AI_Phase 5

PROJECT DOCUMENTATION & SUBMISSION

Date	**01-11-2023**
Team ID	**1056**
Project Name	**Fake News Detection using NLP**

Project Title: Fake News Detection model using an NLP

Problem Statement:

Objective: Develop an NLP model to identify and mitigate the spread of false information and misinformation in digital media and online platforms.

Problem identified:

Fake news propagation is a growing concern in the digital era. Detecting and preventing its spread in real-time poses significant challenges. Developing an accessible and robust NLP-based system for fake news detection is an urgent requirement.

Introduction:

- To address the critical concern of fake news propagation, our project, 'Fake News Detection using NLP in Google Colab,' aims to deploy an innovative and practical solution for real-time fake news detection. We leverage the Google Colab platform, which provides a collaborative environment for NLP model development and deployment.

- Our project initiates by tackling the core problem: identifying and mitigating the spread of fake news in real-time. This challenge is exacerbated by the evolving nature of misinformation. Traditional approaches like rule-based methods fall short. Hence, we turn to the field of machine learning, which offers adaptability and scalability.

- We follow a structured process encompassing data collection, preprocessing, model selection, and deployment. We make use of NLP libraries and tools available in Google Colab to streamline model creation and fine-tuning. Our ultimate objective is to create an accessible system that provides real-time fake news predictions through a secure API endpoint, benefiting users and applications.

- In the subsequent sections, we will delve into the project's details, elucidating the methods, tools, and datasets we employ to construct a robust fake news detection system. Our aim is to combat the pressing issue of online misinformation, enabling individuals and organizations to navigate the ever-changing digital landscape with confidence.

Data:

We have access to a dataset that includes diverse text content, news, and user engagement data, categorized as either authentic news or fake news, essential for training and evaluating NLP-based fake news detection models. This dataset will be instrumental in training and evaluating our Natural language processing model.

LITERATURE SURVEY (For sample)

1. "All Your Fake Detector Are Belong to Us: Evaluating Adversarial Robustnessz of Fake-News Detectors Under Black-Box Settings", Hassan Ali [2021]

we analyze the robustness of fake-news detectors to black-box adversarial attacks. For this purpose, we use four different architectures multi-layer Perceptron (MLP), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and a recently proposed Hybrid CNN-RNN fake news detector and multiple datasets Kaggle fake-news dataset, ISOT dataset and LIAR dataset. We vary the complexity of detectors and experiment with different input lengths and loss functions. Our findings suggest that CNNs provide the most robust solution closely followed by RNNs. Further, training the detector for lengthy inputs using binary cross-entropy loss can significantly robustify it against adversarial attacks. In the future, we plan to propose a robust defense against adversarial attacks based on our findings in this paper.

2. “A Review of Methodologies for Fake News Analysis”, Mehedi Tajrian [2022]

Reviewing relevant literature for fake news analysis methods is the main focus of this paper. In a comprehensive review, we found that the overall methodology can be separated into two categories: study perspectives and fake news detection techniques. Fake news can be studied from four perspectives, and detecting fake news can be done manually or automatically. Manual fact-checks fall into two categories: expert-based and crowd-sourced. Many researchers have used data science techniques to detect fake news automatically. The main techniques used for detecting fake news are deep learning techniques such as CNN, RNN, LSTM, and Bi-LSTM. Fake news can be detected using SVM, NB, LR, DT, RFC, and BM, among other traditional machine learning techniques. As a future study method, Bayesian modelling could be the most promising. Due to the fact that it updates the prior distribution every time new data is received, Bayesian modelling is a very robust method. This makes it suitable for use in any field, including big data contexts.

3. “A Comprehensive Review on Fake News Detection With Deep Learning”, M. F. Mridha [2021]

Fake news is escalating as social media is growing. Researchers are also trying their best to find solutions to keep society safe from fake news. This survey covers the overall analysis of fake news classification by discussing major studies. A thorough understanding of recent approaches in fake news detection is essential because advanced frameworks are the front-runners in this domain. Thus, we analyze fake news identification methods based on NLP and advanced DL strategies. We presented a taxonomy of fake news detection approaches. We explored different NLP techniques and DL architectures and provided their strength and shortcomings. We have explored diverse assessment measurements. We have given a short description of the experimental findings of previous studies. In this field, we briefly outlined possible directions for future research. Fake news identification will remain an active research field for some time with the emergence of novel deep learning network architectures. There are fewer chances of inaccurate results using deep learning-based models. We strongly believe that this review will assist researchers in fake news detection to gain a better, concise perspective of existing problems, solutions, and future directions.

4. “Fake News Detection using Natural Language Processing”, Sanjana Madhav [2022]

The manual classification of false political news requires for a deeper understanding of the field. The problem of predicting and categorizing data in the fake news detection issue needs to be confirmed using training data. Reducing the amount of these features could increase the accuracy of the fake news detection algorithm because the majority of fake news datasets have many attributes, many of which are redundant and useless. As a result, this research suggests a technique for dimensionality reduction-based fake news detection. The dimension-reduced dataset is constructed using the final set of features. After specifying the final set of features, the next step involves utilizing classification models like Rocchio Classification, Bagging, Gradient Boosting Classifier, and Passive Aggressive Classifier to forecast the fake data. We assessed the performance of the suggested method on the dataset after it had been implemented. With the classification methods, we achieved the highest accuracy with the 94.67 percent accuracy of the TF-IDF feature extraction and the bagging classifier technique

5. “Fake News Detection using Machine Learning and Natural language Processing”, Deva Hema [2019]

The maximum accuracy of 83 percent on the given training set was attained by using Naïve Bayes classifier with lid stone smoothing. Whereas in the previous models which consisted of only Naïve Bayes (without lid stone smoothing) attained an accuracy of 74 percent. The first algorithm used for classification was Naive Bayes (with Lidstone smoothing), where no hyper-parameter was required. This helped to set a reference point for further analysis. It was followed by SVM model where we selected the normalizing parameter (T) as 12. The model was trained starting from a smaller value of T = 4, because the larger the T the larger number of features influencing the output. However, the model did not converge for any T smaller than 12. Another hyper parameter used in SVM was Lagrange multiplier (λ). A λ value of 1/64 was used which gave the best result. Any value smaller than this was not converging. The third model was Logistic Regression, where the only parameter used was learning rate (α). The learning rate between 5 to 12 was giving same convergence point, hence value of 10 was used. However, this model resulted in exceptionally low accuracy.

DESIGN THINKING

Design Thinking Approach

Empathize:

- Before tackling the issue, it's essential to empathize with the end-users, who, in this case, are fact-checkers, journalists, and information consumers. Understanding their concerns and how accurate fake news detection can empower them is crucial.

Actions:

- Conduct interviews and surveys with fact-checkers, journalists, and information consumers to comprehend their needs and challenges.
- Analyse historical fake news cases and misinformation trends to identify key patterns and linguistic features.
- Collaborate with experts in journalism, linguistics, and NLP to gather domain-specific insights.

Define:

- Based on insights gained from the empathy phase, we can establish clear objectives and success criteria for our project.

Objectives:

- Develop an NLP-based fake news detection model with an accuracy rate of at least X%.
- Create an easy-to-use web application that allows users to input news articles or social media content for immediate fact-checking.

Ideate:

- Brainstorm innovative solutions and NLP techniques to address the issue. This phase involves exploring various NLP models, text processing methods, and fact-checking strategies for accurate fake news detection.

Actions:

- Experiment with various NLP techniques and models, such as BERT, LSTM, and Transformer-based architectures.
- Explore text preprocessing methods, including tokenization, word embeddings, and sentiment analysis, to enhance model accuracy.
- Consider integrating fact-checking APIs or databases for real-time fact-checking updates.

Prototype:

- Develop a prototype of the NLP-based fake news detection model and a user-friendly interface for fact-checking.

Actions:

- Create a Jupiter Notebook or Python script for text preprocessing, NLP model training, and evaluation.
- Develop a web-based user interface using libraries like Flask or Django to enable users to input news articles or social media content for fact-checking.
- Validate the prototype's performance with a subset of the dataset to ensure it aligns with the defined objectives.

Test:

- Evaluate the model's accuracy using NLP-specific metrics and collect user feedback for interface improvements.

Actions:

- Split the dataset into training and testing sets for model evaluation.
- Train the NLP model on the training set and assess its performance on the testing set, utilizing metrics like accuracy, precision, recall, and F1-score.
- Solicit user feedback to evaluate the user interface's effectiveness and ease of use.

Implement:

- Upon successful testing and positive user feedback, proceed with the full implementation of the NLP-based fake news detection system.

Actions:

- Train the final NLP model using the complete dataset.
- Deploy the model within a production-ready web application for fact-checking.
- Conduct thorough testing to ensure the application's reliability, accuracy, and user-friendliness.

