<br>unique values | 10<br>total values |
| 1 | COVID-19 Vaccine Causes Autism | John Doe | A controversial study claims a link between COVID-19 vaccines and autism, sparking debate... | 1 |
| 2 | NASA Confirms Water on the Moon | Jane Smith | In a groundbreaking discovery, NASA scientists have confirmed the presence of water on the moon, ope... | 0 |
| 3 | Politician Found Stealing Votes | Anonymous | Social media is buzzing with rumors of vote tampering by a well-known politician, though no evidence... | 1 |
| 4 | Breakthrough in Cancer Research | Dr. Emily Roe | Scientists have announced a promising new treatment that could revolutionize the way we approach can... | 0 |

Figure7:Dataset in Kaggle

## Main Code



INSTALLING KAGGLE AND DOWNLOADING DATASET

```
! pip install kaggle
```

```
Requirement already satisfied: kaggle in /usr/local/lib/python3.10/dist-packages (1.6.17)
Requirement already satisfied: six>=1.10 in /usr/local/lib/python3.10/dist-packages (from kaggle) (1.16.0)
Requirement already satisfied: certifi>=2023.7.22 in /usr/local/lib/python3.10/dist-packages (from kaggle) (2024.8.30)
Requirement already satisfied: python-dateutil in /usr/local/lib/python3.10/dist-packages (from kaggle) (2.8.2)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from kaggle) (2.32.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from kaggle) (4.66.6)
Requirement already satisfied: python-slugify in /usr/local/lib/python3.10/dist-packages (from kaggle) (8.0.4)
Requirement already satisfied: urllib3 in /usr/local/lib/python3.10/dist-packages (from kaggle) (2.2.3)
Requirement already satisfied: bleach in /usr/local/lib/python3.10/dist-packages (from kaggle) (6.2.0)
Requirement already satisfied: webencodings in /usr/local/lib/python3.10/dist-packages (from bleach->kaggle) (0.5.1)
Requirement already satisfied: text-unidecode>=1.3 in /usr/local/lib/python3.10/dist-packages (from python-slugify->kaggle) (1.3)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->kaggle) (3.4.0)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->kaggle) (3.10)
```

```
from google.colab import drive
drive.mount('/content/drive')
```

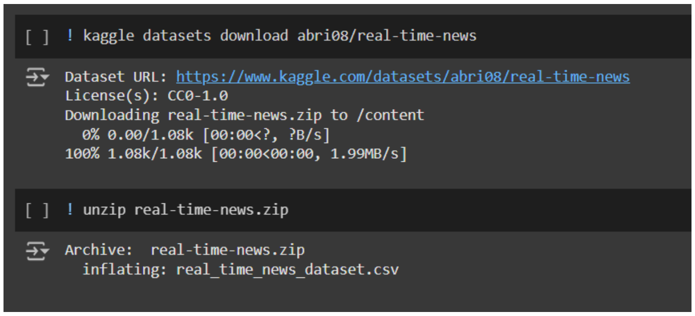
```
Mounted at /content/drive
```

```
! mkdir ~/.kaggle
```

```
! cp /content/drive/MyDrive/Kaggle_API/kaggle.json ~/.kaggle/
```

```
! chmod 600 ~/.kaggle/kaggle.json
```

