FAKE NEWS DETECTION USING MACHINE LEARNING

BY – PREETI RANJAN

AGENDA

- OBJECTIVE
- PROBLEM STATEMENT
- APPROACH TO SOLVING PROBLEM
    - REAL WORLD IMPACT OF FAKE NEWS
- MODEL SUMMARY
    - DATASET DESCRIPTION
    - DATA PRERROCESSING STEPS
    - FEATURE ENGINEERING (TF - IDF)
    - MACHINE LEARNING APPROACH
    - MODEL SELECTION
    - DATA SPLITTING AND TRAINING
    - MODEL EVALUATION
- RESULTS
    - PERFORMACE METRICS
    - CONFUSION MATRIX AND GRAPHS
- KEY OBSERVATION, STRENTHS AND LIMITATIONS
- INFERENCE/CONCLUSION
- REFERENCES
- FUTURE SCOPE AND IMPROVEMENT

OBJECTIVE:

To build a model that can differentiate real news from fake news.

- Fake news has become a serious problem in the digital age.
- Misinformation spreads rapidly through social media and online platforms.
- Machine learning can help detect and classify fake news.

PROBLEM STATEMENT

What is the problem?

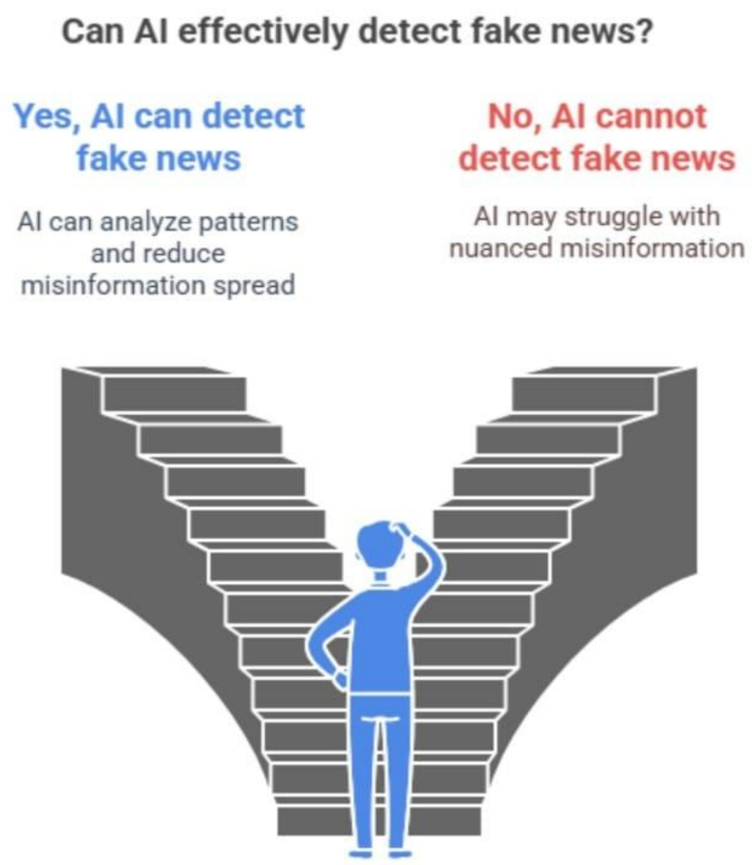
- Fake news misleads people and influences public opinion.
- Can impact elections, finance, and global stability.

Why is it important?

- Manual fact-checking is not scalable.
- Automated machine learning solutions can help.

APPROACH TO SOLVING THE PROBLEM

- Rise of social media has increased fake news circulation.
- Misinformation affects politics, public health, and businesses.
- Key Question: Can AI effectively detect fake news?
- Our goal: Develop a machine learning model for fake news classification.



Real-World Impact of Fake News

Real-World Impact of Fake News

- **Political Influence:** Fake news can manipulate elections.
- **Health Risks:** Spreading misinformation about COVID-19 and vaccines.
- **Economic Impact:** Stock market fluctuations due to fake financial news.

