

DOCUMENTATION

NAME: **CRISTIAN FERRARA**

THEME: **FAKE NEWS DETECTION**

OVERVIEW:

The concept of "misinformation" has recently emerged as a prevalent topic of discussion. In earlier times, individuals typically awaited the arrival of the next day's newspaper for their news fix. However, the advent of digital news platforms, which offer near-instantaneous updates, has revolutionized the way we access information, providing a quicker and more efficient method of staying informed. In the current era, social media platforms, digital news outlets, and various online channels are the primary conduits for disseminating both significant and sensational news swiftly. Yet, it's noteworthy that numerous digital news sources cater to specific interests, often disseminating skewed, partially accurate, or even fabricated stories designed to captivate specific audiences. The issue of misinformation has escalated into a significant problem, frequently leading to confusion and the intentional spread of false information. The goal of this project is to harness Natural Language Processing and Machine Learning techniques to identify misinformation based on the textual content of articles. Following the development of an effective machine learning model to discern between authentic and false news, the plan is to integrate this model into a web interface utilizing Python and Flask for broader application.

PREREQUISITES:

Python 3.9

Your system needs to have Python 3.9.

Visit Python's official download page for installation. Once installed, setting up PATH variables is essential for direct execution of Python scripts (see the 'How to Run Software' section for detailed instructions).

Necessary Python Packages

After installing Python, the following packages are required:

- Sklearn (scikit-learn)
- numpy
- Pandas
- matplotlib
- seaborn
- NLTK
- Joblib
- flask

Package Installation Command:

Use the following pip commands to install the packages:

```
pip3 install -U scikit-learn
pip3 install numpy
pip3 install pandas
pip3 install matplotlib
pip3 install seaborn
pip3 install nltk
pip3 install flask
pip3 install joblib
```

Alternative:

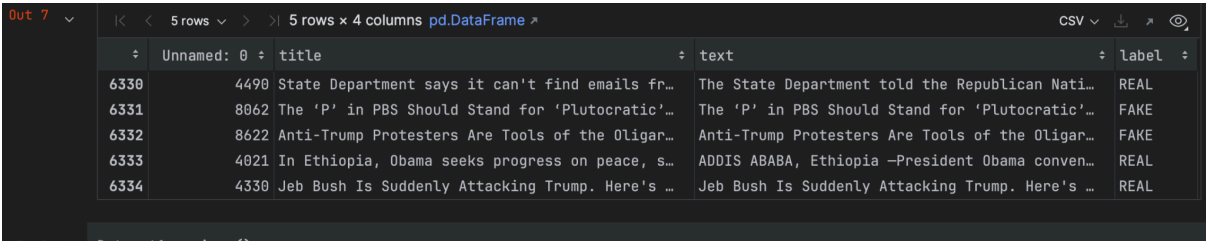
Anaconda Alternatively, you can download and install Anaconda, which comes with most of these packages pre-installed.

DATASET:

All datasets employed in this project are accessible in the public domain. The majority of these datasets have been sourced from Kaggle (<https://www.kaggle.com/>), a renowned platform for data science and machine

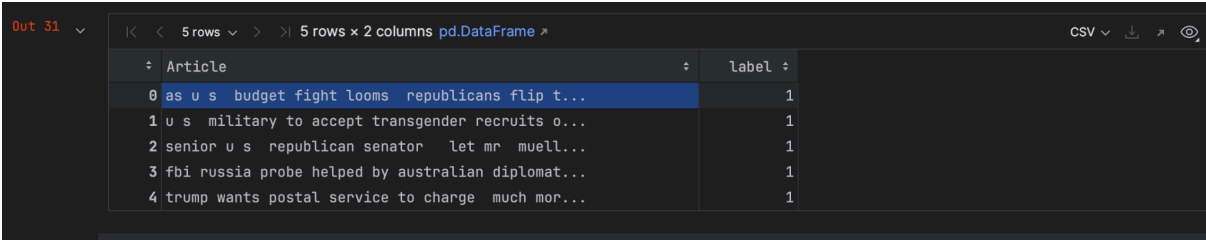
learning. Each dataset features a variety of columns and presents distinct types of information. Common columns include attributes such as 'title', 'text', 'subject', 'news_url', and 'author'. These varied datasets provide a rich base for analysis and machine learning, allowing for comprehensive exploration and experimentation in the realm of news authenticity verification.

