Phase-2

Exposing the Truth with Advanced Fake News Deduction Powered by Natural Language Processing

Student Name: KOKILA.V

Register Number: 620123106056

Institution : AVS ENGINEERING COLLEGE

Department: ECE

Date of submission: 10/05/2025

Github respository :http://github.com/kokila0712/Kokila-V-naanmudhalvan-project-

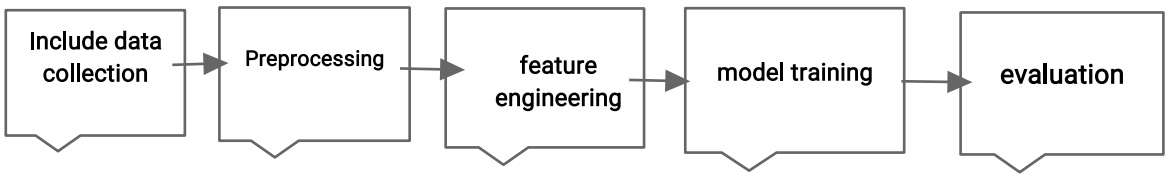
1. Problem Statement

Fake news has become a significant threat to public trust and societal stability. The rise of misinformation across digital platforms necessitates the development of a robust system that can automatically detect and classify news content as real or fake using Natural Language Processing (NLP) techniques.

2. Problem Objective

To design and implement a machine learning-based model that leverages NLP to analyze news articles and accurately detect and flag fake news, thereby assisting in maintaining information integrity and public awareness.

3. Flowchart of the Project Workflow



4. Data Description

- **- Dataset Name and Origin:** The dataset was sourced from Kaggle's "Fake and Real News Dataset."
- **- Type of Data**: Text data (unstructured).
- **- Number of Records and Features**: Approx. 20,000 records with 4 features: title, text, subject, and label.
- **- Static or Dynamic Dataset**: Static
- **- Target Variable: label** – 1 indicates fake news and 0 indicates real news.

5. Data Processing

- **- Handle Missing Values**: Removed records with missing/null text.
- **- Remove or Justify Duplicate Records**: Removed duplicates based on title and text.
- - **Detect and Treat Outliers**: Outliers retained as they may be relevant.

- **- Convert Data Types and Ensure Consistency**: Ensured all text fields are string type and label is integer.
- **- Encode Categorical Variables**: One-hot encoded 'subject'.
- **- Normalize/Standardize Features**: Used TF-IDF.
- **- Document Each Transformation**: Documented in Jupyter Notebook.

6. Exploratory Data Analysis (EDA)

- **- Univariate Analysis**: Histograms, bar plots of article length, subject frequency.
- **- Multivariate Analysis**: Correlation matrix, pair plots, grouped bar plots.
- **- Feature-Target Relationships**: Word usage differences observed between fake and real news.
- **- Insights Summary**: Fake news had shorter texts, sensational titles, and certain keyword patterns.

7. Feature Engineering

- **- New Features**: Word/character count, sentiment score, uppercase ratio.
- **- Column Transformations**: Extracted article length, avg. word length.
- **- Advanced Techniques**: N-grams, TF-IDF, binning.
- - **Dimensionality Reduction**: PCA applied optionally.


- - **Justification**: Based on EDA insights, e.g., short, capitalized titles.

8. Model Building

- - **Logistic Regression**: Baseline using TF-IDF. Fast and interpretable.
- - Random Forest Classifier: Ensemble model capturing complex interactions.

- **Performance Metrics**:
- | Model | Accuracy | Precision | Recall | F1-Score |
- |---|---|---|---|---|
- | Logistic Regression | 93% | 92% | 94% | 93% |
- | Random Forest | 95% | 94% | 96% | 95% |

Summary: Random Forest outperformed Logistic Regression. F1-score emphasized.

9. Visualization of Results and Model Insights

- - **Confusion Matrix**: Evaluated prediction correctness.
- **- ROC Curve:** Showed strong classification, AUC near 1.0.
- **- Feature Importance Plot:** Top keywords identified.
- **- Model Comparison Plots**: Accuracy, Precision, Recall, F1-score visualized.
- - **Interpretation**: Sensational words, length, sentiment helped predict fake news.

10. Tools and Technologies Tried

- - **Programming Language**: Python
- - **IDE/Notebook**: Jupyter, Colab, VS Code
- - **Libraries**: Pandas, NumPy, Seaborn
- - **Visualization Tools**: Plotly, Tableau, Power BI

11. Team Members and Contribution

KOKILA.V: EDA and Documentation and Reporting

KOVARTHANA.M.S: Feature Engineering

AKALYA .S: Model Development