MODEL SUMMARY

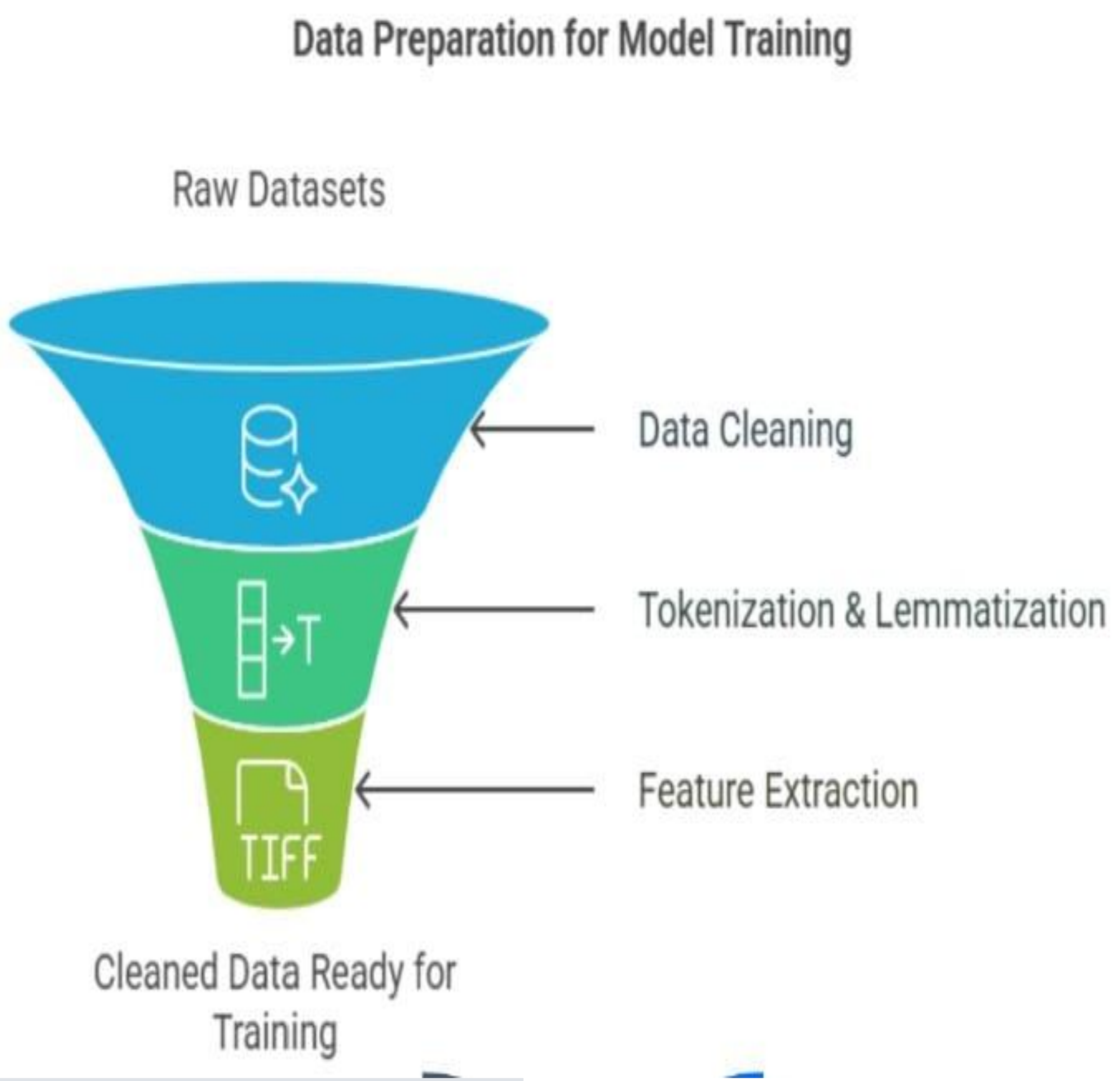
❑ Dataset Description

- **Source: Kaggle Fake News Dataset.**
- **Number of Articles:**
  - o Fake News: ~24,000 articles
  - o Real News: ~21,000 articles

- **Attributes in Dataset:**
  - o Title: The headline of the article.
  - o Text: The body content of the article.
  - o Subject: The category (Politics, World, etc.).
  - o Date: The publication date.

DATA PREPROCESSING STEPS

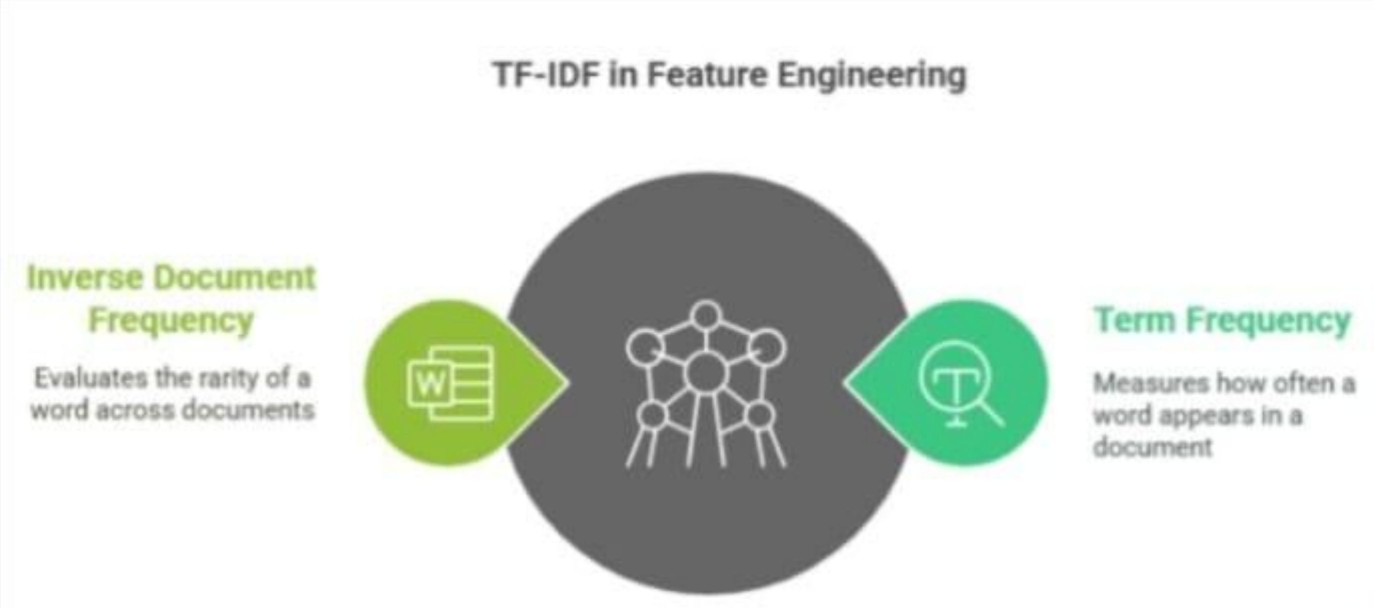
- **Removing Unnecessary Elements:** Punctuation, numbers, and special characters.
- **Lowercasing: Standardizing text.**
- **Removing Stop words:** Words that do not add meaning (e.g., "the", "is", "and").
- **Tokenization:** Splitting text into individual words.
- **Lemmatization:** Converting words to their base form (e.g., "running" → "run").