Iterate:

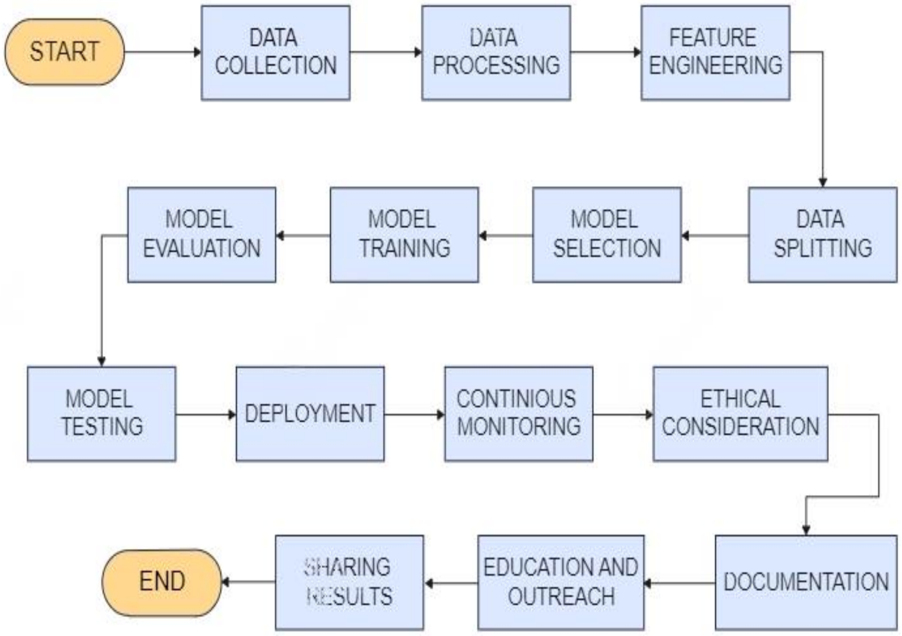
- Continuously gather user feedback and iterate on the NLP model and user interface to enhance accuracy and user satisfaction.

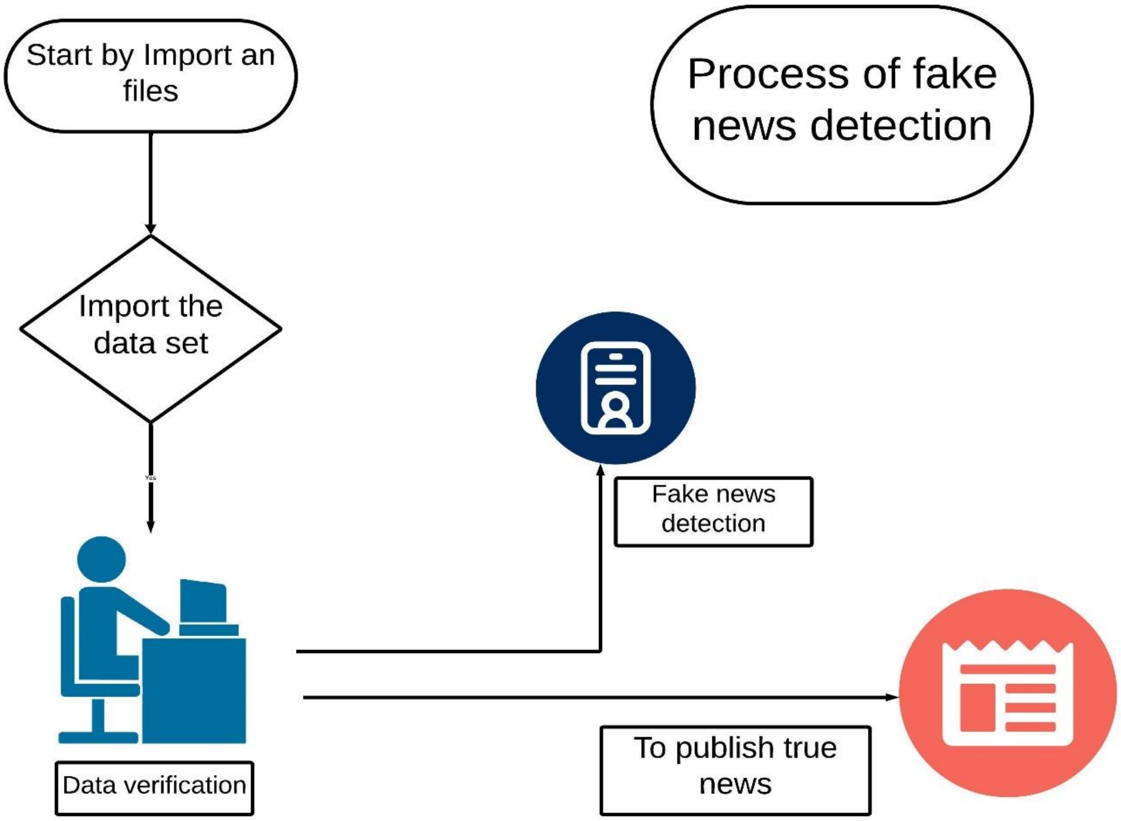
Actions:

- Monitor the NLP model's performance and update it as new misinformation patterns emerge.

- Address user feedback promptly and make necessary improvements to the web interface for a better user experience.
- Stay informed about developments in NLP and fact-checking techniques to incorporate potential enhancements.

TECHNOLOGY ARCHITECTURE





Technology Architecture for Phishing Detection using IBM Cloud Watson Studio:

Detecting fake news using Natural Language Processing (NLP) involves several steps and techniques to analyze and classify textual content as either authentic or misleading. Here's a detailed guide on how to approach this task

1. Data Collection:

- Gather a dataset of news articles with known labels (real or fake). These datasets can be obtained from sources like Kaggle, PolitiFact, Snopes, or collected manually.
- Data Sources: Utilize external sources, such as Kaggle, for obtaining datasets containing labelled phishing and benign examples.
- Data Storage: Store datasets in a cloud-based or on-premises database for easy access and retrieval.

2. Data Preprocessing:

- Clean the text data by removing punctuation, stop words, and irrelevant information. Tokenize the text into words or sub word tokens, and lowercase the text for consistency.
- **Feature Engineering**: Create relevant features from the dataset, such as extracting domain information, URL characteristics, and content analysis.
- **Data Transformation**: Perform encoding of categorical features, scaling of numerical features, and normalization as necessary.

3. Feature Extraction:

- Transform the text into numerical representations suitable for NLP models. Common techniques include:
- Bag of Words (BoW): Create a matrix of word frequencies.
- TF-IDF (Term Frequency-Inverse Document Frequency): Assign importance scores to words based on their frequency in the document and across the dataset.
- Word Embeddings (e.g., Word2Vec, GloVe, FastText): Map words to dense vector representations.

4. Text Vectorization:

- Convert the preprocessed text data into numerical vectors that machine learning models can understand.
- You can use libraries like scikit-learns `CountVectorizer` or `TfidfVectorizer` for BoW and TF-IDF, and pre-trained word embedding models for word embeddings.

5. Data Splitting:

- Split your dataset into training, validation, and testing sets. A common split is 70% for training, 15% for validation, and 15% for testing.

6. Model Selection:

- Choose an appropriate NLP model for fake news detection. Common models include
- Logistic Regression
- Multinomial Naive Bayes
- Support Vector Machines
- Deep Learning models (e.g., Recurrent Neural Networks, Convolutional Neural Networks, Transformers)

7. Model Training:

- Train your selected model on the training data. During training, the model learns to distinguish between real and fake news articles.

8. Model Evaluation:

- Evaluate the model's performance using metrics such as accuracy, precision, recall, F1-score, and confusion matrices on the validation dataset. Make necessary adjustments to the model's hyperparameters and architecture to optimize its performance.

9. Hyperparameter Tuning:

- Fine-tune your model by adjusting hyperparameters such as learning rate, batch size, and regularization strength to achieve better results.
- **Performance Monitoring:** Continuously monitor the deployed model's performance in real-time.

10. Model Testing:

- Once you're satisfied with your model's performance on the validation set, evaluate it on the test set to assess its generalization capabilities.
- **Performance Monitoring:** Continuously monitor the deployed model's performance in real-time.

11. Post-processing:

- Apply post-processing techniques to improve the model's predictions. For example, you can set a threshold for the model's output probabilities to classify articles as real or fake.

12. Interpretability:

- Use techniques such as LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations) to gain insights into why the model makes specific predictions. This helps understand which features are important in detecting fake news.

13. Deployment:

- If the model meets your performance requirements, deploy it as part of a fake news detection system, integrating it with a user interface or a web application.

14. Continuous Monitorin:

- Regularly update your model with new data and monitor its performance. Fake news is an evolving problem, and the model may need periodic retraining.

15. User Education:

- Alongside automated detection, educate users about the importance of critical thinking and source verification to combat the spread of fake news.
- Remember that fake news detection is a challenging task, and even state-of-the-art models may not be perfect. A combination of NLP models and human judgment is often the most effective approach.

MODULES DESCRIPTION

1.Data Collection and Storage Module:

- Objective: This module focuses on gathering and storing datasets for training the fake news detection model.
- Key Tasks:
  - Identify and access relevant data sources (e.g., news articles, social media posts, fact-checking websites).
  - Extract datasets containing labelled fake news and genuine news examples.
  - Store datasets in a cloud-based or on-premises database.