- VIEW OF DATASET 1



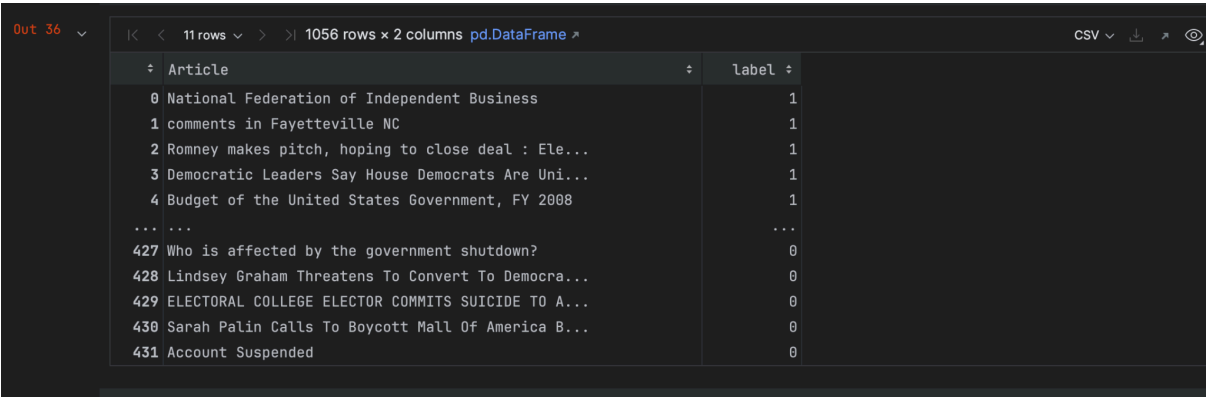
	Unnamed: 0	title	text	label
6330	4490	State Department says it can't find emails fr...	The State Department told the Republican Nati...	REAL
6331	8062	The 'P' in PBS Should Stand for 'Plutocratic'...	The 'P' in PBS Should Stand for 'Plutocratic'...	FAKE
6332	8622	Anti-Trump Protesters Are Tools of the Oligar...	Anti-Trump Protesters Are Tools of the Oligar...	FAKE
6333	4021	In Ethiopia, Obama seeks progress on peace, s...	ADDIS ABABA, Ethiopia –President Obama conven...	REAL
6334	4330	Jeb Bush Is Suddenly Attacking Trump. Here's ...	Jeb Bush Is Suddenly Attacking Trump. Here's ...	REAL

- VIEW OF DATASET 2



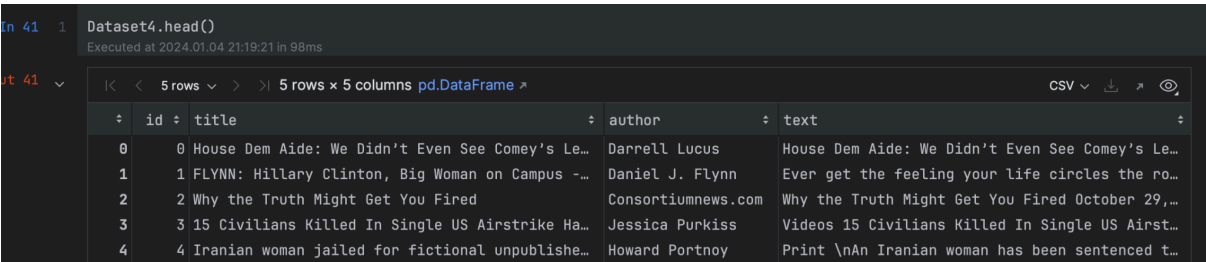
	Article	label
0	as u s budget fight looms republicans flip t...	1
1	u s military to accept transgender recruits o...	1
2	senior u s republican senator let mr muell...	1
3	fbi russia probe helped by australian diplomat...	1
4	trump wants postal service to charge much mor...	1