**Figure8:Installing kaggle and downloading dataset in google colab part1**



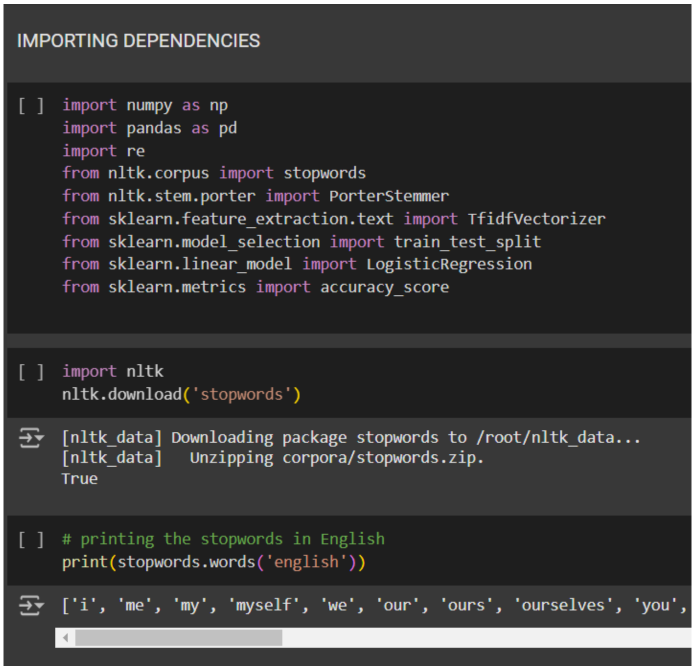
```
! kaggle datasets download abri08/real-time-news
```

```
Dataset URL: https://www.kaggle.com/datasets/abri08/real-time-news
License(s): CC0-1.0
Downloading real-time-news.zip to /content
  0% 0.00/1.08k [00:00<?, ?B/s]
100% 1.08k/1.08k [00:00<00:00, 1.99MB/s]
```

```
! unzip real-time-news.zip
```

```
Archive:  real-time-news.zip
  inflating: real_time_news_dataset.csv
```

**Figure9:Installing kaggle and downloading dataset in google colab part2**



IMPORTING DEPENDENCIES

```
import numpy as np
import pandas as pd
import re
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

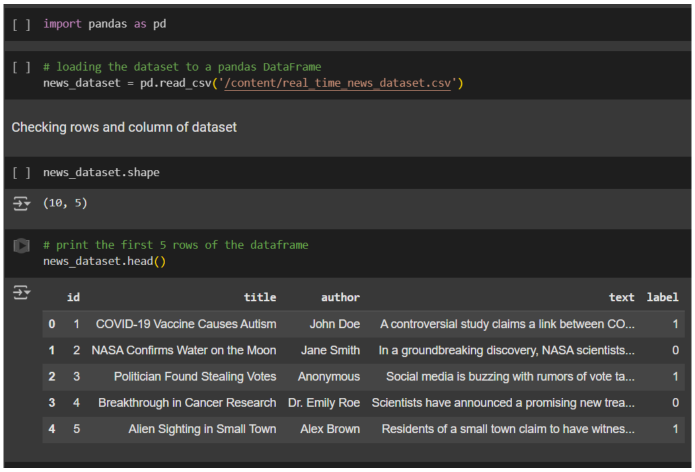
```
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
True
```

```
# printing the stopwords in English
print(stopwords.words('english'))
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you',
```

**Figure10:Importing dependencies**



```
import pandas as pd
```

```
# loading the dataset to a pandas DataFrame
news_dataset = pd.read_csv('/content/real_time_news_dataset.csv')
```

Checking rows and column of dataset

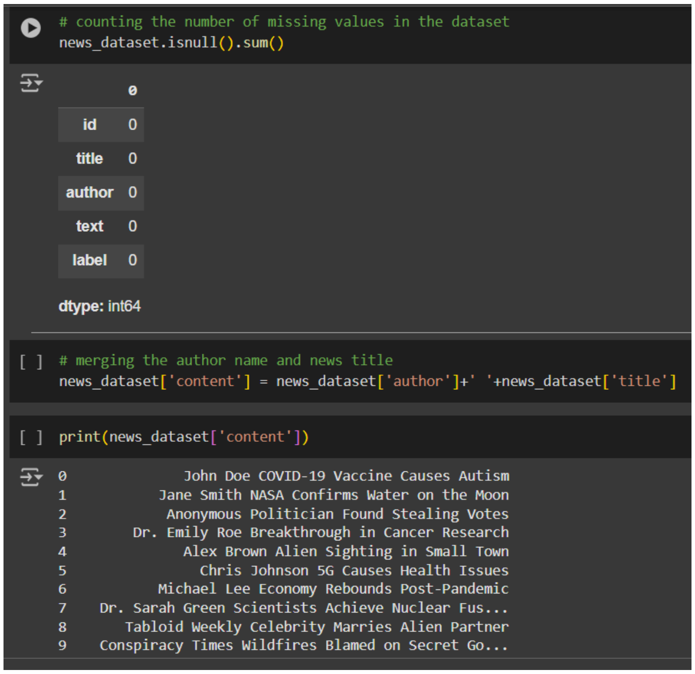
```
news_dataset.shape
```