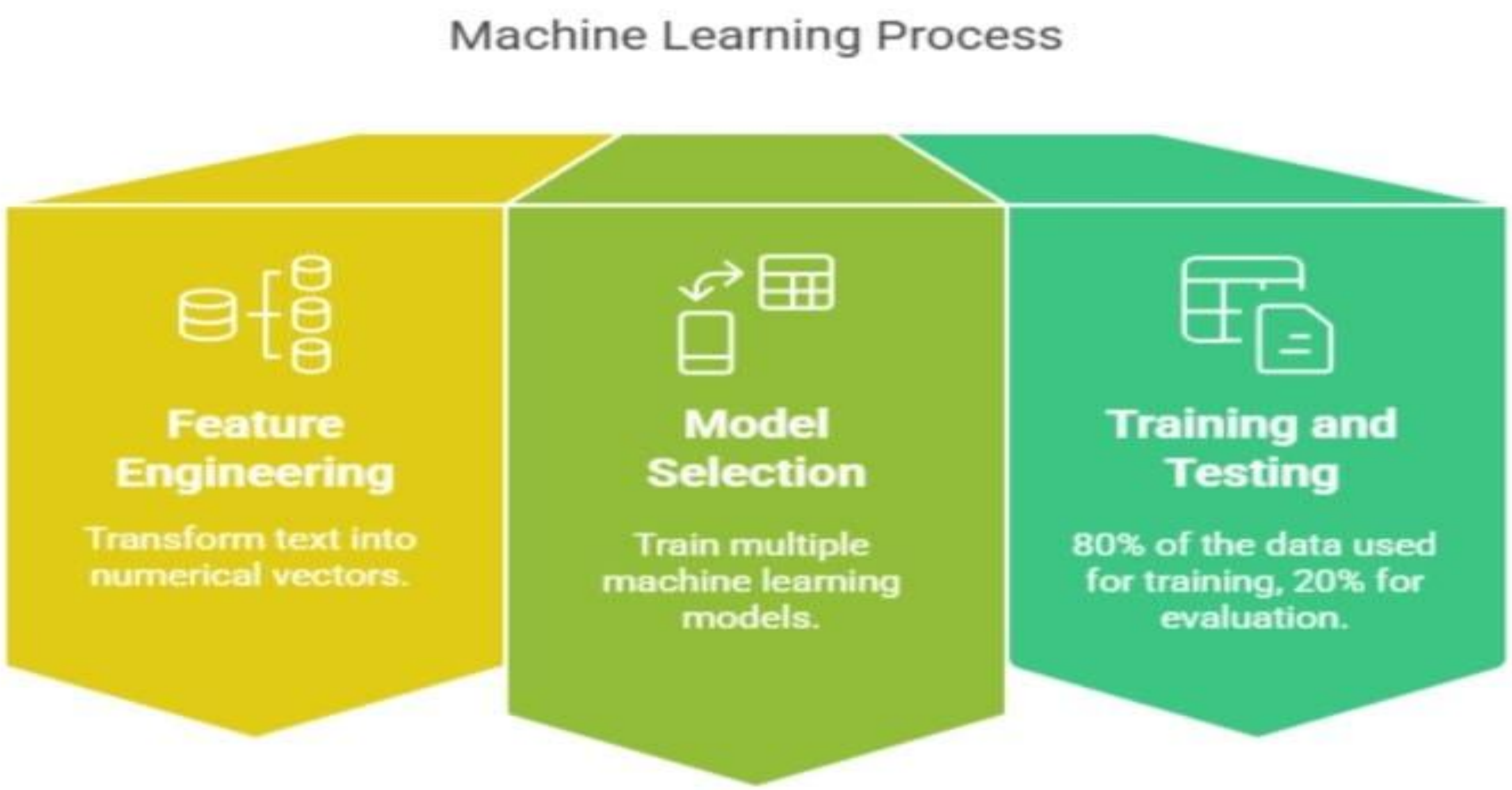
FEATURE ENGINEERING (TF - IDF)

TF-IDF (Term Frequency-Inverse Document Frequency)

- Assigns importance to words based on their frequency.
- Helps convert text into numerical features for machine learning.



