Fake News Detection System Report

Prepared By:

Shaleen Dutta

Course / Subject:

Introduction to Problem solving Programming

Institution:

VIT BHOPAL

Date:24/11/25

2. Introduction

Fake news refers to false or misleading information presented as news. With the rapid spread of digital media, misinformation has become a major societal challenge. Fake news detection systems apply machine learning, natural language processing, and data analysis to automatically classify news content as real or fake. This report outlines the design, development, and evaluation of a fake news detection system.

3. Problem Statement

The rapid dissemination of misleading information through social media and online platforms poses risks to society, affecting public opinion, decisions, and trust. Manual verification is time-consuming and infeasible at scale. Thus, there is a need for an automated system capable of detecting fake news efficiently and accurately.

4. Functional Requirements

1. The system must accept news text as input.

2. The system must preprocess text (tokenization, stopword removal, stemming/lemmatization).

3. The system must classify news as "Real" or "Fake" using a trained model.

4. The system must display results to the user.

5. The system must allow dataset upload for retraining (optional).

5. Non-functional Requirements

1. Performance: High classification accuracy and fast response time.

2. Scalability: Ability to handle a large dataset for training.

3. Usability: Simple and user-friendly interface.

4. Security: Ensure dataset and user inputs remain confidential.

5. Reliability: Consistent results across multiple inputs.

6. System Architecture

Input Layer → Text Preprocessing Module → Feature Extraction (TF-IDF / Word Embeddings) → Classification Model (Logistic Regression / SVM / LSTM) → Output Layer

7. Design Decisions & Rationale

- Machine Learning Model: Logistic Regression chosen for its simplicity and high performance on text classification tasks.

- Feature Extraction: TF-IDF selected for effective vectorization of textual data.

- Dataset: A labeled dataset (e.g., LIAR or FakeNewsNet) chosen for training and evaluation.

- Architecture Choice: Modular architecture to allow easy model replacement.

8. Implementation Details

- Programming Language: Python

- Libraries Used: NumPy, Pandas, Scikit-learn, NLTK, Matplotlib

- Preprocessing steps: Tokenization, stopword removal, TF-IDF vectorization

- Model Training: Logistic Regression classifier

- Evaluation Metrics: Accuracy, Precision, Recall, F1-Score

9. Screenshots / Results

(Add application screenshots or classification results here)

10. Testing Approach

- Unit Testing for preprocessing functions

- Functional Testing for input and output flow

- Model Evaluation using test dataset

- Cross-validation for performance stability

11. Challenges Faced

- Handling imbalanced datasets

- Choosing optimal feature extraction method

- Processing large text datasets efficiently

- Ensuring high accuracy while avoiding overfitting

12. Learnings & Key Takeaways

- Gained understanding of machine learning pipelines

- Improved knowledge of text preprocessing techniques

- Understood importance of evaluation metrics

- Learned the challenges of real-world dataset handling

13. Future Enhancements

- Deep learning models like LSTMs, BERT, or RoBERTa

- Multi-language fake news detection

- Browser extension for real-time analysis

- Integration with fact-checking APIs

14. References

- Research papers on fake news detection

- Scikit-learn documentation

- NLP preprocessing guides

- Public datasets like LIAR, FakeNewsNet