FAKE NEWS DETECTION USING NLP AND MACHINE LEARNING

Project Report

1. Abstract

This project presents a complete Fake News Detection System using Natural Language Processing

(NLP), TF-IDF text vectorization, Logistic Regression and Naive Bayes models, and a deployed

Streamlit web application. The system achieves approximately 99 percent accuracy in distinguishing

real news from fake news.

2. Introduction

Fake news has become a major concern in the digital era due to its ability to influence public behavior.

This project aims to build a highly accurate classifier, deploy it through an easy-to-use interface, and

analyze its strengths, weaknesses, and performance on multiple news types.

3. Dataset Description

The project uses the Kaggle Fake and Real News dataset.

Fake.csv contains around 23,481 fabricated news articles.

True.csv contains around 21,417 verified articles from sources like Reuters.

Both datasets span politics, world news, economy, and social topics.

4. Data Preprocessing

Cleaning included lowercasing, removing URLs, removing punctuation, removing emails, tokenization,

stopword removal, and optional lemmatization.

Example:

Input: "Trump Sends Out Embarrassing New Year Tweet!"

Output: "trump send embarrass new year tweet"

5. Feature Engineering

TF-IDF vectorization was used with max_features=15000, ngram_range (1,2), and min_df=3. This

representation captures important unigrams and bigrams relevant for classification.

6. Model Development

Two models were trained:

Logistic Regression: Primary model, high accuracy, robust with sparse vectors.

Naive Bayes: Secondary model, fast, slightly lower accuracy.

7. Model Evaluation

Logistic Regression performance:

- Accuracy: 0.99
- Precision: 0.99
- Recall: 0.99
- F1-score: 0.99
- ROC AUC: 0.9985

Naive Bayes performance:

- Accuracy: 0.95
- Precision: 0.94
- Recall: 0.95
- F1-score: 0.95
- ROC AUC: 0.9867

8. Types of News Predicted and Accuracy

The model performs well across:

- Politics: 98–99 percent
- World news: 96–98 percent
- Economy: 97–99 percent
- Science/Technology: 95–98 percent
- Health: 95–98 percent
- Conspiracy topics: 99 percent or higher

Some unrealistic factual claims may still score as real due to stylistic similarity to real news writing.

9. When Accuracy Drops

- Accuracy decreases when:
- Statements are grammatically correct but factually incorrect.
- Headlines are too short.
- Topics are completely unseen or futuristic.

10. Streamlit Web Application

A full user interface allows:

- Single-article prediction.
- Batch CSV prediction.
- Top contributing words display.
- Model performance summary.
- Models required include TF-IDF vectorizer, logistic regression, optional Naive Bayes, and evaluation images or CSVs.

11. Errors Encountered and Fixes

- Version mismatch with spaCy and h5py: fixed via conda-forge reinstall.
- Missing clean_text column: fixed by adding the column through SQL.

- Streamlit not finding models: fixed through consistent folder structure.
- Unrealistic predictions documented as a limitation.

12. Limitations

- The system does not perform factual verification.
- Cannot detect sarcasm or humor.
- Dependent on dataset biases.
- Cannot verify entities or real-world facts.

13. Future Improvements

- Integration with factual verification APIs.
- Upgrade to transformer-based models like BERT or RoBERTa.
- Real-time deployment on cloud platforms.
- Expanding dataset diversity.

14. Conclusion

This project demonstrates a complete endntonend pipeline for fake news detection, covering text cleaning, feature extraction, model training, evaluation, and deployment. The system shows strong performance and provides a solid foundation for future enhancements in verifiable fact-checking.