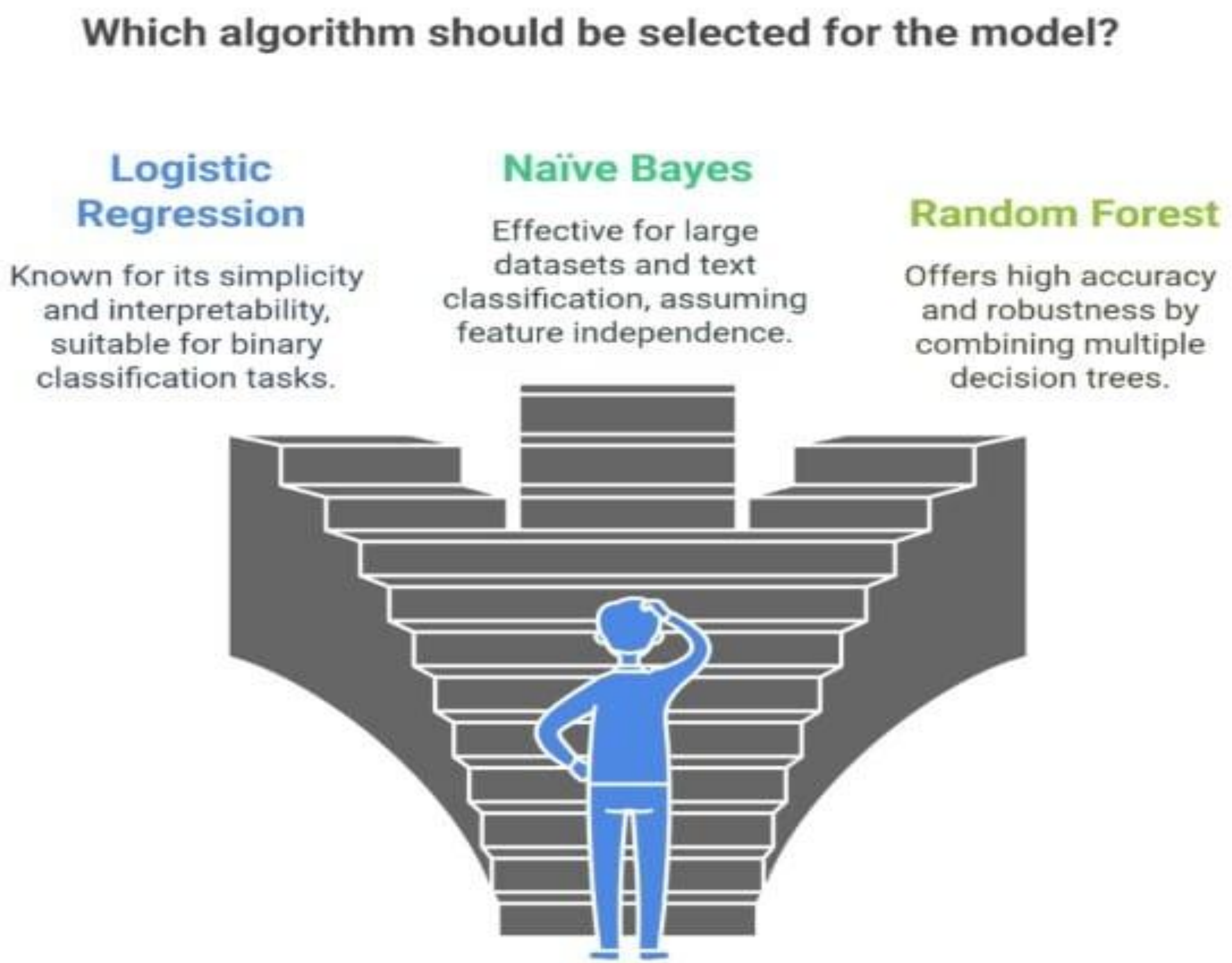
MACHINE LEARNING APPROACH



MODEL SELECTION

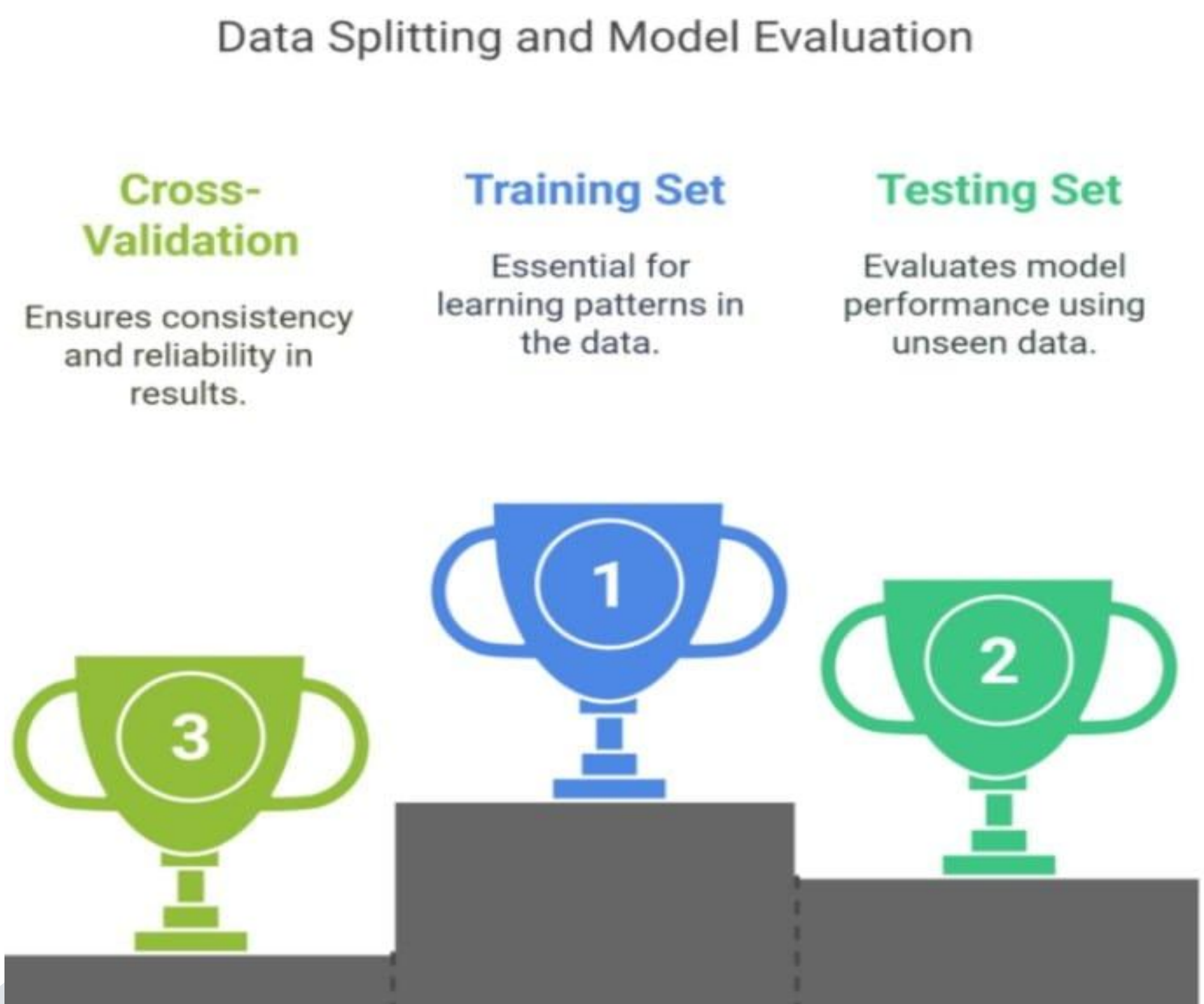
ALGORITHMS USED:

- LOGISTIC REGRESSION
- NAÏVE BAYES
- RANDOM FOREST



DATA SPLITTING AND TRAINING

- **Training Set (80%)** – Used for learning patterns in data.
- **Testing Set (20%)** – Used to evaluate model performance.
- **Cross-Validation:** Used to ensure consistency in results.



MODEL EVALUATION

Metrics Used for Evaluation:

- Accuracy – Correct predictions out of total cases.
- Precision – Percentage of true fake news predictions.
- Recall – Percentage of actual fake news correctly identified.
- F1 Score – Balance between precision and recall.
- Confusion Matrix – Breakdown of model's correct and incorrect classifications.