- VIEW OF DATASET 3



	Article	label
0	National Federation of Independent Business	1
1	comments in Fayetteville NC	1
2	Romney makes pitch, hoping to close deal : Ele...	1
3	Democratic Leaders Say House Democrats Are Uni...	1
4	Budget of the United States Government, FY 2008	1
...
427	Who is affected by the government shutdown?	0
428	Lindsey Graham Threatens To Convert To Democra...	0
429	ELECTORAL COLLEGE ELECTOR COMMITS SUICIDE TO A...	0
430	Sarah Palin Calls To Boycott Mall Of America B...	0
431	Account Suspended	0

- VIEW OF DATASET 4



```
In 41 1 Dataset4.head()
      Executed at 2024.01.04 21:19:21 in 98ms
```

	id	title	author	text
0	0	House Dem Aide: We Didn't Even See Comey's Le...	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Le...
1	1	FLYNN: Hillary Clinton, Big Woman on Campus -...	Daniel J. Flynn	Ever get the feeling your life circles the ro...
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29,...
3	3	15 Civilians Killed In Single US Airstrike Ha...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airst...
4	4	Iranian woman jailed for fictional unpublishe...	Howard Portnoy	Print \nAn Iranian woman has been sentenced t...

● VIEW OF DATASET 5

46 1 Dataset5 = pd.read_csv("data.csv")
Executed at 2024.01.04 21:21:25 in 283ms

47 1 Dataset5
Executed at 2024.01.04 21:21:27 in 780ms

47 ▾ |< < 11 rows > >| 4009 rows x 4 columns pd.DataFrame CSV ▾ ⬇ ⌕ ⋮

	URLs	Headline	Body
0	http://www.bbc.com/news/world-us-canada-41419...	Four ways Bob Corker skewered Donald Trump	Image copyright Getty Imag
1	https://www.reuters.com/article/us-filmfestiv...	Linklater's war veteran comedy speaks to mode...	LONDON (Reuters) - "Last F
2	https://www.nytimes.com/2017/10/09/us/politic...	Trump's Fight With Corker Jeopardizes His Leg...	The feud broke into public
3	https://www.reuters.com/article/us-mexico-oil...	Egypt's Cheiron wins tie-up with Pemex for Me...	MEXICO CITY (Reuters) - Eg
4	http://www.cnn.com/videos/cnnmoney/2017/10/08...	Jason Aldean opens 'SNL' with Vegas tribute	Country singer Jason Aldea
...
4004	http://beforeitsnews.com/sports/2017/09/trend...	Trends to Watch	Trends to Watch\n% of read
4005	http://beforeitsnews.com/u-s-politics/2017/10...	Trump Jr. Is Soon To Give A 30-Minute Speech ...	Trump Jr. Is Soon To Give
4006	https://www.activistpost.com/2017/09/ron-paul...	Ron Paul on Trump, Anarchism & the AltRight	NaN
4007	https://www.reuters.com/article/us-china-phar...	China to accept overseas trial data in bid to...	SHANGHAI (Reuters) - China
4008	http://beforeitsnews.com/u-s-politics/2017/10...	Vice President Mike Pence Leaves NFL Game Bec...	Vice President Mike Pence

In constructing the model, only two pieces of data are essential: the text content and its corresponding label. Consequently, the final dataset will be streamlined to include just two columns: ['Article', 'Label'].

Article Column: This will be a newly created column titled 'Article', which will be a composite of the news item's header and the main text body. This amalgamation ensures that both the headline and the textual content are considered in the analysis, providing a more comprehensive view of each news piece.

Label Column: The labeling system is binary and straightforward. A value of '1' will represent authentic or true news, while a value of '0' will signify that the news is fake or fabricated. This binary classification simplifies the model's task, focusing it on distinguishing between genuine and false information. The aim is to provide a clear, unambiguous dataset that allows the machine learning model to effectively learn and identify patterns indicative of either genuine or fake news articles.

DATA PROCESSING:

1. Eliminate Unnecessary Columns:

- Begin by removing columns that are not essential for the analysis. This step streamlines the dataset, focusing solely on the relevant data, which in this case are the 'Article' and 'Label' columns.

2. Handle Missing Values:

- Identify and eliminate records that contain missing values. This ensures the integrity and completeness of the dataset, as incomplete records could skew the results or hinder the model's performance.

