جامعة أم القرى
UMM AL-QURA UNIVERSITY

REPORT
DATA ANALYSIS (2)
DS3114

Dr. Omaima Fallatah

Team Members:

- Faten Matouq Almowallad 444000266
- Ayah Bakur Alhawsawi 444006678

https://colab.research.google.com/drive/1TkJMUcwVtpcQECNOlm8KJ9PsEV14eGML?usp=sharing#scrollTo=RuBK2piFlgkg

Tasks: 3

Text Analysis

(Fake News)

INTRODUCTION

In recent years, the spread of fake news has become a significant issue due to its potential to mislead the public and influence opinions on various subjects. Fake news detection aims to classify news articles as either fake or real using machine learning techniques. This project employs classification algorithms to analyze textual data and differentiate between fake and real news articles with high accuracy.

ABOUT DATASET

The dataset used in this project consists of two CSV files: one containing fake news and the other containing real news articles. The main columns of the dataset are:

- **title:** The title of the news article.
- **text:** The body text of the article.
- **subject**: The category or subject of the news (e.g., politics, world news).
- **date:** The date of publication.
- **class:** A label indicating if the article is fake (0) or real (1).

The dataset contains:

- Fake News Samples: 23,481 entries.
- Real News Samples: 21,417 entries.

ANALYSIS & RESULTS

This project involves various machine learning algorithms for text classification:

- **Logistic Regression**: Achieved an accuracy of 98.6%.
- **Gradient Boosting Classifier**: Achieved an accuracy of 99.6%, making it the best-performing model.

The performance of each model was evaluated using a confusion matrix, classification report (including precision, recall, and F1-score), and ROC curve analysis. These metrics help in understanding the model's ability to accurately classify fake and real news.

Data Preprocessing

To prepare the text data for model training, several preprocessing steps were applied:

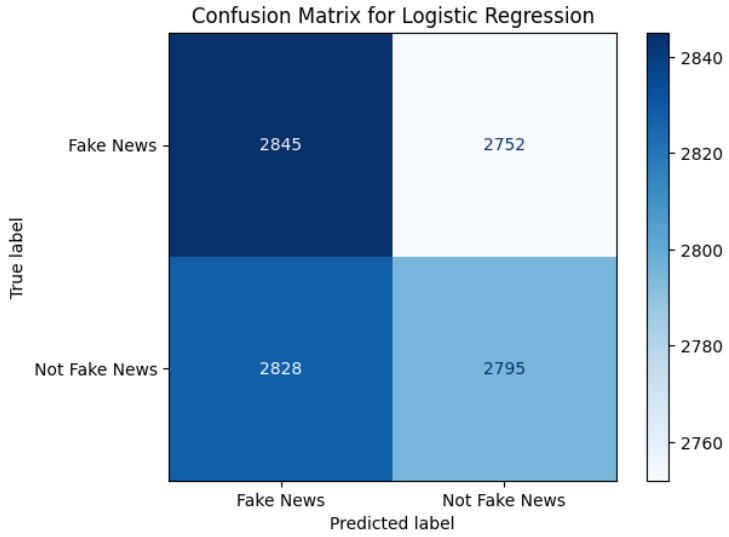
- **Lowercasing**: Converted all text to lowercase for uniformity.
- **Removing Special Characters**: Removed punctuation, URLs, and special characters.
- **Tokenization**: Split the text into individual words.
- **Stopwords Removal**: Removed common English stopwords that do not add value to the model.
- **TF-IDF Vectorization**: Transformed the text data into numerical form using Term Frequency-Inverse Document Frequency (TF-IDF). This method emphasizes important words while reducing the impact of less informative words.

The data was split into training and testing sets (75% training, 25% testing) to validate the model's performance.

