# Project Title

## Fake News Detection and Evaluation

Submitted By:
Himadri Dhang
BSc Statistics, Asutosh College
Affiliated to Calcutta University

Period of Internship: 25th August 2025 - 19th September 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

# Table of Contents

# 1. Abstract

My Project Work is about detecting fake news from a dataset and evaluate the dataset, using Machine Learning Techniques. A classification model is trained in this project, on textual data, and its effectiveness is evaluated through a confusion matrix to assess accuracy and misclassification patterns.

# 2. Introduction

In today's world, lots of rumors and fake news are spreading there. So we have to be sincere about the data we are receiving from anywhere, we've to recheck the data accuracy twice.

In this project work a dataset is used, which is compiled from real-world sources; the genuine articles were scraped from Reuters.com (a reputable news website). And the fake news articles were gathered from various unreliable platforms identified by Politifact (a U.S.-based fact-checking organization) and Wikipedia. The collection covers articles on diverse subjects, though most of them center around politics and world news.

I've applied Random Forest Classifier and Confusion Matrix in Python to check the accuracy of the data, which I have learnt throughout the internship program. Also I've learned few more things during the internship:

- Handling Lists, loops
- Creating Data Structures
- Introduction to Python OOP part
- Defining Class, Functions
- Introduction to Numpy, Pandas like libraries and more
- Machine Learning
- Introduction to LLM

# 3. Project Objective

Main objective of doing this project is very simple, doing experiment with data with Python libraries and learning through it. Here I classified the objective into two categories:

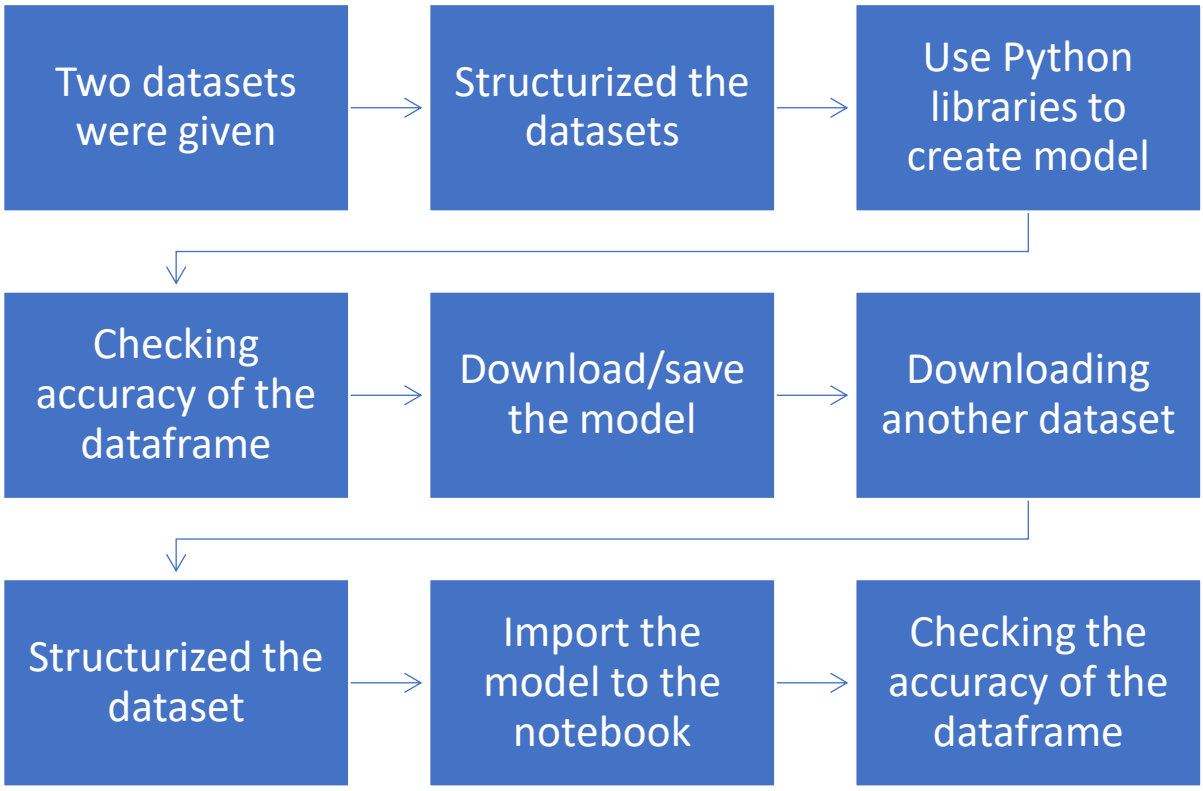
- Learning data handling
    - Use of Numpy, Pandas to structure the data
    - Use of Matplotlib to visualize

- Accuracy Checking
  - Use of Scikitlearn to check the accuracy

# 4. Methodology

To perform the project well, I'm provided two textual datasets and a well-designed Google Colab Notebook, for writing codes and evaluating the datasets. Also I've collected an other dataset for testing the Random Forest Classifier model I developed.

