Fake News Detection – Project Report

1. Introduction

In today's digital world, social media and online platforms have become the primary sources of news and information. However, the spread of **fake news** has emerged as a major problem, influencing public opinion and creating misinformation.
This project aims to **build a machine learning model to detect fake news articles** and classify them as either **Real** or **Fake** based on their textual content.

Objectives of the Project:

- Preprocess and clean news article text.
- Train multiple machine learning models for classification.
- Compare model performance based on evaluation metrics.
- Visualize results using graphs and plots.

---

2. Dataset

The dataset used in this project is the **Fake and Real News Dataset** from Kaggle.

- **Columns used:**
  - `content` → textual content of the news article
  - `label` → target variable (Fake / Real)

Dataset Statistics:

- Total Articles: ~44,000
- Real News: ~21,000
- Fake News: ~23,000

---

3. Methodology

3.1 Data Preprocessing

Steps applied to clean and prepare the text:

1. Removed punctuation, numbers, and special characters.

2. Converted text to lowercase.
3. Tokenized sentences into words.
4. Removed **stopwords** (e.g., “the”, “and”, “is”).
5. Applied **lemmatization** to reduce words to their root form.

3.2 Feature Extraction

- Used **TF-IDF Vectorizer** with bigrams (`ngram_range=(1,2)`) and a maximum of 20,000 features.
- Converted text into numerical feature vectors.

3.3 Models Used

We trained and evaluated the following models:

1. Logistic Regression
2. Linear SVC
3. Multinomial Naïve Bayes
4. Random Forest Classifier
5. Gradient Boosting Classifier

---

4. Experiments and Results

4.1 Evaluation Metrics

We used:

- **Accuracy** – overall correctness
- **Precision** – reliability of positive predictions
- **Recall** – ability to capture fake news correctly
- **F1-Score** – balance between precision and recall
- **ROC-AUC** – area under the ROC curve

4.2 Results Summary

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.99	0.95	0.94	0.95
Linear SVC	0.99	0.94	0.93	0.94
Multinomial NB	0.96	0.89	0.91	0.90
Random Forest	0.01	0.92	0.91	0.92
Gradient Boosting	0.99	0.93	0.92	0.93

Best Model → Logistic Regression (99% Accuracy)

---

5. Visualizations

1. **Model Accuracy Comparison**
   Bar chart comparing accuracy of all models.
2. **Confusion Matrices**
   For each model, confusion matrices were plotted to show correct vs incorrect classifications.
3. **ROC Curves**
   Plots showing trade-off between True Positive Rate and False Positive Rate for each model.
4. **Word Clouds**
   Word clouds for *Fake* and *Real* news to highlight most frequent terms.

---

6. Discussion

- Logistic Regression performed the best due to its efficiency in handling high-dimensional sparse text data.
- Naïve Bayes was fast but less accurate compared to other models.
- Ensemble models like Random Forest and Gradient Boosting achieved strong performance but required more computational resources.
- Proper preprocessing significantly improved the model's ability to distinguish between Fake and Real news.

---

7. Conclusion

This project demonstrates the use of **machine learning for Fake News Detection**. By applying text preprocessing, feature extraction, and training multiple classifiers, we successfully achieved high accuracy in distinguishing between fake and real news.

Key Takeaways:

- Logistic Regression with TF-IDF is the most effective approach for this dataset.
- Visualization techniques helped better understand data distribution and model performance.
- The project can be extended by using **deep learning (LSTMs, Transformers)** for better accuracy and generalization.

8. Future Work

- Incorporating metadata such as publication source and author credibility.
- Using advanced NLP models like **BERT, RoBERTa, or GPT** for contextual understanding.
- Deploying the trained model as a **web application** for real-time fake news detection.