```
(10, 5)
```

```
# print the first 5 rows of the dataframe
news_dataset.head()
```

| | id | title | author | text | label |
|---|---|---|---|---|---|
| 0 | 1 | COVID-19 Vaccine Causes Autism | John Doe | A controversial study claims a link between CO... | 1 |
| 1 | 2 | NASA Confirms Water on the Moon | Jane Smith | In a groundbreaking discovery, NASA scientists... | 0 |
| 2 | 3 | Politician Found Stealing Votes | Anonymous | Social media is buzzing with rumors of vote ta... | 1 |
| 3 | 4 | Breakthrough in Cancer Research | Dr. Emily Roe | Scientists have announced a promising new trea... | 0 |
| 4 | 5 | Alien Sighting in Small Town | Alex Brown | Residents of a small town claim to have witnes... | 1 |

**Figure11:Data pre processing**



```
# counting the number of missing values in the dataset
news_dataset.isnull().sum()
```

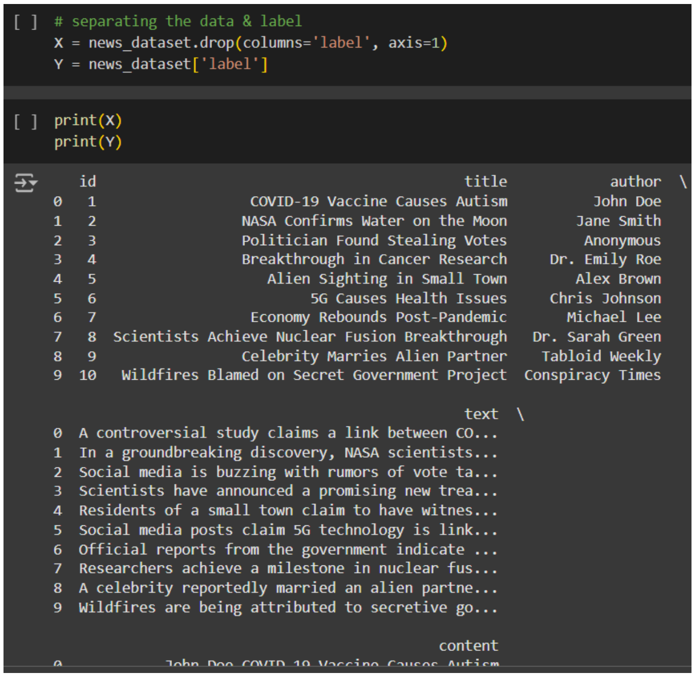
```
         0
    id   0
 title   0
author   0
  text   0
 label   0

dtype: int64
```

```
[ ] # merging the author name and news title
    news_dataset['content'] = news_dataset['author']+' '+news_dataset['title']
```

```
[ ] print(news_dataset['content'])
```

```
0              John Doe COVID-19 Vaccine Causes Autism
1           Jane Smith NASA Confirms Water on the Moon
2            Anonymous Politician Found Stealing Votes
3         Dr. Emily Roe Breakthrough in Cancer Research
4              Alex Brown Alien Sighting in Small Town
5                Chris Johnson 5G Causes Health Issues
6           Michael Lee Economy Rebounds Post-Pandemic
7     Dr. Sarah Green Scientists Achieve Nuclear Fus...
8       Tabloid Weekly Celebrity Marries Alien Partner
9     Conspiracy Times Wildfires Blamed on Secret Go...
```

**Figure12:Checking missing values and column merging**

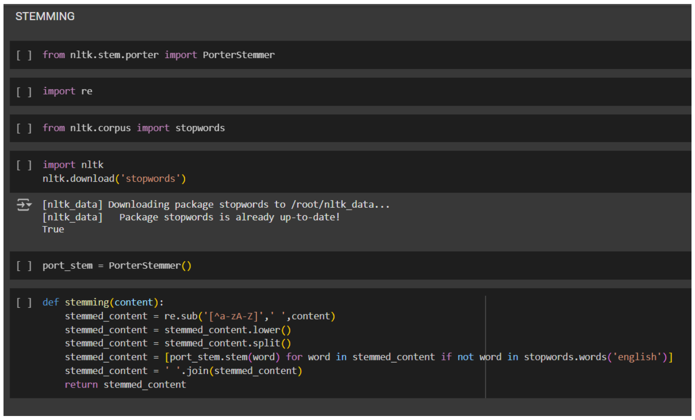