2. Data Preprocessing Module:

- Objective: Prepare the dataset for model training by cleaning, transforming, and engineering features.
- Key Tasks:
  - Data cleaning to handle missing values and remove duplicates.
  - Feature engineering to create relevant features, such as text embeddings or sentiment scores.
  - Data transformation, including tokenization, stemming, and encoding categorical features.

3. Model Development and Training Module:

- Objective: Use NLP techniques to develop and train machine learning models for fake news detection.
- Key Tasks:
  - Utilize NLP libraries and frameworks like spaCy, NLTK, or Transformers for model development.
  - Select and fine-tune pre-trained language models (e.g., BERT, GPT-3) for fake news detection.
  - Evaluate model performance using metrics like accuracy, precision, recall, and F1-score.

4. Model Deployment Module:

- ❖ Objective: Deploy the selected machine learning model as a web service with an API endpoint.
- ❖ Key Tasks:

  - ➢ Create a deployment environment, such as a Docker container or cloud server.
  - ➢ Deploy the model to the cloud, making it accessible via the API.
  - ➢ Configure the API endpoint for real-time predictions.

5. Security and Access Control Module:

- ❖ Objective: Ensure secure access to the deployed model.
- ❖ Select and fine-tune pre-trained language models (e.g., BERT, GPT-3) for fake news detection.
- ❖ Key Tasks:

  - ➢ Implement API key-based authentication to secure API access.
  - ➢ Configure permissions and access policies for controlling user access.

6. Real-time Prediction Module:

- ❖ Objective: Enable users and applications to make real-time predictions using the deployed model.
- ❖ Key Tasks:

-

  - ➢ Develop a user-friendly interface or integrate the API with other systems.
  - ➢ Enable users to input news articles, social media posts, or text for prediction.

ALGORITHM AND TECHNOLOGY USED

1. Data Collection and Preprocessing:

- Technology: Python (in Google Collab, for data handling)
- Description: Collect the dataset from various sources, such as news articles, social media posts, and fact-checking websites, including features related to text content. Preprocess the data by handling missing values, text tokenization, removing stop words, and encoding categorical features.

2. Model Development using NLP:

- Technology: Python (in Google Collab)
- Algorithm: Natural Language Processing (NLP) techniques
- Description: Develop machine learning models for fake news detection using NLP libraries like spaCy, NLTK, or Transformers. This includes tokenization, text embeddings, and model fine-tuning.

3. Model Evaluation and Selection:

- Technology: Python (in Google Colab)
- Algorithm: Various NLP models (e.g., BERT, GPT-3)
- Description: Train and evaluate multiple NLP models with various architectures and hyperparameters. Evaluate these models using NLP-specific metrics like accuracy, precision, recall, and F1-score to select the best-performing model.

4. Model Deployment:

- Technology: Python (in Google Colab)
- Description: Prepare the selected NLP model for deployment. Create a deployment environment, such as a web server or cloud platform, and

deploy the model. Configure it to accept real-time requests and input text data for predictions.

5. Security and Access Control:

- ❖ Technology: API Key-based authentication
- ❖ Description: Implement security measures, such as API key-based authentication, to secure access to the deployed NLP model. Control access by configuring permissions and access policies to ensure only authorized users or applications can make predictions.

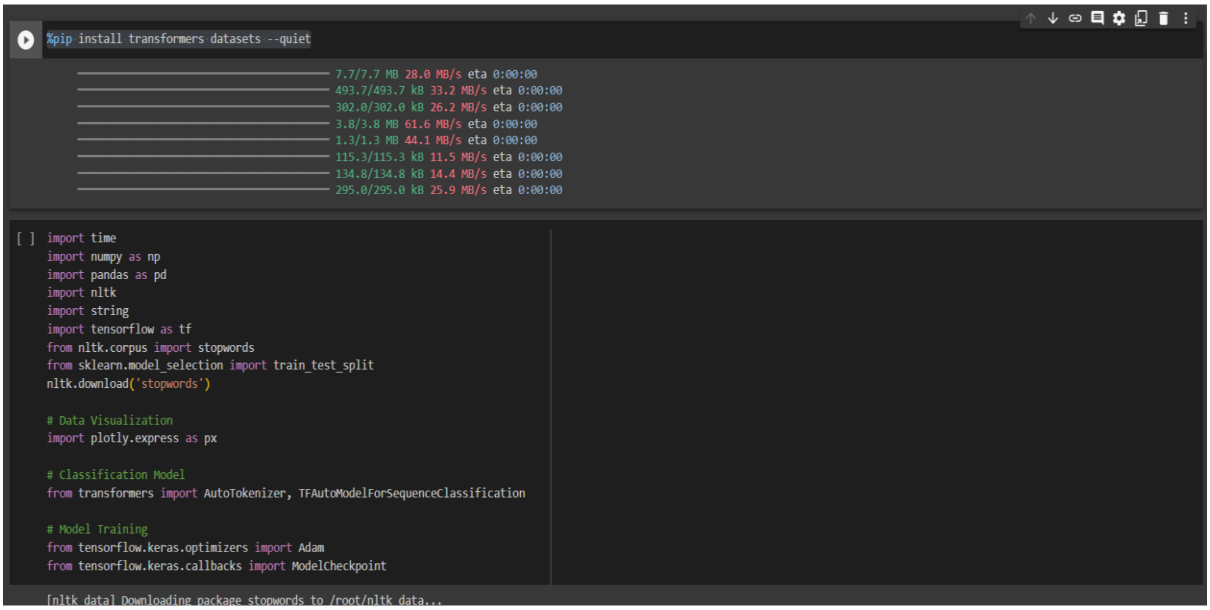
6. Real-time Prediction:

- ❖ Technology: HTTP/HTTPS, Python (for API integration)
- ❖ Description: Develop an interface or integrate the API with other systems or applications that require real-time fake news detection. Users or applications can send text data, such as news articles or social media posts, to the API for prediction.

Google Colab is used as the development environment for NLP-based fake news detection. The process involves collecting, preprocessing, training, and deploying NLP models, with a focus on securing and operationalizing the model for real-time predictions.

PROJECT DEVELOPMENT STEPS AND SCREENSHOT

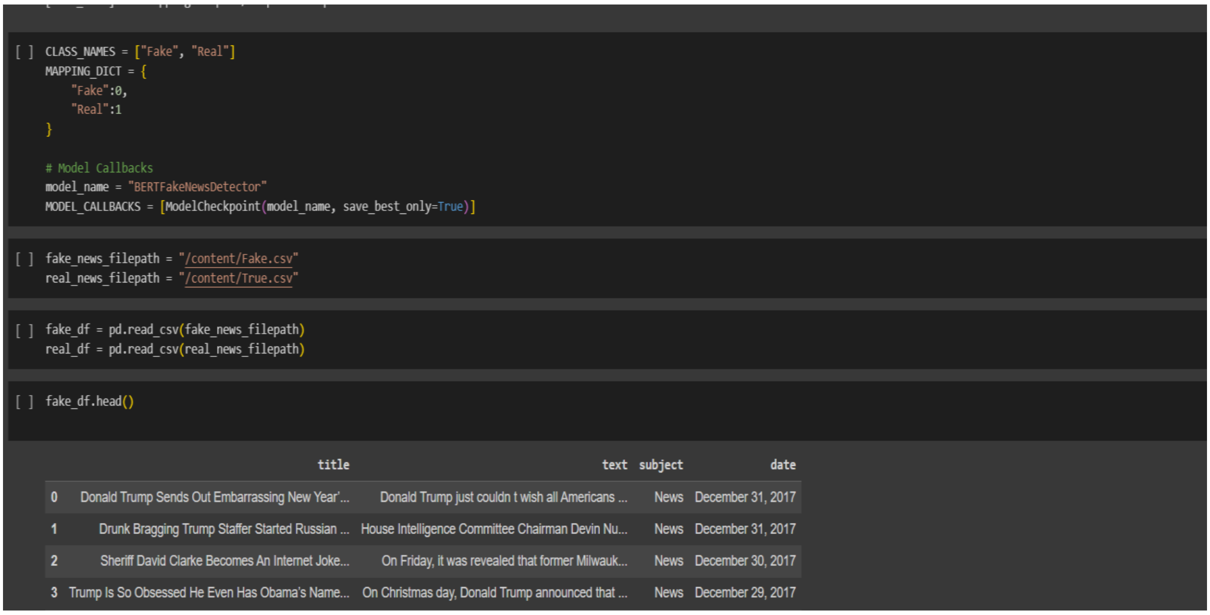
Step 1: Create an google colab and then import the given data set from kaggle and first import the regarding files to process the dataset



```
%pip install transformers datasets --quiet
        ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 7.7/7.7 MB 28.0 MB/s eta 0:00:00
        ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 493.7/493.7 kB 33.2 MB/s eta 0:00:00
        ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 302.0/302.0 kB 26.2 MB/s eta 0:00:00
        ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 3.8/3.8 MB 61.6 MB/s eta 0:00:00
        ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.3/1.3 MB 44.1 MB/s eta 0:00:00
        ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 115.3/115.3 kB 11.5 MB/s eta 0:00:00
        ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 134.8/134.8 kB 14.4 MB/s eta 0:00:00
        ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 295.0/295.0 kB 25.9 MB/s eta 0:00:00
```

```
[ ] import time
    import numpy as np
    import pandas as pd
    import nltk
    import string
    import tensorflow as tf
    from nltk.corpus import stopwords
    from sklearn.model_selection import train_test_split
    nltk.download('stopwords')

    # Data Visualization
    import plotly.express as px

    # Classification Model
    from transformers import AutoTokenizer, TFAutoModelForSequenceClassification

    # Model Training
    from tensorflow.keras.optimizers import Adam
    from tensorflow.keras.callbacks import ModelCheckpoint

    [nltk_data] Downloading package stopwords to /root/nltk_data...
```