3. Textual Cleanup:

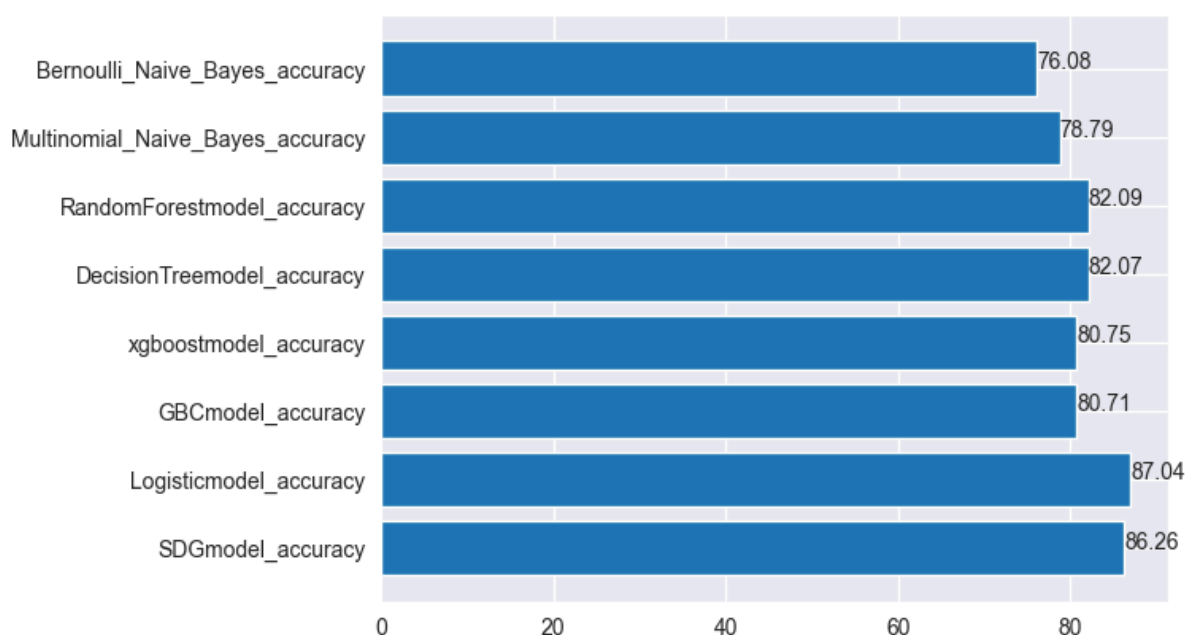
- Remove extraneous elements from the text. This includes eliminating punctuation marks like commas, apostrophes, quotes, and question marks. The goal is to have clean, uncluttered text that aids in more accurate analysis.

4. Exclude Non-Textual Elements:

- Strip the text of any numeric figures and URLs. Removing these elements is crucial as they could introduce noise into the data, potentially confusing the model and impacting its ability to accurately classify news as true or fake.

These preprocessing steps are vital for preparing the data for effective analysis. By cleaning and refining the dataset, you enhance the model's potential for accurate and reliable predictions.

Training and Building the Machine Learning Model



1. Classifier Selection and Feature Feeding:

- A range of classifiers was chosen to predict fake news, including Logistic Regression, Stochastic Gradient Descent, Random Forest, Gradient Boosting Classifier (GBC), XGBoost, Decision Tree, Multinomial Naive Bayes, and Bernoulli Naive Bayes. Each classifier received the same set of extracted features from the preprocessed dataset for analysis.

2. Model Fitting and Evaluation:

- After fitting the models with the data, their performance was evaluated based on accuracy scores and confusion matrices. These metrics provide insight into each model's effectiveness at correctly classifying news articles.

3. Accuracy Score and Model Selection:

- The highest recorded accuracy score was 87.04%. This was achieved after training the models with a substantial dataset comprising over 61,000 records, ensuring robustness and reliability. The Logistic Regression model emerged as the most effective, outperforming others in terms of accuracy.

4. Model Saving and Deployment:

- The Logistic Regression model, identified as the best performer, was saved to disk under the name 'model.pkl'. This saved model can be transferred to other systems by cloning the repository, facilitating easy deployment. The model is designed to accept a news article as input, analyze it, and predict whether it is true or fake.