```
[ ] # separating the data & label
    X = news_dataset.drop(columns='label', axis=1)
    Y = news_dataset['label']
```

```
[ ] print(X)
    print(Y)
```

```
   id                                      title            author  \
0   1                COVID-19 Vaccine Causes Autism          John Doe
1   2               NASA Confirms Water on the Moon        Jane Smith
2   3               Politician Found Stealing Votes         Anonymous
3   4               Breakthrough in Cancer Research     Dr. Emily Roe
4   5                  Alien Sighting in Small Town        Alex Brown
5   6                     5G Causes Health Issues     Chris Johnson
6   7                Economy Rebounds Post-Pandemic       Michael Lee
7   8  Scientists Achieve Nuclear Fusion Breakthrough  Dr. Sarah Green
8   9               Celebrity Marries Alien Partner    Tabloid Weekly
9  10   Wildfires Blamed on Secret Government Project Conspiracy Times

                                               text  \
0  A controversial study claims a link between CO...
1  In a groundbreaking discovery, NASA scientists...
2  Social media is buzzing with rumors of vote ta...
3  Scientists have announced a promising new trea...
4  Residents of a small town claim to have witnes...
5  Social media posts claim 5G technology is link...
6  Official reports from the government indicate ...
7  Researchers achieve a milestone in nuclear fus...
8  A celebrity reportedly married an alien partne...
9  Wildfires are being attributed to secretive go...

                                         content
0        John Doe COVID-19 Vaccine Causes Autism
```

**Figure13:Separating data and label**



STEMMING

```
[ ] from nltk.stem.porter import PorterStemmer
```

```
[ ] import re
```

```
[ ] from nltk.corpus import stopwords
```

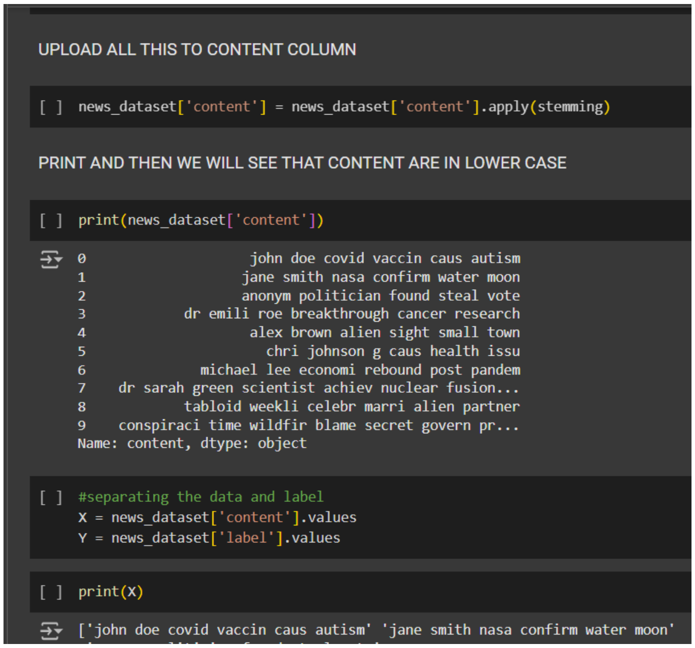
```
[ ] import nltk
    nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
True
```

```
[ ] port_stem = PorterStemmer()
```

```
[ ] def stemming(content):
        stemmed_content = re.sub('[^a-zA-Z]',' ',content)
        stemmed_content = stemmed_content.lower()
        stemmed_content = stemmed_content.split()
        stemmed_content = [port_stem.stem(word) for word in stemmed_content if not word in stopwords.words('english')]
        stemmed_content = ' '.join(stemmed_content)
        return stemmed_content
```

Figure14:Stemming



UPLOAD ALL THIS TO CONTENT COLUMN