Visual Analysis

Visual analysis provided insights into the data and model performance:

- **Confusion Matrix:** Displayed the number of true positives, true negatives, false positives, and false negatives, offering a clear view of model accuracy.
- **Distribution of Fake vs. Real News:** A bar chart showed the counts of fake and real news, illustrating the balance of the dataset.
- **Feature Importance:** For the Gradient Boosting Classifier, feature importance scores were plotted, showing the words that contributed most to the model's decisions.
- **ROC Curve:** Illustrated the trade-off between true positive rates and false positive rates, allowing for threshold adjustments to improve model sensitivity.

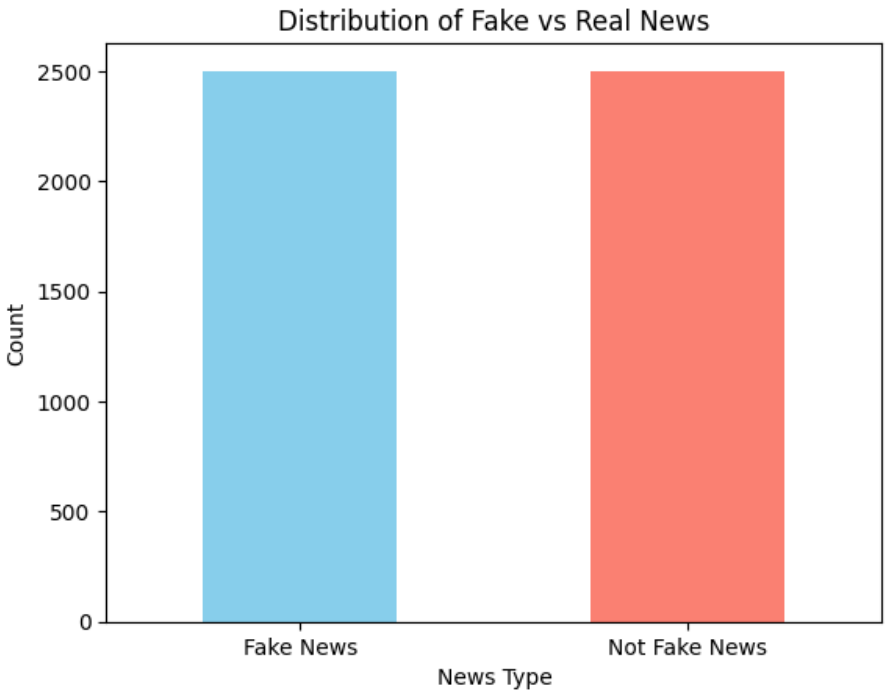


the confusion matrix for the Logistic Regression model. It shows the number of true positives, true negatives, false positives, and false negatives. This visualization helps in assessing how well the model distinguishes between "Fake News" and "Not Fake News":

- **True Positives (Bottom-right)**: Correctly identified "Not Fake News."
- **True Negatives (Top-left)**: Correctly identified "Fake News."
- **False Positives (Top-right)**: Misclassified "Fake News" as "Not Fake News."
- **False Negatives (Bottom-left)**: Misclassified "Not Fake News" as "Fake News."