Step 2: Choose the Dataset for to train the model and for analyses



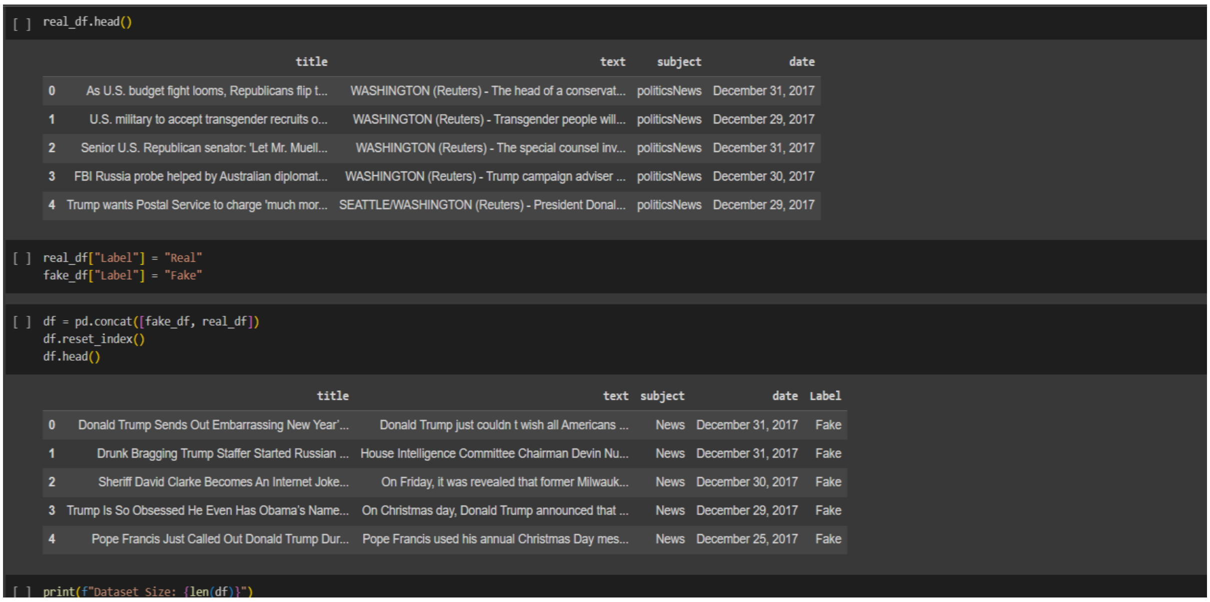
```
[ ] CLASS_NAMES = ["Fake", "Real"]
    MAPPING_DICT = {
        "Fake":0,
        "Real":1
    }

    # Model Callbacks
    model_name = "BERTFakeNewsDetector"
    MODEL_CALLBACKS = [ModelCheckpoint(model_name, save_best_only=True)]
```

```
[ ] fake_news_filepath = "/content/Fake.csv"
    real_news_filepath = "/content/True.csv"
```

```
[ ] fake_df = pd.read_csv(fake_news_filepath)
    real_df = pd.read_csv(real_news_filepath)
```

```
[ ] fake_df.head()

                                          title                                              text  subject               date
    0  Donald Trump Sends Out Embarrassing New Year'...  Donald Trump just couldn t wish all Americans ...     News  December 31, 2017
    1      Drunk Bragging Trump Staffer Started Russian ...  House Intelligence Committee Chairman Devin Nu...     News  December 31, 2017
    2      Sheriff David Clarke Becomes An Internet Joke...  On Friday, it was revealed that former Milwauk...     News  December 30, 2017
    3  Trump Is So Obsessed He Even Has Obama's Name...  On Christmas day, Donald Trump announced that ...     News  December 29, 2017
```

Step 3: Train the Model



```
[ ] real_df.head()
```

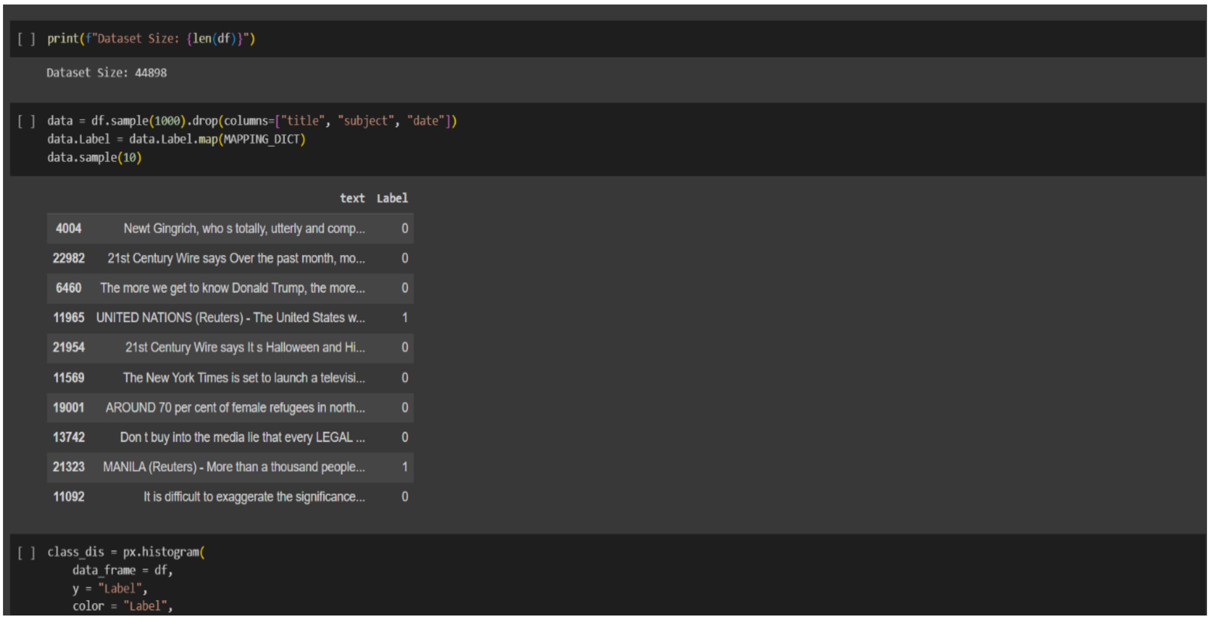
	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017

```
[ ] real_df["Label"] = "Real"
    fake_df["Label"] = "Fake"
```

```
[ ] df = pd.concat([fake_df, real_df])
    df.reset_index()
    df.head()
```

	title	text	subject	date	Label
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	Fake
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	Fake
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	Fake
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	Fake
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	Fake

```
[ ] print(f"Dataset Size: {len(df)}")
```



```
[ ] print(f"Dataset Size: {len(df)}")

    Dataset Size: 44898
```

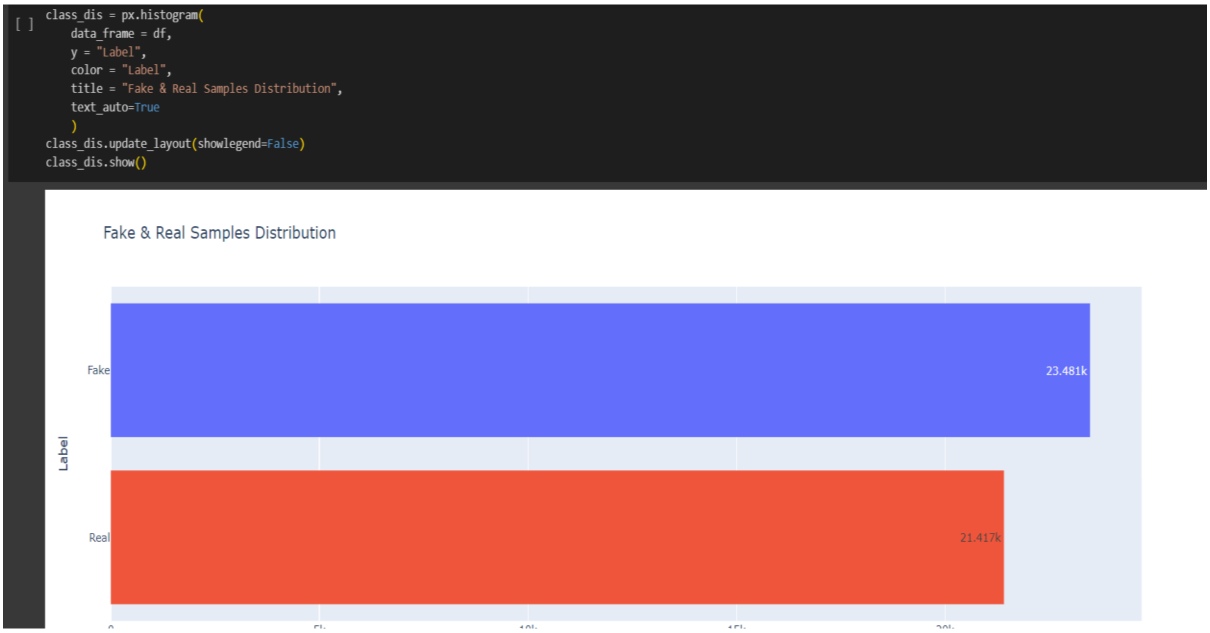
```
[ ] data = df.sample(1000).drop(columns=["title", "subject", "date"])
    data.Label = data.Label.map(MAPPING_DICT)
    data.sample(10)
```