```
[ ] news_dataset['content'] = news_dataset['content'].apply(stemming)
```

PRINT AND THEN WE WILL SEE THAT CONTENT ARE IN LOWER CASE

```
[ ] print(news_dataset['content'])
```

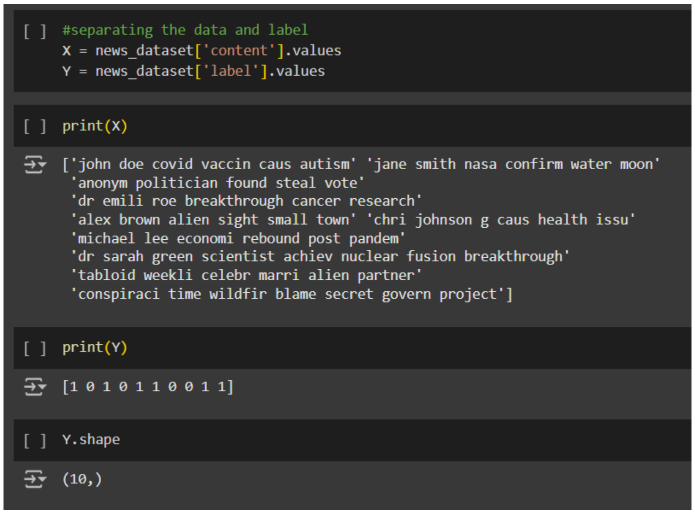
```
0                       john doe covid vaccin caus autism
1                      jane smith nasa confirm water moon
2                      anonym politician found steal vote
3               dr emili roe breakthrough cancer research
4                       alex brown alien sight small town
5                         chri johnson g caus health issu
6                  michael lee economi rebound post pandem
7        dr sarah green scientist achiev nuclear fusion...
8               tabloid weekli celebr marri alien partner
9        conspiraci time wildfir blame secret govern pr...
Name: content, dtype: object
```

```
[ ] #separating the data and label
    X = news_dataset['content'].values
    Y = news_dataset['label'].values
```

```
[ ] print(X)
```

```
['john doe covid vaccin caus autism' 'jane smith nasa confirm water moon'
```

Figure15:Stemming function used in content



```
#separating the data and label
X = news_dataset['content'].values
Y = news_dataset['label'].values
```

```
print(X)
```

```
['john doe covid vaccin caus autism' 'jane smith nasa confirm water moon'
 'anonym politician found steal vote'
 'dr emili roe breakthrough cancer research'
 'alex brown alien sight small town' 'chri johnson g caus health issu'
 'michael lee economi rebound post pandem'
 'dr sarah green scientist achiev nuclear fusion breakthrough'
 'tabloid weekli celebr marri alien partner'
 'conspiraci time wildfir blame secret govern project']
```

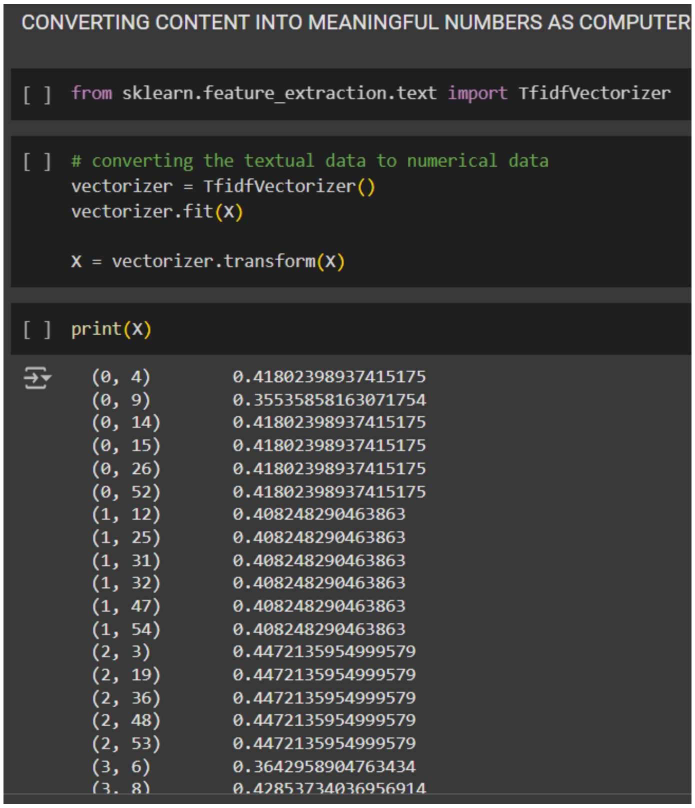
```
print(Y)
```