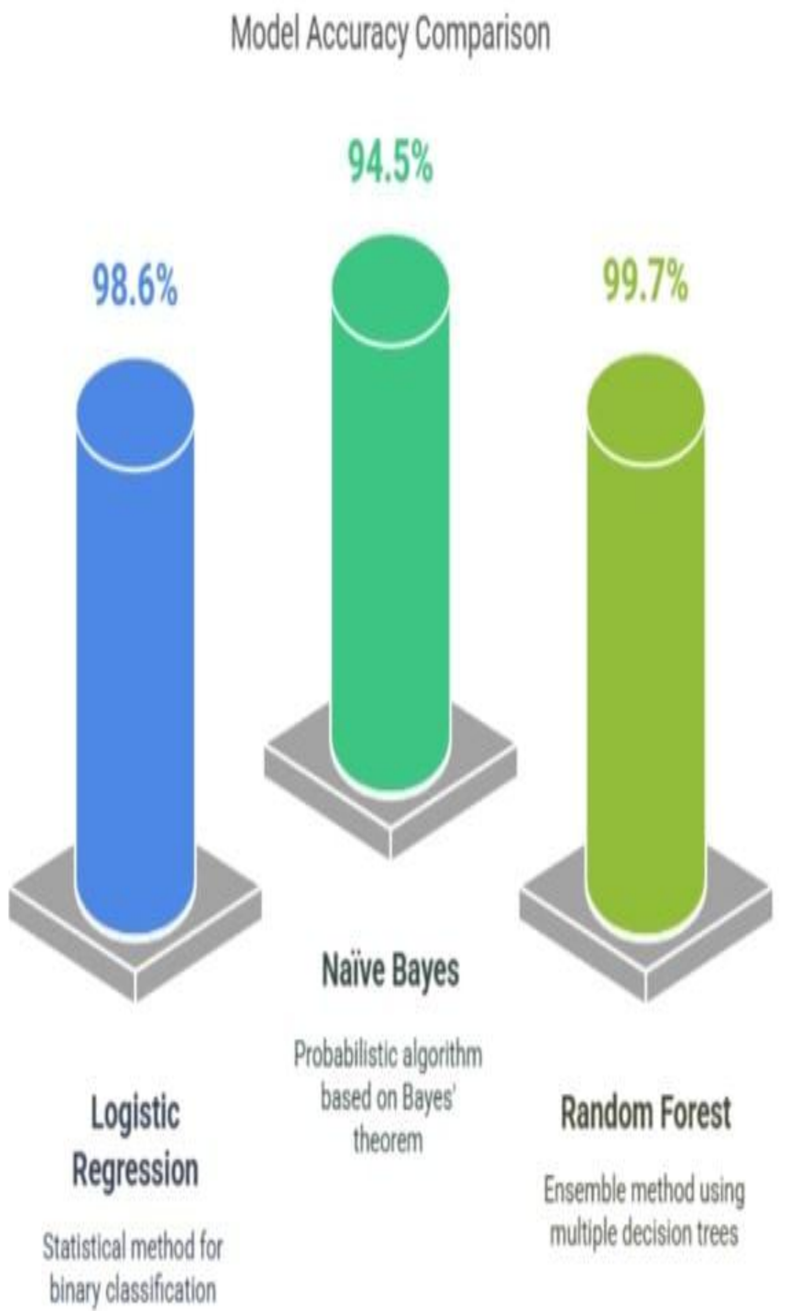
RESULTS

ACCURACY

Logistic Regression Accuracy: 0.9878619153674832
Naïve Bayes Accuracy: 0.9239420935412027
Random Forest Accuracy: 0.9973273942093541

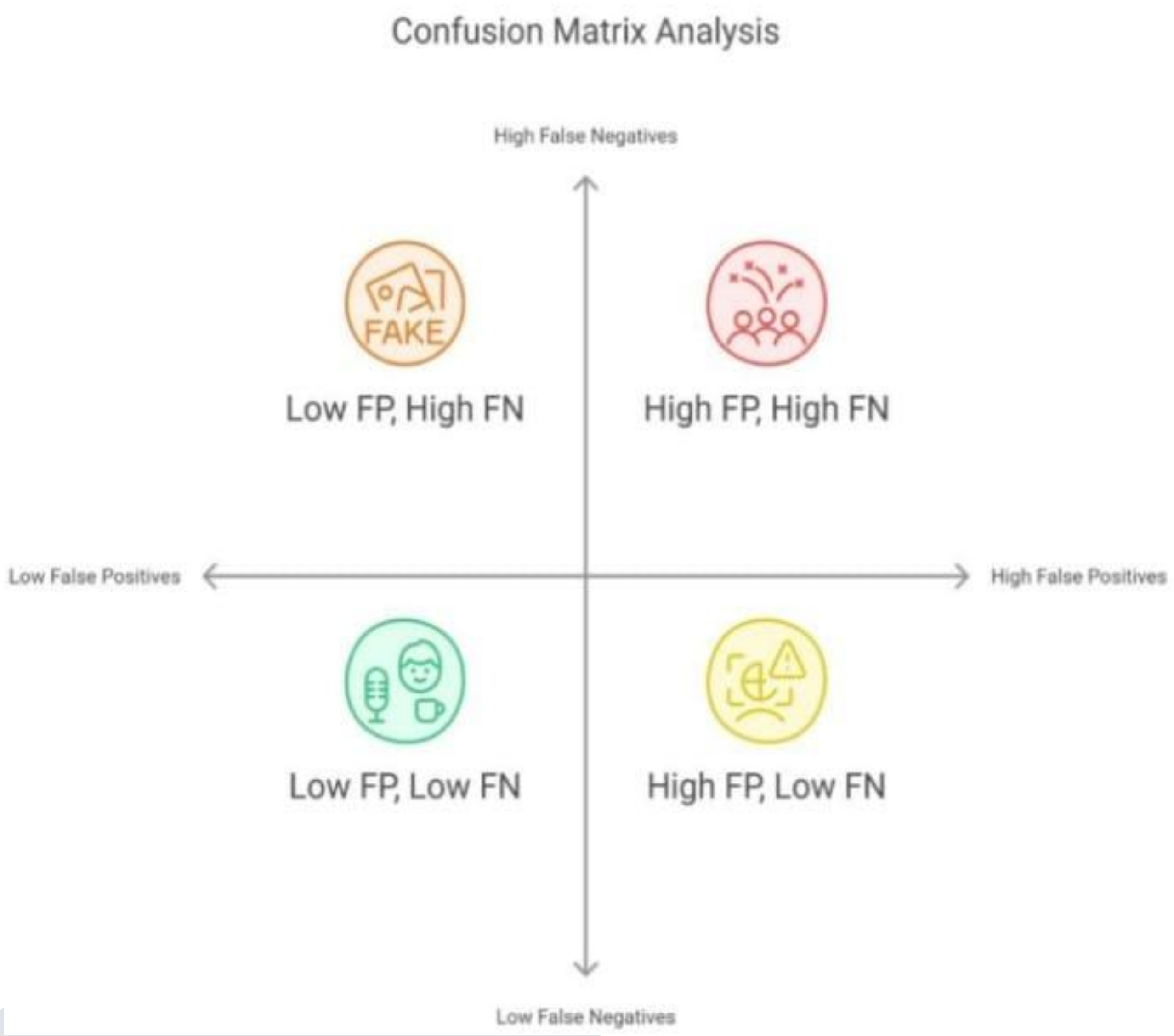
Comparison of Models:

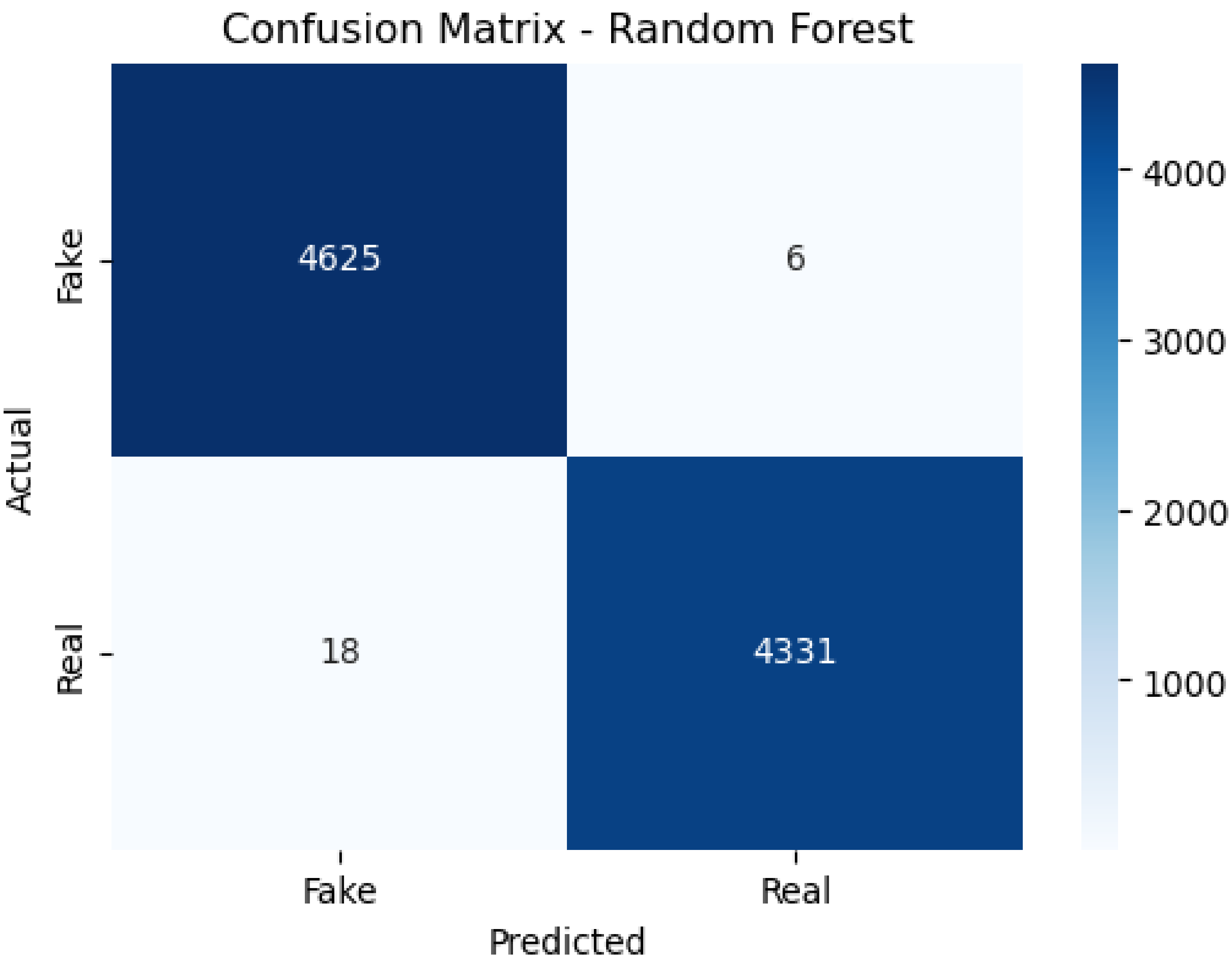
- Conclusion:
  Random Forest performed the best!