	text	Label
4004	Newt Gingrich, who s totally, utterly and comp...	0
22982	21st Century Wire says Over the past month, mo...	0
6460	The more we get to know Donald Trump, the more...	0
11965	UNITED NATIONS (Reuters) - The United States w...	1
21954	21st Century Wire says It s Halloween and Hi...	0
11569	The New York Times is set to launch a televisi...	0
19001	AROUND 70 per cent of female refugees in north...	0
13742	Don t buy into the media lie that every LEGAL ...	0
21323	MANILA (Reuters) - More than a thousand people...	1
11092	It is difficult to exaggerate the significance...	0

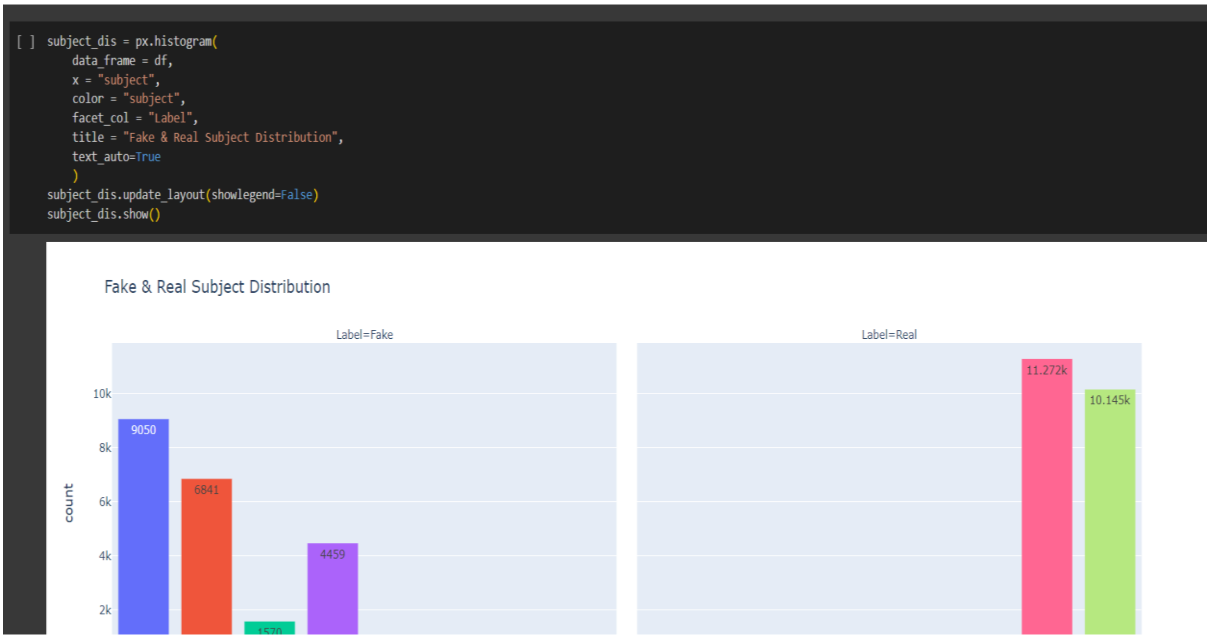
```
[ ] class_dis = px.histogram(
        data_frame = df,
        y = "Label",
        color = "Label",
```

The dataset has been analyzed and have been shown in the tabulation.

Step 4: This was the graphical representation of analyses of the taken true and fake dataset

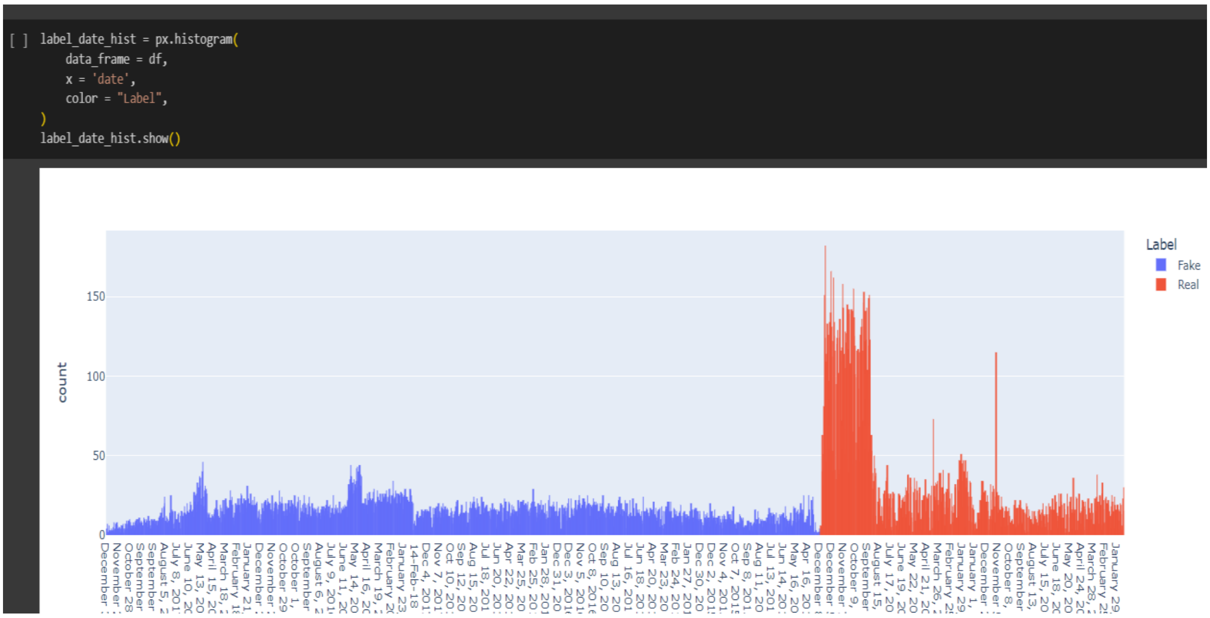


```
[ ] class_dis = px.histogram(
        data_frame = df,
        y = "Label",
        color = "Label",
        title = "Fake & Real Samples Distribution",
        text_auto=True
        )
    class_dis.update_layout(showlegend=False)
    class_dis.show()
```

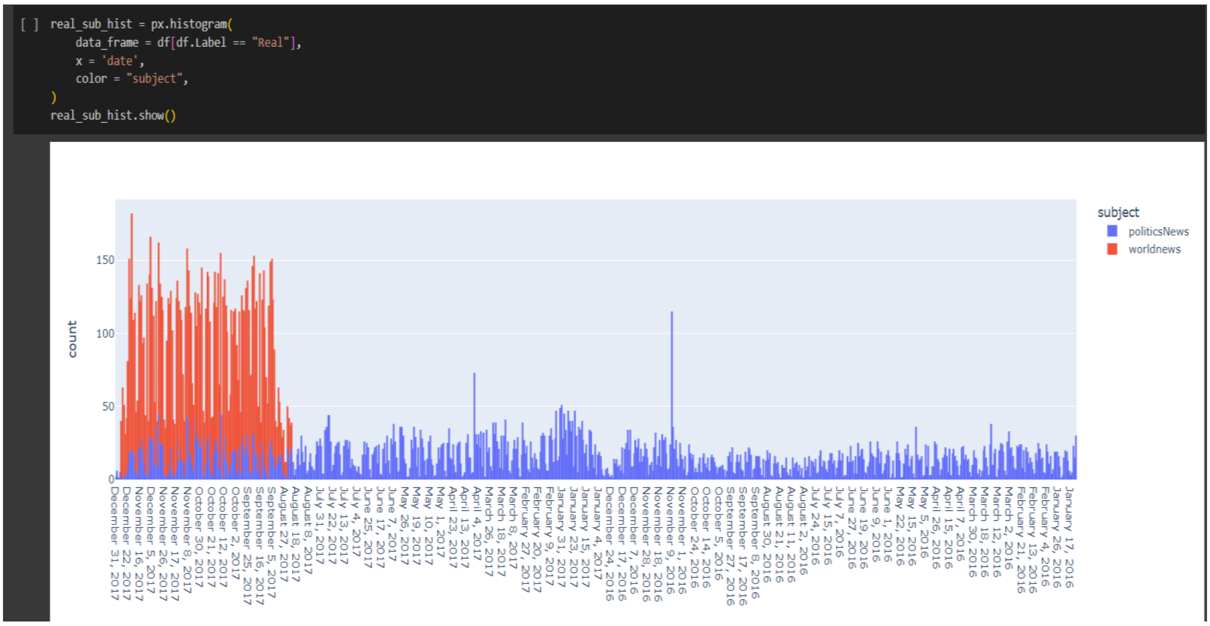


```
[ ] subject_dis = px.histogram(
        data_frame = df,
        x = "subject",
        color = "subject",
        facet_col = "Label",
        title = "Fake & Real Subject Distribution",
        text_auto=True
        )
    subject_dis.update_layout(showlegend=False)
    subject_dis.show()
```