```
[1 0 1 0 1 1 0 0 1 1]
```

```
Y.shape
```

```
(10,)
```

Figure16:Separating content and label



CONVERTING CONTENT INTO MEANINGFUL NUMBERS AS COMPUTER

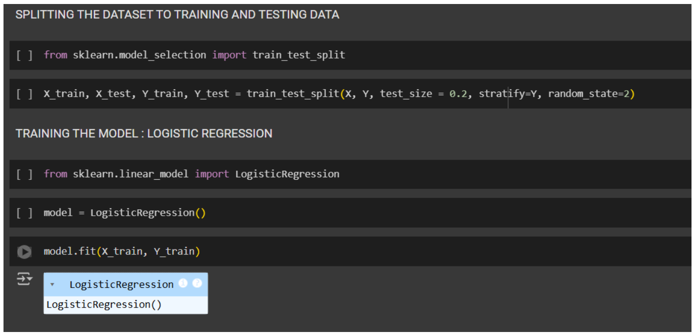
```
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
# converting the textual data to numerical data
vectorizer = TfidfVectorizer()
vectorizer.fit(X)

X = vectorizer.transform(X)
```

```
print(X)
```

```
  (0, 4)	0.41802398937415175
  (0, 9)	0.35535858163071754
  (0, 14)	0.41802398937415175
  (0, 15)	0.41802398937415175
  (0, 26)	0.41802398937415175
  (0, 52)	0.41802398937415175
  (1, 12)	0.408248290463863
  (1, 25)	0.408248290463863
  (1, 31)	0.408248290463863
  (1, 32)	0.408248290463863
  (1, 47)	0.408248290463863
  (1, 54)	0.408248290463863
  (2, 3)	0.4472135954999579
  (2, 19)	0.4472135954999579
  (2, 36)	0.4472135954999579
  (2, 48)	0.4472135954999579
  (2, 53)	0.4472135954999579
  (3, 6)	0.3642958904763434
  (3, 8)	0.42853734036956914
```

Figure17:Tfidfvectorizer



SPLITTING THE DATASET TO TRAINING AND TESTING DATA

```
[ ] from sklearn.model_selection import train_test_split
```

```
[ ] X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, stratify=Y, random_state=2)
```

TRAINING THE MODEL : LOGISTIC REGRESSION

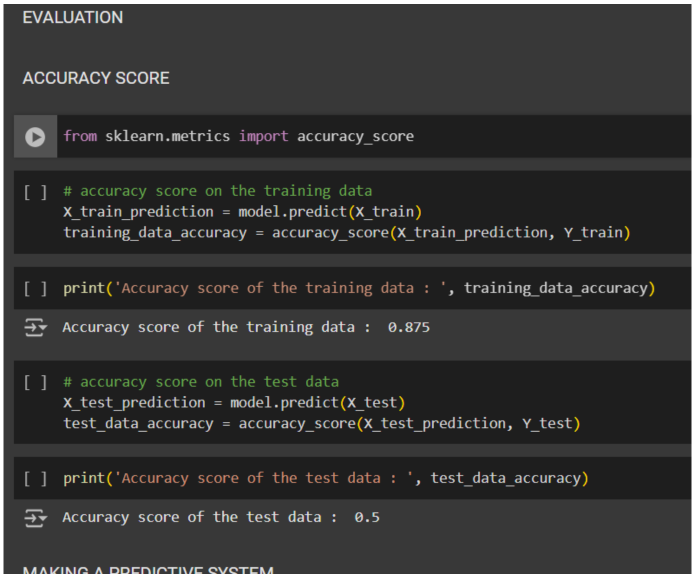
```
[ ] from sklearn.linear_model import LogisticRegression
```