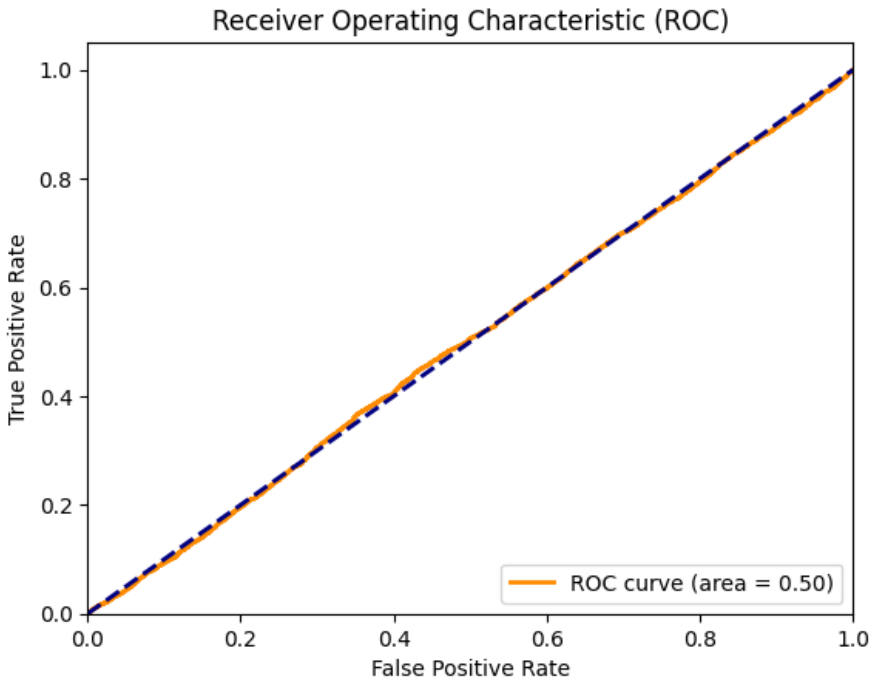
Distribution of Fake vs. Real News:

A bar plot to visualize the counts of "Fake News" and "Not Fake News" in the dataset.



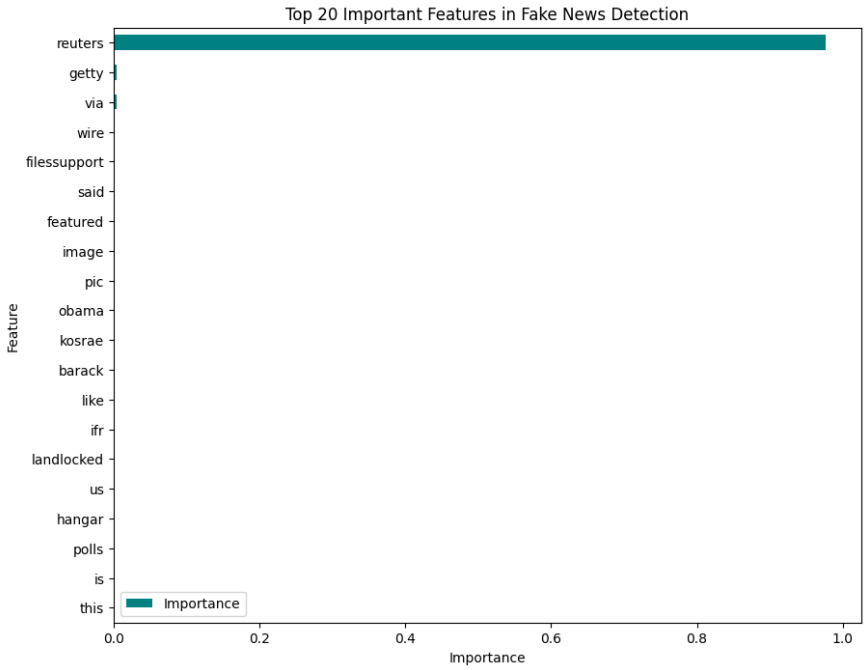
ROC Curve:

Visualize the trade-off between the true positive rate (TPR) and false positive rate (FPR).



Feature Importance (For Gradient Boosting Classifier):

Display which words (features) are most important in determining if the news is fake or real.



CONCLUSION

The fake news detection project demonstrates the effectiveness of machine learning in text classification tasks. By utilizing advanced algorithms like Gradient Boosting Classifier and text preprocessing techniques such as TF-IDF, the project achieved a high accuracy rate in distinguishing between fake and real news articles. The visualizations provided valuable insights into the model's behavior and the nature of the data. With further tuning and testing, this approach could be applied to real-time news verification systems, helping to combat the spread of misinformation.

REFERENCES

YETIM, E. (2022) *Fake News Detection Datasets*. Available at: https://www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets (Accessed: 26 October 2024).