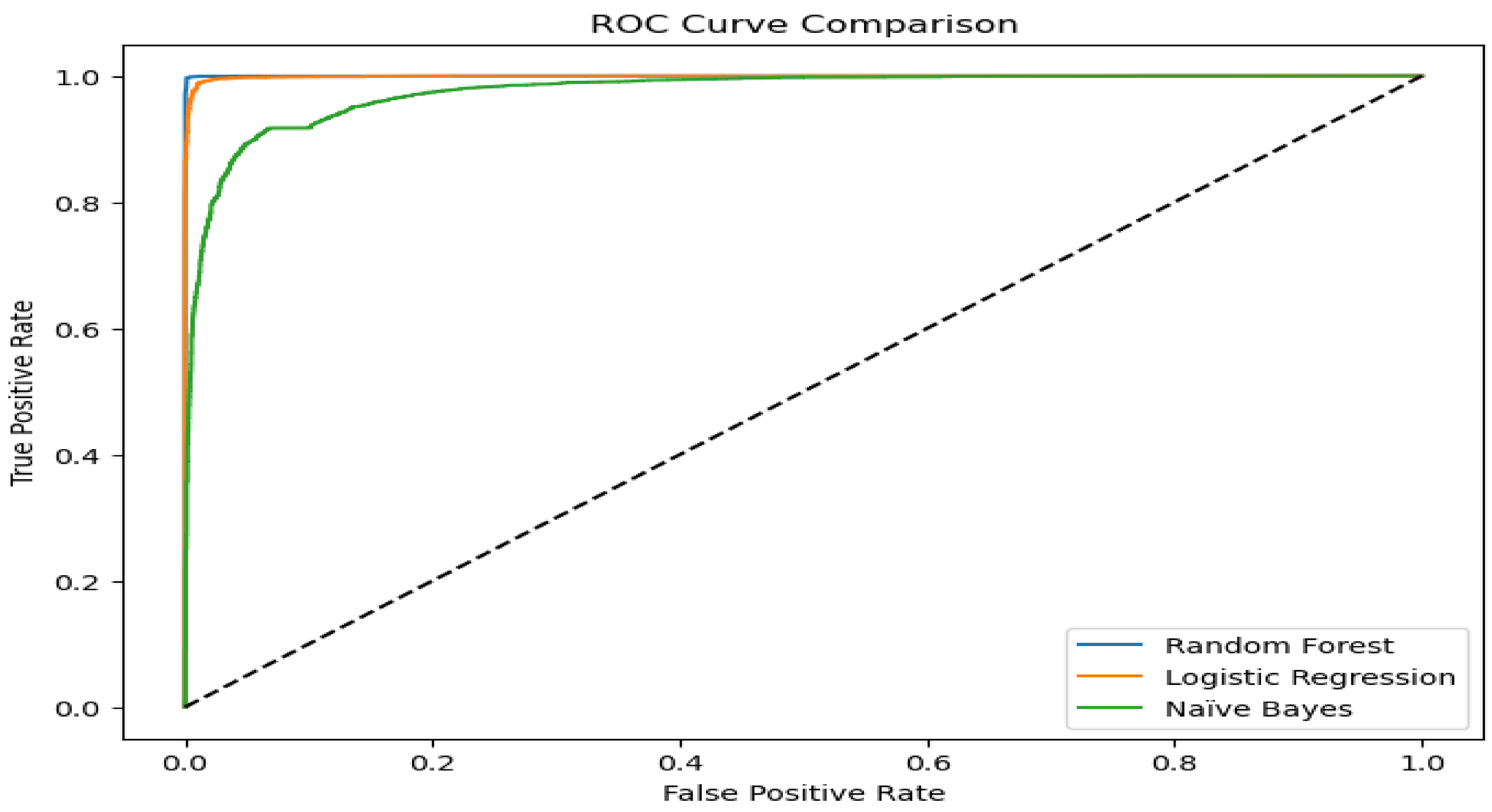


CONFUSION MATRIX AND GRAPHS

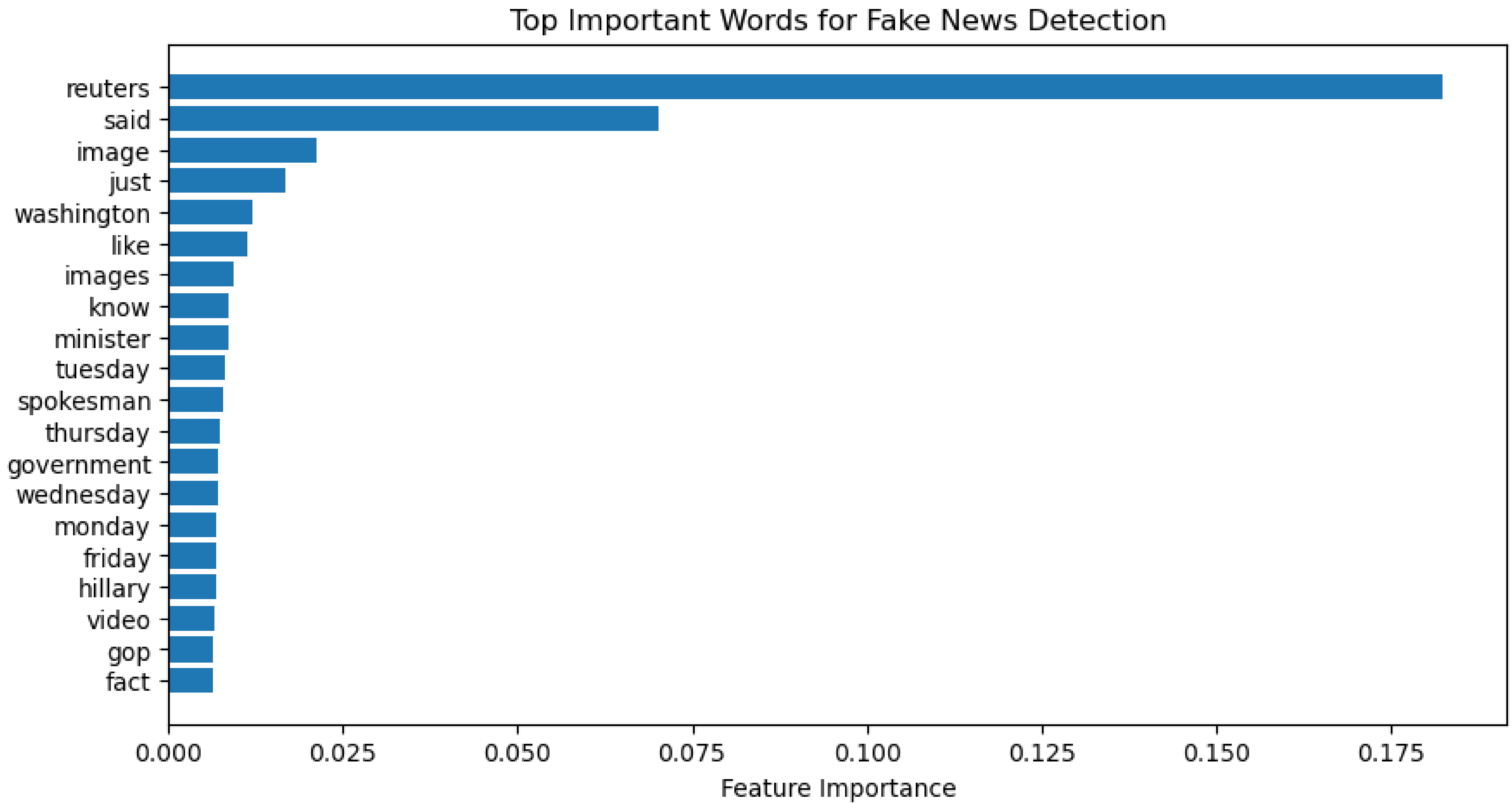
- Confusion Matrix: Visual representation of model performance.
- True Positives (TP) – Correctly identified real news.
- True Negatives (TN) – Correctly identified fake news.
- False Positives (FP) – Real news misclassified as fake.
- False Negatives (FN) – Fake news misclassified as real.
- ROC Curve: Graph showing the trade-off between true positives and false positives.




Confusion Matrix - Random Forest
Actual
Fake
Real
4625
6
18
4331
Fake
Real
Predicted
4000
3000
2000
1000



ROC Curve Comparison
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Random Forest
Logistic Regression
Naïve Bayes



Top Important Words for Fake News Detection
reuters
said
image
just
washington
like
images
know
minister
tuesday
spokesman
thursday
government
wednesday
monday
friday
hillary
video
gop
fact
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Feature Importance

KEY OBSEVATIONS, STRENGHTS AND LIMITATIONS

Key Observations

- Random Forest outperformed other models.
- Feature importance analysis highlighted key words in fake news.
- TF-IDF played a crucial role in distinguishing fake vs. real news.

Strengths

- High accuracy achieved with Random Forest.
- Effective data preprocessing improved model performance.

Limitations

- Dataset is limited to English-language news.
- Fake news evolves, requiring continuous model updates.