```
[ ] model = LogisticRegression()
```

```
model.fit(X_train, Y_train)
```

```
LogisticRegression
LogisticRegression()
```

Figure18:Splitting and training



EVALUATION

ACCURACY SCORE

```
from sklearn.metrics import accuracy_score
```

```
[ ] # accuracy score on the training data
    X_train_prediction = model.predict(X_train)
    training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

```
[ ] print('Accuracy score of the training data : ', training_data_accuracy)
```

```
Accuracy score of the training data :  0.875
```

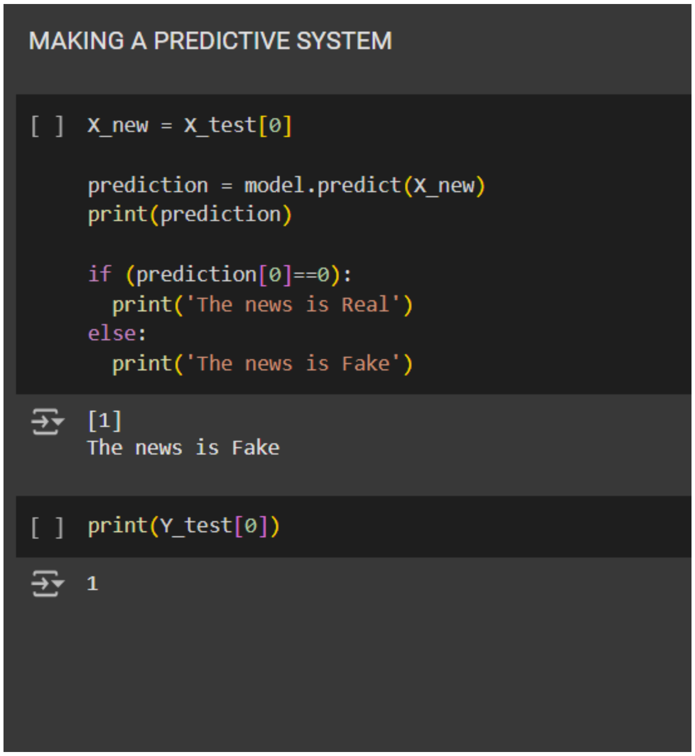
```
[ ] # accuracy score on the test data
    X_test_prediction = model.predict(X_test)
    test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

```
[ ] print('Accuracy score of the test data : ', test_data_accuracy)
```

```
Accuracy score of the test data :  0.5
```

MAKING A PREDICTIVE SYSTEM

Figure19:Accuracy


MAKING A PREDICTIVE SYSTEM

```
[ ] X_new = X_test[0]

    prediction = model.predict(X_new)
    print(prediction)

    if (prediction[0]==0):
      print('The news is Real')
    else:
      print('The news is Fake')
```

```
[1]
The news is Fake
```

```
[ ] print(Y_test[0])
```

```
1
```

**Figure20:Prediction**

## 0.10 Conclusion

The project aims to leverage machine learning and NLP techniques to build an efficient fake news detection system. The model can be improved with more diverse data and advanced algorithms like deep learning .As there are some limitations on logistic regression as well as working with real time data . There are some work that needs to be done. So, in future i will work to improve in this specific sectors . Hopefully it will enhance the reliability of information and contributes to combating the spread of misinformation.

# Bibliography

[1] Zeba Khanam, BN Alwasel, H Sirafi, and Mamoon Rashid. Fake news detection using machine learning approaches. In *IOP conference series: materials science and engineering*, volume 1099, page 012040. IOP Publishing, 2021.

[2] John Levine. *Flex & Bison: Text Processing Tools.* " O'Reilly Media, Inc.", 2009.

[3] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, page 587–601, New York, NY, USA, 2017. Association for Computing Machinery.

[4] I. Stančin and A. Jović. An overview and comparison of free python libraries for data mining and big data analysis. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 977–982, 2019.

[5] Anmol Uppal, Vipul Sachdeva, and Seema Sharma. Fake news detection using discourse segment structure analysis. In *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 751–756. IEEE, 2020.

[6] Real time news. (2024, November 18). Kaggle.