Project Report: News Article Classification (Fake vs. Real)

Internship: Elevate Labs (AI/ML Track)
Name: Jeon Jiju

1. Introduction
This project involves classifying news articles as either real or fake using Natural Language Processing

2. Dataset
Two Excel datasets were used:
- True.csv.xlsx – Contains authentic news articles.
- Fake.csv.xlsx – Contains fabricated news articles.
These datasets were merged into a single dataset with a new label column:
- 1 → Real News
- 0 → Fake News

3. Data Preprocessing
The articles were cleaned through these steps:
- Conversion to lowercase
- Removal of punctuation and special characters
- Elimination of stopwords via NLTK
- Stemming using PorterStemmer
This prepared the text data for feature extraction and model training.

4. Feature Extraction
TF-IDF (Term Frequency–Inverse Document Frequency) was employed to convert text into numerical f
- Top 5000 frequent terms were considered.
- Feature matrix shape: (41000, 5000)

5. Model Training
The dataset was split as follows:
- 80% for training
- 20% for testing

Model: LogisticRegression()
The classifier was trained on the TF-IDF features for binary classification.

6. Evaluation
Achieved Accuracy: ~98.9%
Performance Metrics:
- Fake News (0): Precision = 0.99, Recall = 0.99, F1-Score = 0.99 (Support: 4476)
- Real News (1): Precision = 0.99, Recall = 0.99, F1-Score = 0.99 (Support: 4504)
- Overall: Precision = 0.99, Recall = 0.99, F1-Score = 0.99 (Support: 8980)

7. Example Output
Input: "The government announces a new employment policy for rural workers."
Predicted Output: REAL

8. Output Files
- news_model.pkl: Trained Logistic Regression model
- vectorizer.pkl: TF-IDF vectorizer
- True.csv.xlsx / Fake.csv.xlsx: Raw datasets
- News_Classification_Project_Output.pdf: Detailed output report
- README.md: Project summary for GitHub

9. Conclusion