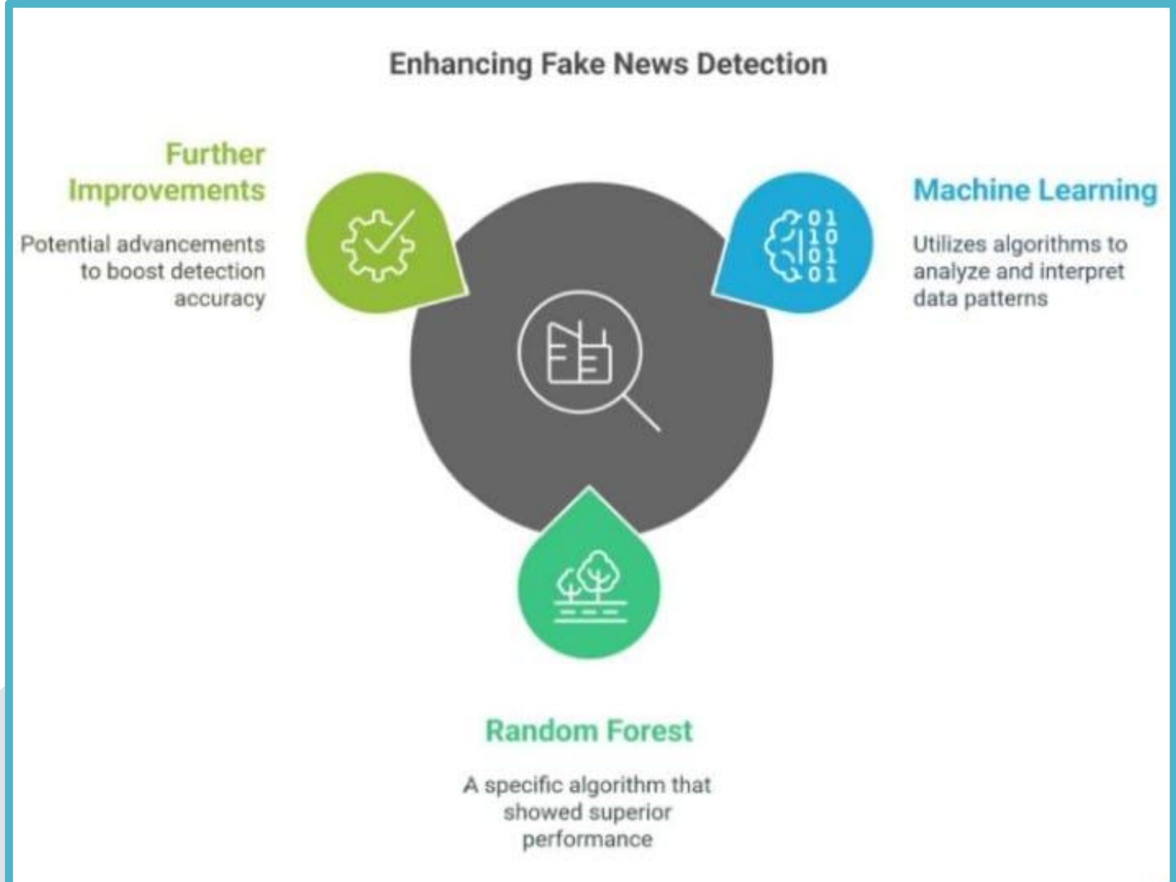
INFERENCE / CONCLUSION

Final Takeaways:

• Machine learning is effective in detecting fake news.

• Random Forest performed best in our experiments.

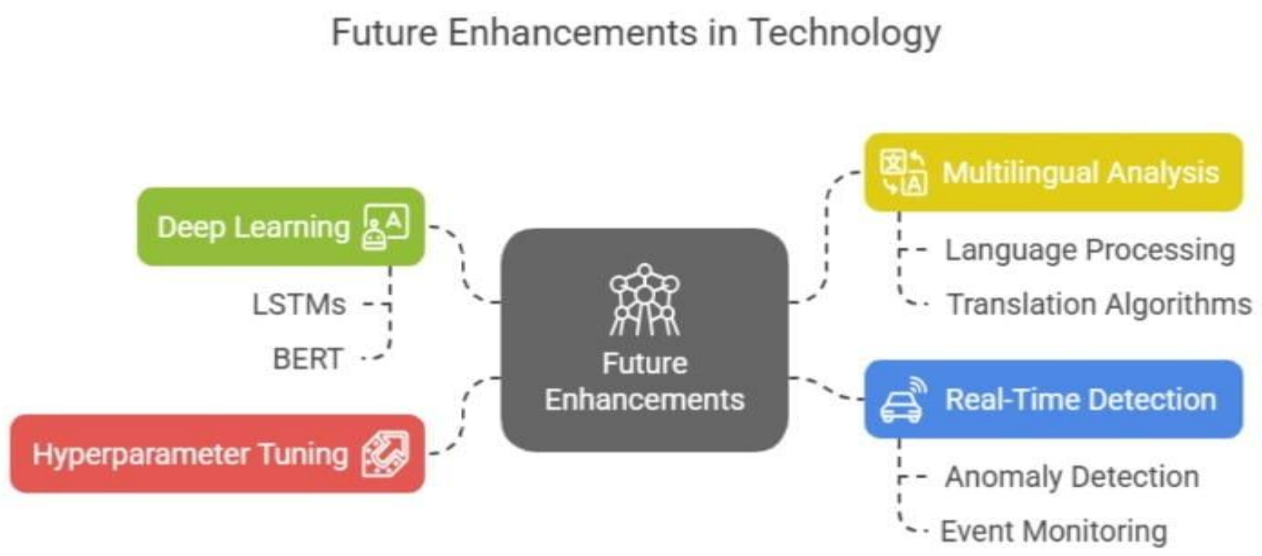
• Further improvements can enhance detection capabilities.



REFERENCES

- ❑ Code And Dataset Link
- https://github.com/preeti2207ranjan/Fake-News-/blob/main/Fake_News_Detection.ipynb
- Dataset code
- https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset/data
- ❑ Hands on Machine Learning with Scikit-Learn, Keras, and Tensorflow-Aurelien Geron

FUTURE SCOPE AND IMPROVEMENTS



1.Deep Learning Approaches:
- Implement LSTMs or BERT for better text understanding.

2.Hyperparameter Tuning:
- Optimizing models to improve accuracy further.

3.Real-Time Fake News Detection:
- Deploying the model as an API for live analysis.

4.Multilingual Analysis:
- Expanding dataset to detect fake news in multiple languages.

THANK YOU

PREETI RANJAN

preeti.ranjan17@gmail.com