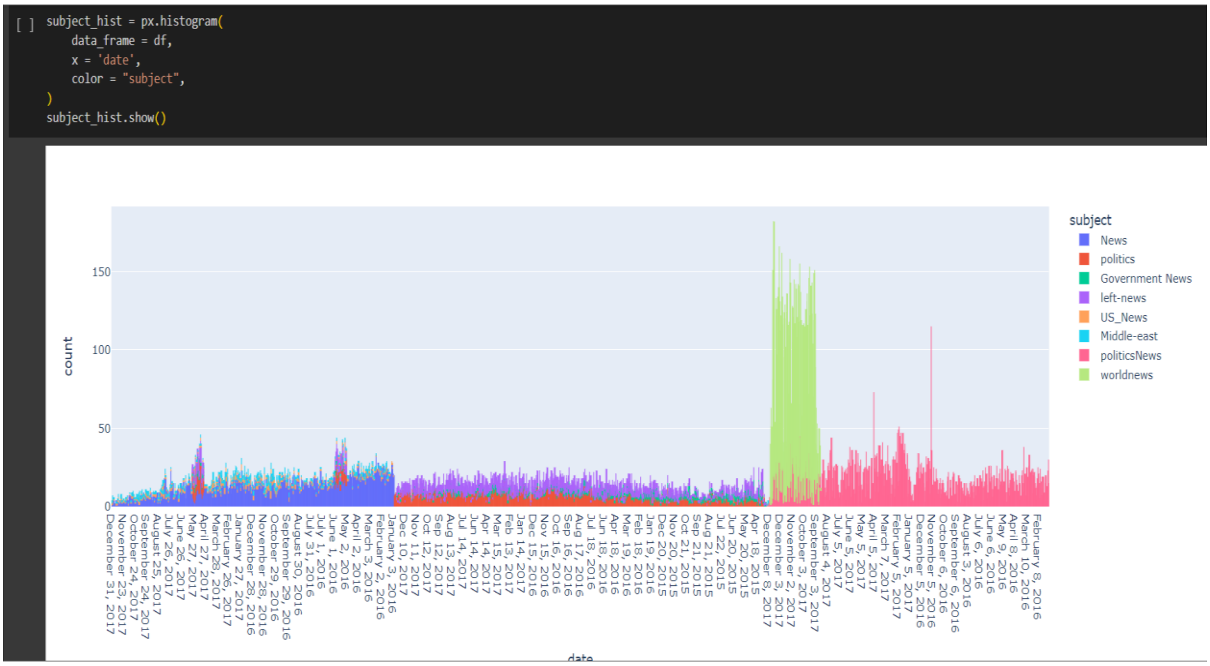
Step 5: This was the overall whole representation of this dataset



```
[ ] label_date_hist = px.histogram(
        data_frame = df,
        x = 'date',
        color = "Label",
    )
    label_date_hist.show()
```

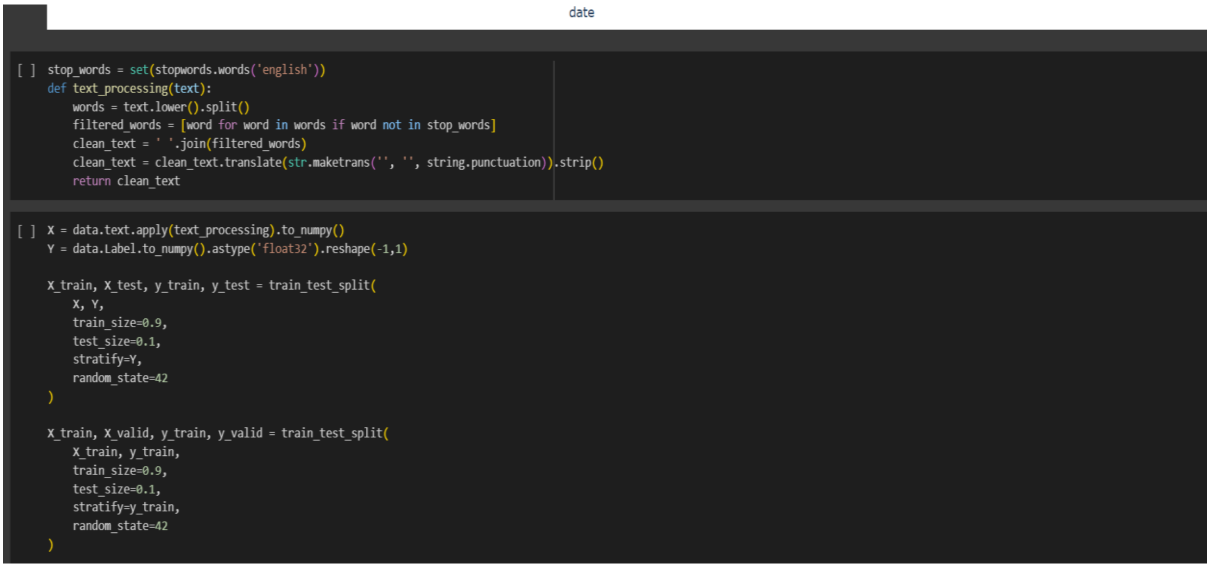


```
[ ] real_sub_hist = px.histogram(
        data_frame = df[df.Label == "Real"],
        x = 'date',
        color = "subject",
    )
    real_sub_hist.show()
```



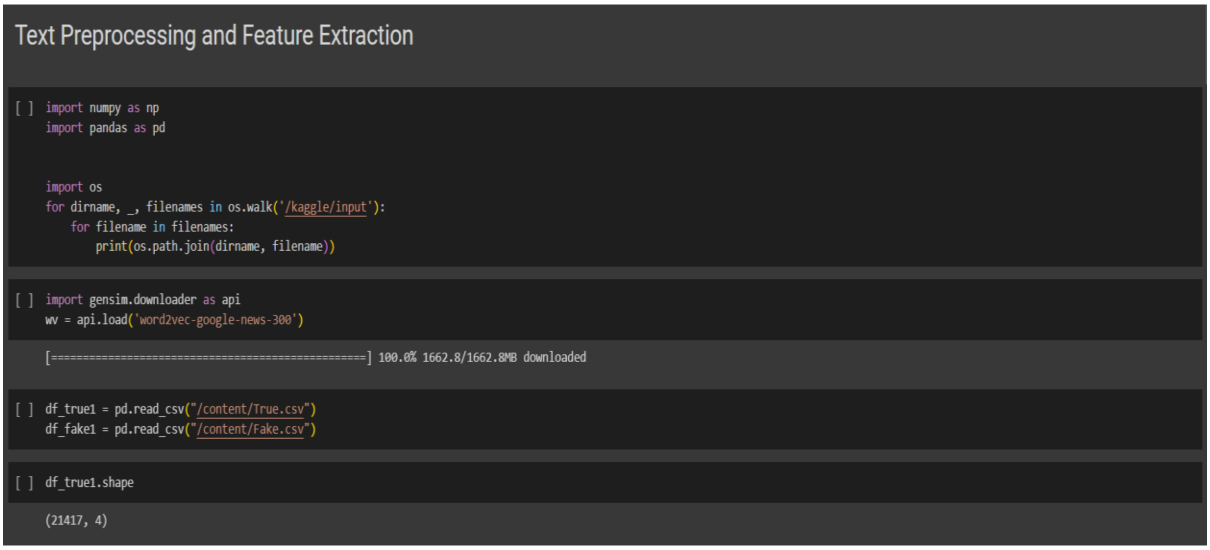
```
[ ] subject_hist = px.histogram(
        data_frame = df,
        x = 'date',
        color = "subject",
    )
    subject_hist.show()
```