5. Deployment with Flask:

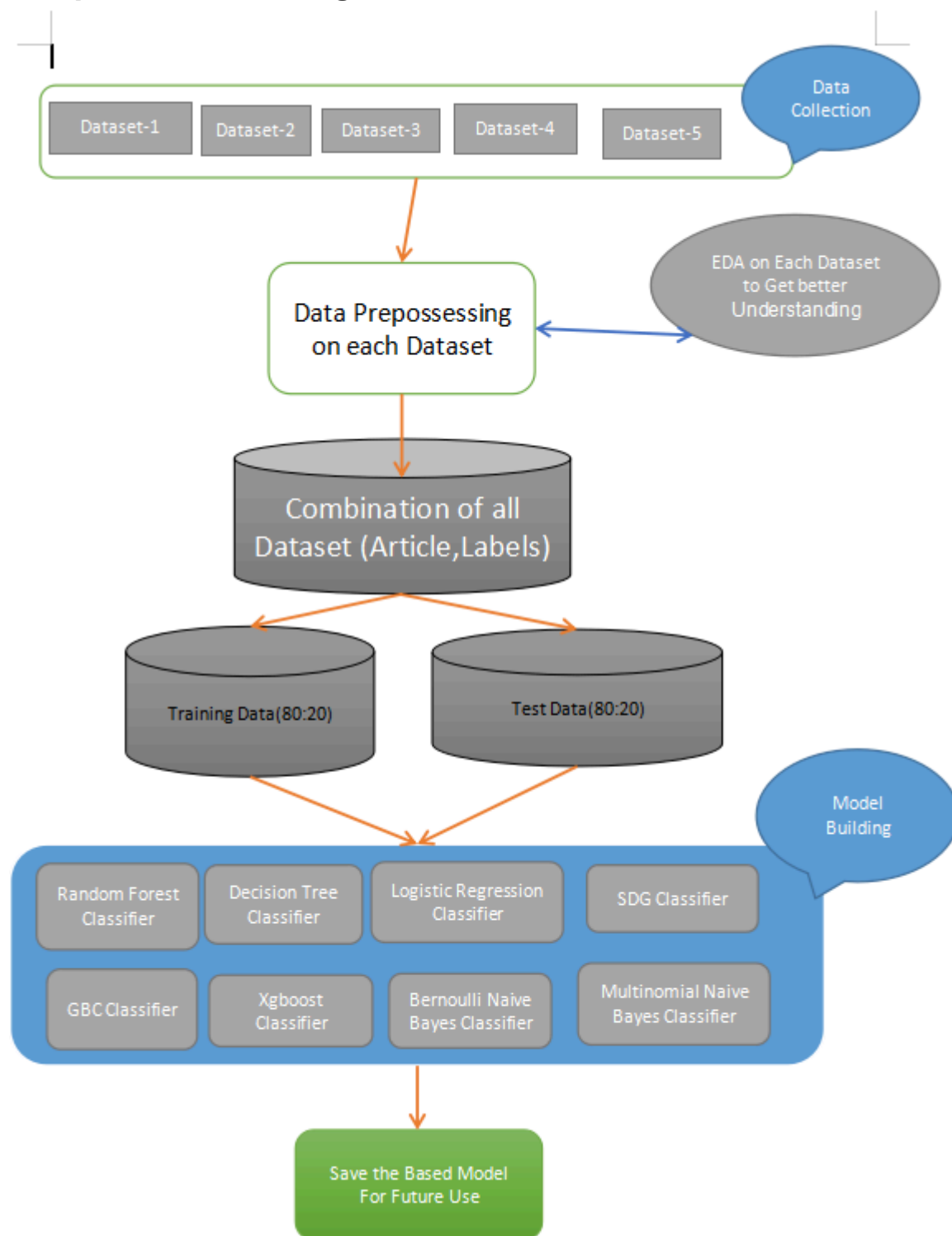
- For deployment, 'model.pkl' is integrated into a web interface using Flask, a Python web framework. This enables users to input news articles and receive immediate predictions regarding their authenticity.

6. Process Flow of Model Building:

- The overall process flow involves data preprocessing, feature extraction, classifier training, model evaluation, selection of the best model based on accuracy, and finally, deployment for user interaction.

This comprehensive approach ensures the development of a robust and accurate system for fake news detection, leveraging the strengths of various classifiers and the efficacy of logistic regression for the final prediction model.

The process flow diagram



1. Data Collection:

- Gathering datasets from various sources.