Here the short flow-chart of my project work.




In this project, it was said to split the Training and Testing sets with 25% test size and the restriction was given to build a Random Forest Classifier for classification of data and to predict the outcomes for test data.

The final Colab Notebook with Python codes for the project can be found here:

https://github.com/HimadriDhang06/IDEAS-TIH_Internship-2025/blob/Colab-Notebook/04-fake_news-detection-and-evaluation.ipynb

# 5. Data Analysis and Results

In terms of Data Analysis, I paste here the finding results by subject sorting and Random Forest Classifier, and the Visualizations by the Confusion Matrix, also the plottings.

Findings:

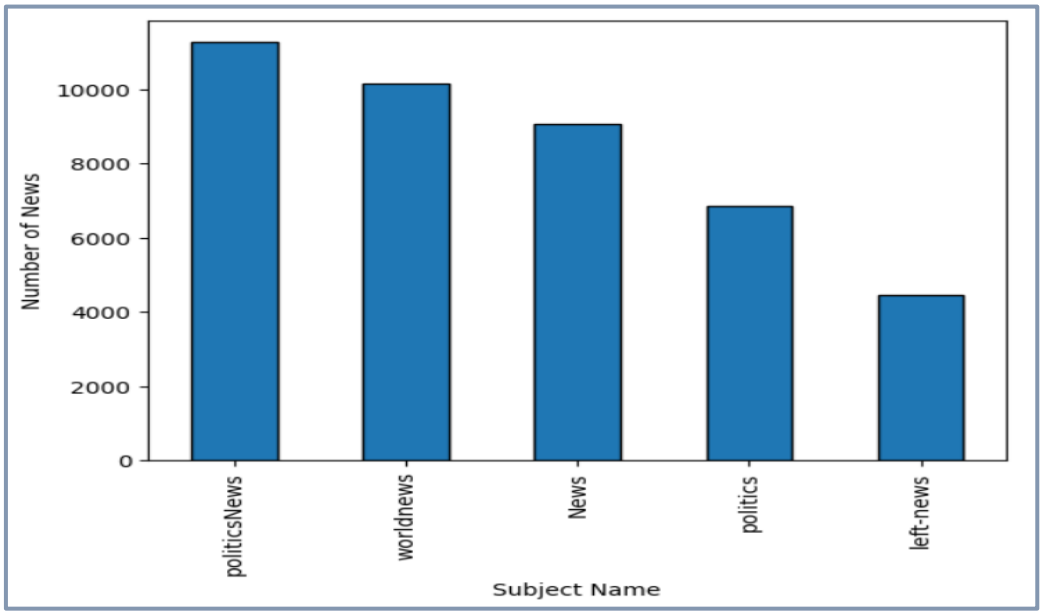
| | |
|---|---|
| Data on Topics | 6. Middle-East News: 1.73%<br>7. US News: 1.74%<br>8. Government News: 3.50%<br>9. Left News: 9.93%<br>10. Politics: 15.24%<br>11. News: 20.16%<br>12. World News: 22.605<br>13. Politics News: 25.11% |
| Accuracy score | • Logistic Model: 0.9399554<br>• Random Forest Classifier: 0.9165256 |
| Precision score | • Logistic Model: 0.9437521<br>• Random Forest Classifier: 0.9259324 |
| Recall score | • Logistic Model: 0.9398483<br>• Random Forest Classifier: 0.9114098 |
| F1 score | • Logistic Model: 0.9417962<br>• Random Forest Classifier: 0.9186137 |

Footnote: Model scores are based on 25% test size

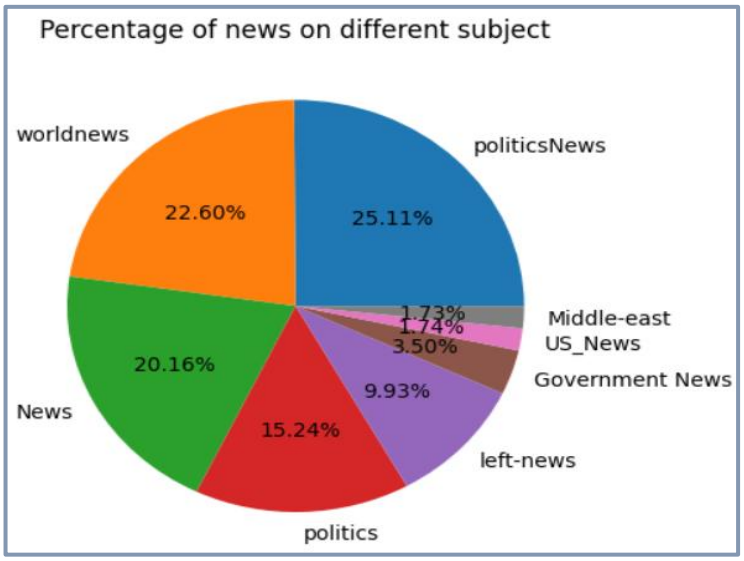
Visualizations:

In the project Histogram, Pie diagram and Confusion matrices for Logistic Model and random Forest Classifier are plotted.

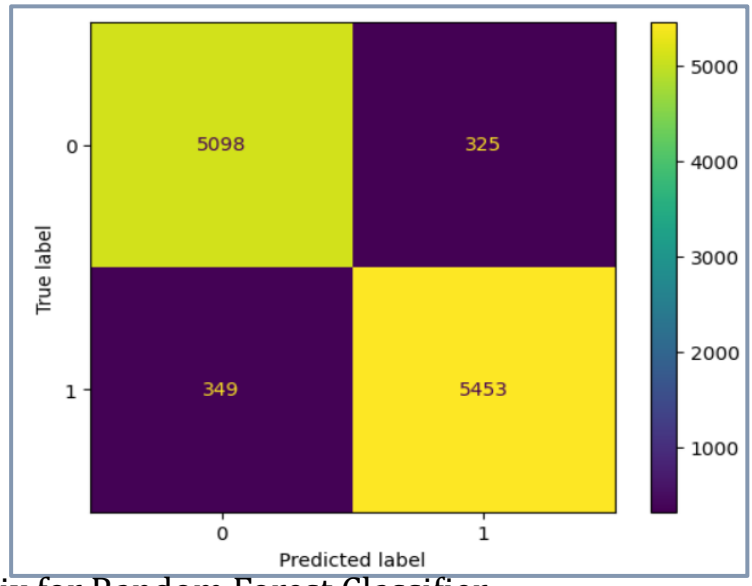
Histogram:



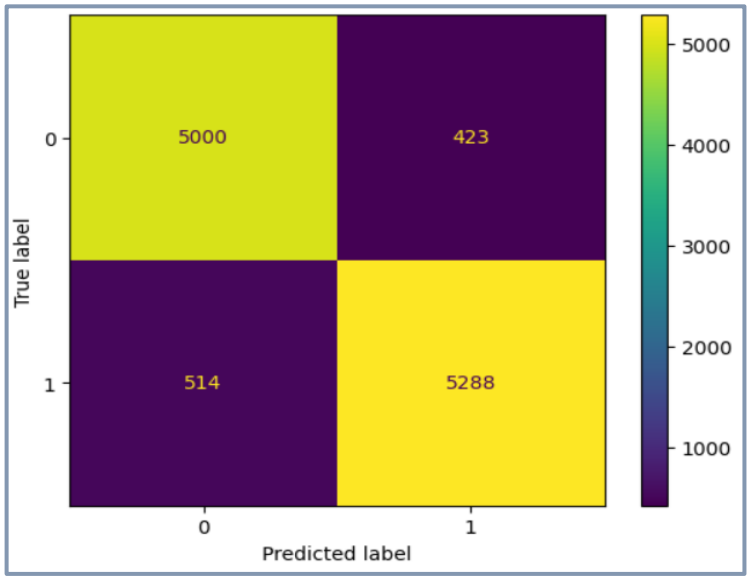
## Pie diagram:




## Confusion matrix of Logistic Model:




## Confusion matrix for Random Forest Classifier:



Comparative Analysis:

Based on the scores obtained from the models, it is clearly seen that the Logistic Model has better accuracy rate than the Random Forest Classifier. Also, the former has better precision, recall and F1 score compared to the latter. So, as far I did the project, I would like to prefer Logistic Model, as it has higher scores for the data; hence it is more workable model.

# 14. Conclusion

While doing the project, I was asked by myself multiple times when I was about to use new code snippet, that is it enough or I need to clarify more; but integrated Gemini handled it very well, it helped me to explain the error occurred and sometimes I had doubts, solved by Gemini.

After the project done, a vast window is open to me about Python language. I was new to it, and now I learned how to plot things, how we can build model and train it for our purposes, and etc. Meantime I've acquired a bit of knowledge for some of very important libraries of Python, which are really impressive. Now, after this project I can work upon several fundamental projects on myself, for gaining better knowledge.

# 15. APPENDICES

- Github Repository Link: https://github.com/HimadriDhang06/IDEAS-TIH_Internship-2025.git