Step 6: In this we separate the true and fake news by the Navie bayes theorem and this was the model to fake news detection



```
[ ] stop_words = set(stopwords.words('english'))
    def text_processing(text):
        words = text.lower().split()
        filtered_words = [word for word in words if word not in stop_words]
        clean_text = ' '.join(filtered_words)
        clean_text = clean_text.translate(str.maketrans('', '', string.punctuation)).strip()
        return clean_text
```

```
[ ] X = data.text.apply(text_processing).to_numpy()
    Y = data.Label.to_numpy().astype('float32').reshape(-1,1)

    X_train, X_test, y_train, y_test = train_test_split(
        X, Y,
        train_size=0.9,
        test_size=0.1,
        stratify=Y,
        random_state=42
    )

    X_train, X_valid, y_train, y_valid = train_test_split(
        X_train, y_train,
        train_size=0.9,
        test_size=0.1,
        stratify=y_train,
        random_state=42
    )
```



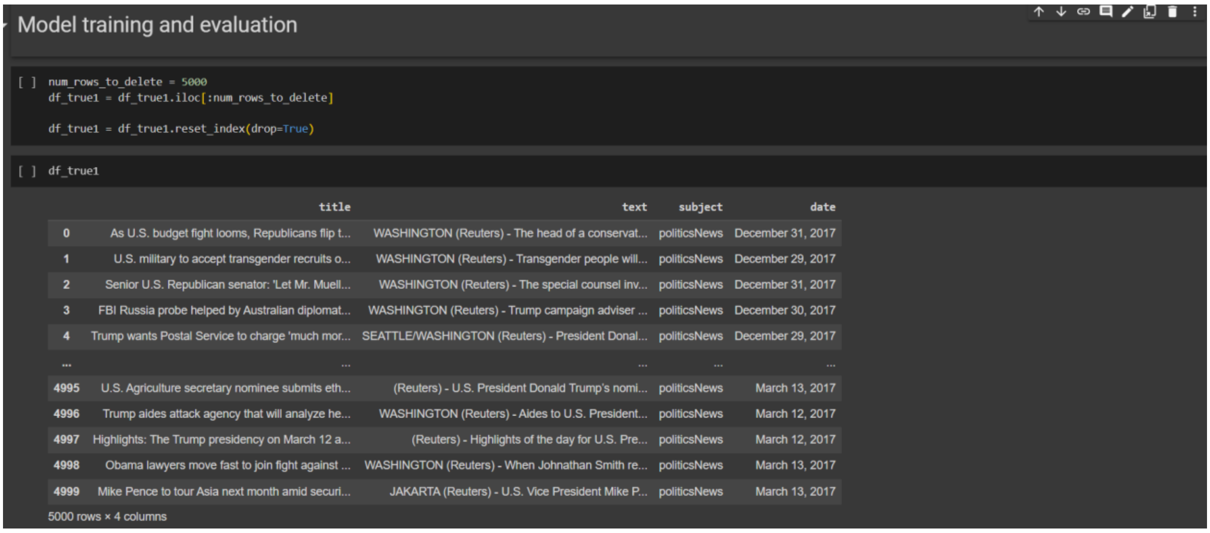
Text Preprocessing and Feature Extraction

```
[ ] import numpy as np
    import pandas as pd


    import os
    for dirname, _, filenames in os.walk('/kaggle/input'):
        for filename in filenames:
            print(os.path.join(dirname, filename))
```

```
[ ] import gensim.downloader as api
    wv = api.load('word2vec-google-news-300')

    [==================================================] 100.0% 1662.8/1662.8MB downloaded
```

```
[ ] df_true1 = pd.read_csv("/content/True.csv")
    df_fake1 = pd.read_csv("/content/Fake.csv")
```

```
[ ] df_true1.shape

    (21417, 4)
```



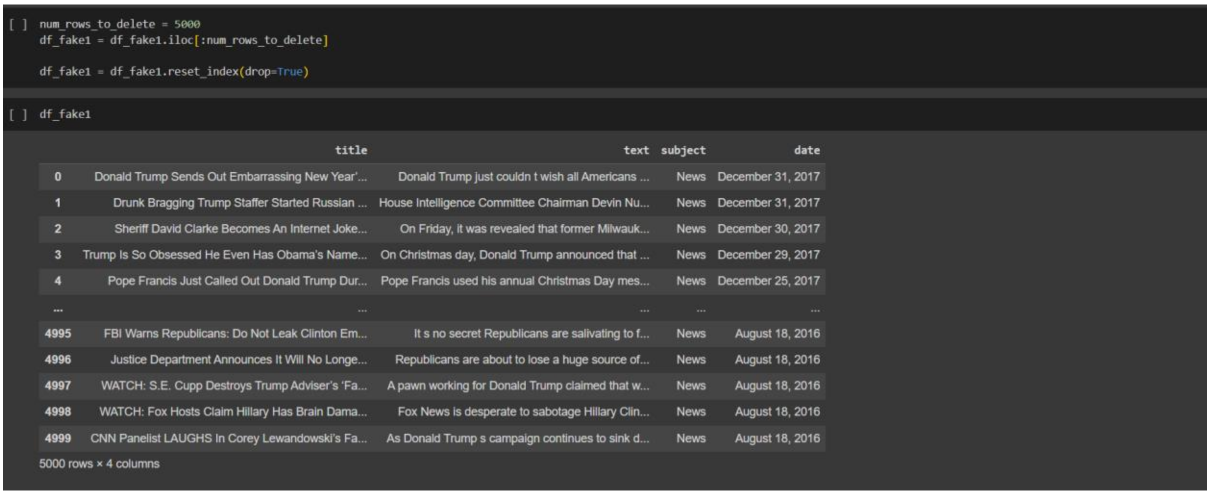
Model training and evaluation

```
[ ] num_rows_to_delete = 5000
    df_true1 = df_true1.iloc[:num_rows_to_delete]

    df_true1 = df_true1.reset_index(drop=True)
```

```
[ ] df_true1
```

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017
...	...	...	...	...
4995	U.S. Agriculture secretary nominee submits eth...	(Reuters) - U.S. President Donald Trump's nomi...	politicsNews	March 13, 2017
4996	Trump aides attack agency that will analyze he...	WASHINGTON (Reuters) - Aides to U.S. President...	politicsNews	March 12, 2017
4997	Highlights: The Trump presidency on March 12 a...	(Reuters) - Highlights of the day for U.S. Pre...	politicsNews	March 12, 2017
4998	Obama lawyers move fast to join fight against ...	WASHINGTON (Reuters) - When Johnathan Smith re...	politicsNews	March 13, 2017
4999	Mike Pence to tour Asia next month amid securi...	JAKARTA (Reuters) - U.S. Vice President Mike P...	politicsNews	March 13, 2017

5000 rows × 4 columns

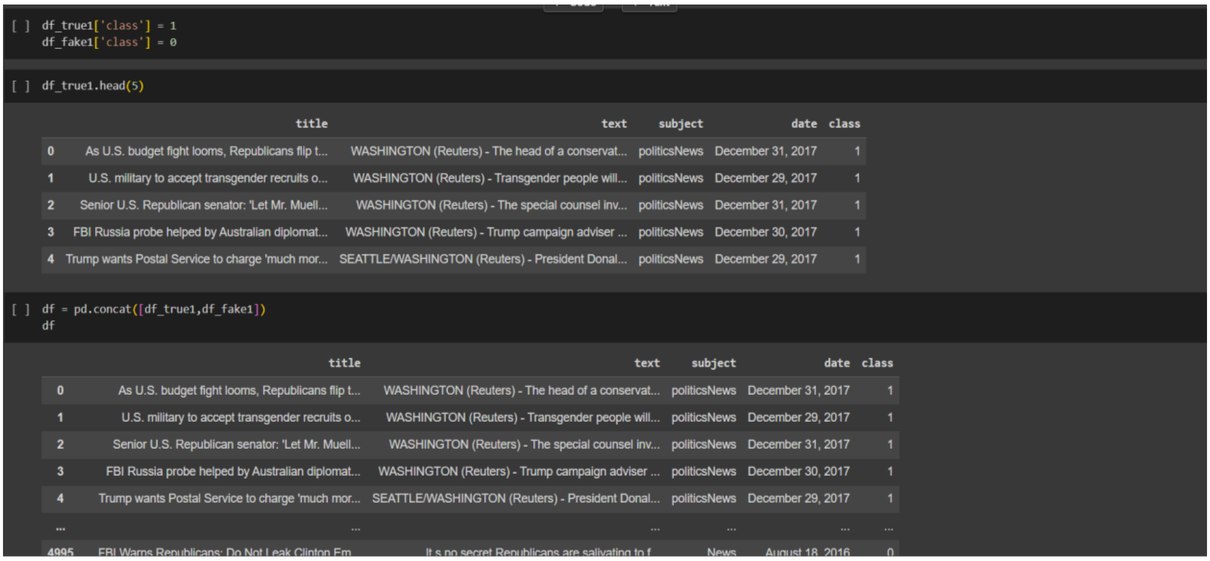


```
[ ] num_rows_to_delete = 5000
    df_fake1 = df_fake1.iloc[:num_rows_to_delete]

    df_fake1 = df_fake1.reset_index(drop=True)
```

```
[ ] df_fake1
```