2. Data Preprocessing:

- Cleaning and preparing each dataset individually, which includes handling missing values, removing unnecessary columns, and text cleaning.

3. Exploratory Data Analysis (EDA):

- Performing EDA on each dataset to gain better understanding and insights.

4. Combining Datasets:

- Merging all datasets into a single dataset with two columns: 'Article' and 'Label'.

5. Splitting Data:

- Dividing the combined dataset into training and test sets, typically with an 80:20 ratio.

6. Model Building:

- Training various classifiers, including Random Forest, Decision Tree, Logistic Regression, Stochastic Gradient

Descent (SGD) Classifier, Gradient Boosting Classifier (GBC), XGBoost, Bernoulli Naive Bayes, and Multinomial Naive Bayes.

7. Model Evaluation and Selection:

- After training, each model is evaluated for performance. The best model is then selected based on the evaluation metrics, which typically include accuracy scores and confusion matrices.

8. Saving the Best Model:

- The best-performing model is saved to disk for future use, ensuring that it can be easily accessed and deployed for predictions.

This structured approach demonstrates a comprehensive methodology for building a machine learning model for the purpose of fake news detection, highlighting the importance of preprocessing and rigorous testing of various algorithms to find the most effective solution.

MODEL DEPLOY:

To deploy, we must develop a basic web interface that allows users to input text and subsequently transmit it to the Flask server. Within the Flask server, we will utilize the stored model model.pkl to ascertain whether the news is genuine or fabricated, and then furnish the outcome to the user via the web interface.

Invia

Entered Text is:

Clinton hates Trump

The AIFAKE Found That is a:

True News

Invia

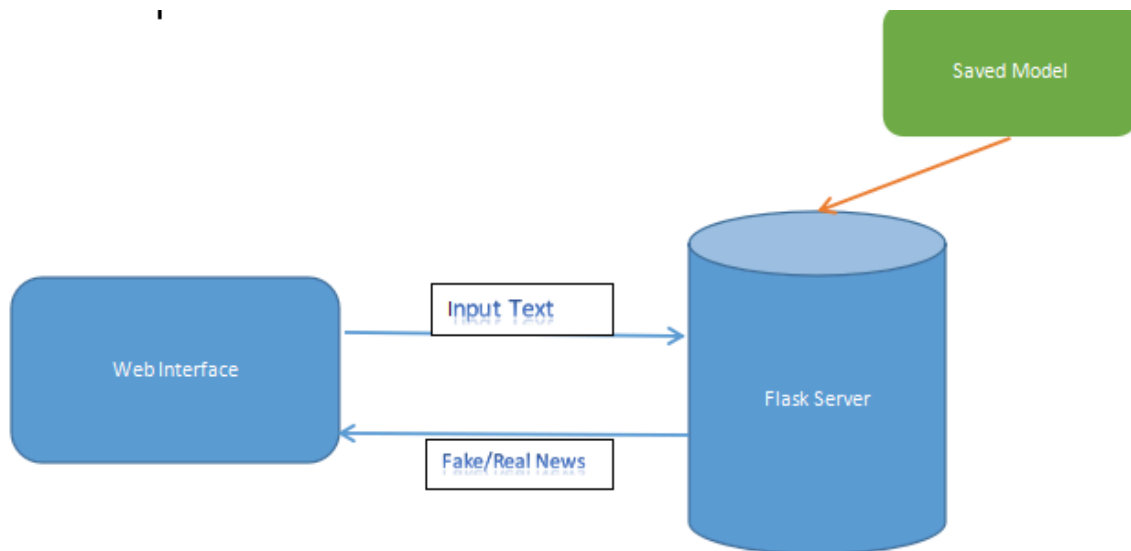
Entered Text is:

trump is dead

The AIFAKE Found That is a:

Fake News

Process Flow of Deploy model:



ss

As we can observe, the top-performing models achieved an 87.04% accuracy score. This can be attributed to the fact that the text still contains stopwords and wordnet, and for classification, is relied on default parameters without exploring Deep Learning-based classification. As we can observe, the top-performing models achieved an 87.04% accuracy score. This can be attributed to the fact that the text still contains stopwords and wordnet, and for classification, is relied on default parameters without exploring Deep Learning-based classification.

The End.