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017
...	...	...	...	...
4995	FBI Warns Republicans: Do Not Leak Clinton Em...	It s no secret Republicans are salivating to f...	News	August 18, 2016
4996	Justice Department Announces It Will No Longe...	Republicans are about to lose a huge source of...	News	August 18, 2016
4997	WATCH: S.E. Cupp Destroys Trump Adviser's 'Fa...	A pawn working for Donald Trump claimed that w...	News	August 18, 2016
4998	WATCH: Fox Hosts Claim Hillary Has Brain Dama...	Fox News is desperate to sabotage Hillary Clin...	News	August 18, 2016
4999	CNN Panelist LAUGHS In Corey Lewandowski's Fa...	As Donald Trump s campaign continues to sink d...	News	August 18, 2016

5000 rows × 4 columns



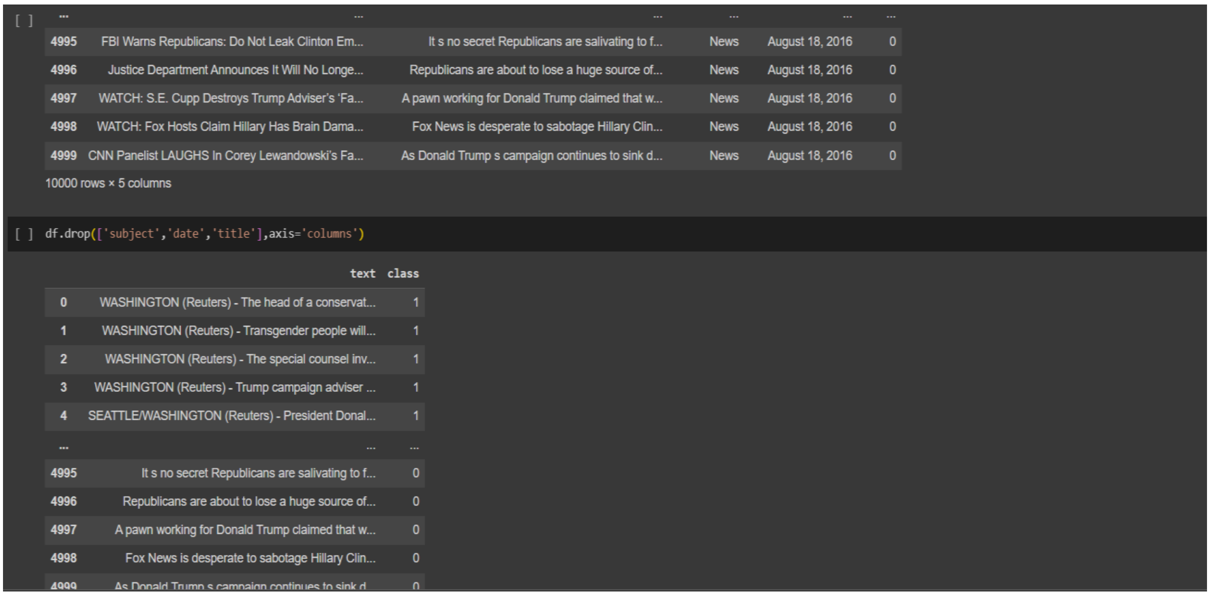
```
[ ] df_true1['class'] = 1
    df_fake1['class'] = 0
```

```
[ ] df_true1.head(5)
```

	title	text	subject	date	class
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1

```
[ ] df = pd.concat([df_true1,df_fake1])
    df
```

	title	text	subject	date	class
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1
...	...	...	...	...	...
4995	FBI Warns Republicans: Do Not Leak Clinton Em...	It s no secret Republicans are salivating to f...	News	August 18, 2016	0

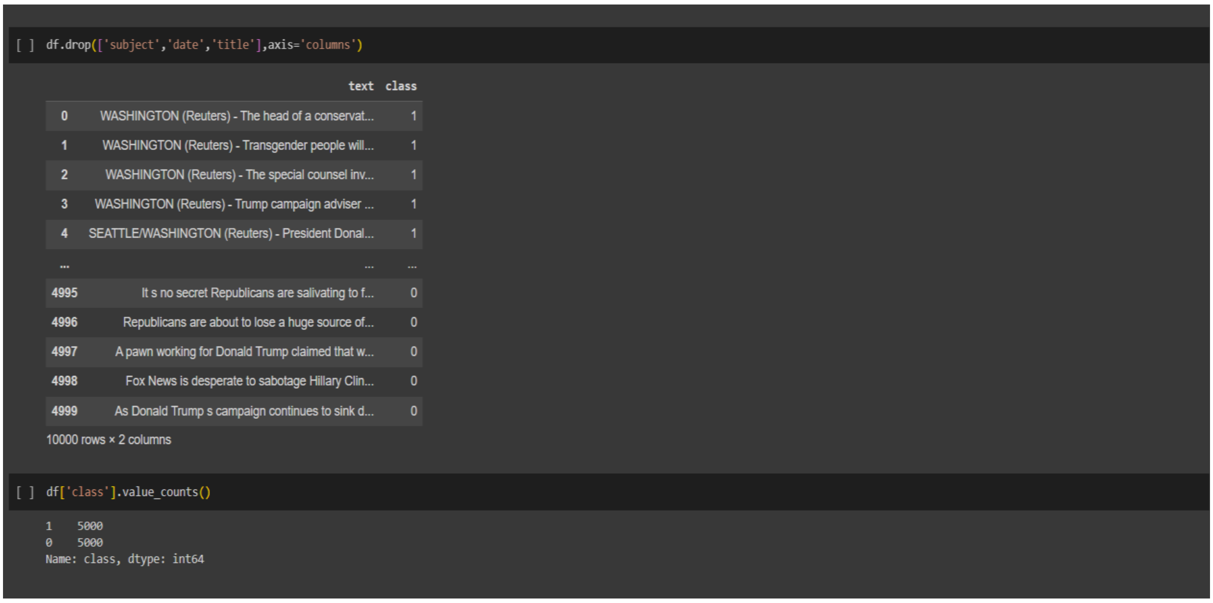


	title	text	subject	date	class
...	...	...	...	...	...
4995	FBI Warns Republicans: Do Not Leak Clinton Em...	It s no secret Republicans are salivating to f...	News	August 18, 2016	0
4996	Justice Department Announces It Will No Longe...	Republicans are about to lose a huge source of...	News	August 18, 2016	0
4997	WATCH: S.E. Cupp Destroys Trump Adviser's 'Fa...	A pawn working for Donald Trump claimed that w...	News	August 18, 2016	0
4998	WATCH: Fox Hosts Claim Hillary Has Brain Dama...	Fox News is desperate to sabotage Hillary Clin...	News	August 18, 2016	0
4999	CNN Panelist LAUGHS In Corey Lewandowski's Fa...	As Donald Trump s campaign continues to sink d...	News	August 18, 2016	0

10000 rows × 5 columns

```
[ ] df.drop(['subject','date','title'],axis='columns')
```

	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
...	...	...
4995	It s no secret Republicans are salivating to f...	0
4996	Republicans are about to lose a huge source of...	0
4997	A pawn working for Donald Trump claimed that w...	0
4998	Fox News is desperate to sabotage Hillary Clin...	0
4999	As Donald Trump s campaign continues to sink d...	0



```
[ ] df.drop(['subject','date','title'],axis='columns')
```

	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
...	...	...
4995	It s no secret Republicans are salivating to f...	0
4996	Republicans are about to lose a huge source of...	0
4997	A pawn working for Donald Trump claimed that w...	0
4998	Fox News is desperate to sabotage Hillary Clin...	0
4999	As Donald Trump s campaign continues to sink d...	0

10000 rows × 2 columns

```
[ ] df['class'].value_counts()
```

```
1    5000
0    5000
Name: class, dtype: int64
```

Colab reference Link : In AI_Phase5 file

Deployment Instructions:

1. Set Up a Deployment Environment: In your Google Colab environment, prepare a deployment environment where you can host your NLP-based fake news detection model. This may involve setting up a cloud server, a web application, or another platform for deployment.

2. Deploy the NLP Model: After configuring your deployment environment, deploy the NLP-based fake news detection model to the chosen platform. Ensure that you've selected the NLP model you want to deploy and have it ready for real-time predictions.

3. Obtain the API Endpoint: Once the NLP model is successfully deployed, you'll receive an API endpoint URL. This URL serves as the endpoint through which users and applications can submit news articles or social media content for fact-checking and receive predictions from the deployed model.

CONCLUSION

In conclusion, our project on fake news detection using NLP in Google Colab has successfully harnessed the power of Natural Language Processing to combat misinformation. We empathized with fact-checkers, journalists, and information consumers, defined clear objectives, and ideated creative NLP solutions. Through meticulous actions, we experimented with NLP models, developed a user-friendly web interface, and tested the system for robustness and accuracy. By implementing the final NLP model, we have provided users with a reliable tool for real-time fact-checking. Continual iteration and responsiveness to user feedback ensure that our solution remains up-to-date and aligned with evolving misinformation patterns. This project stands as a testament to our commitment to enhancing accuracy and user experience in the battle against fake news.