Fake News Detection Using Information Retrieval and Machine Learning

Abstract

The rapid spread of online misinformation has created a critical need for automated fake news detection systems. This project explores how Information Retrieval (IR) techniques and machine learning (ML) models can be applied to classify news articles as *real* or *fake*. The approach begins with text preprocessing and normalization, followed by feature extraction using **TF-IDF** and **Word2Vec embeddings**. Multiple machine learning models—including Linear SVM, Random Forest, Gradient Boosting, XGBoost, and a Bi-LSTM deep learning model—were trained and evaluated. Results indicate that TF-IDF paired with XGBoost achieved the highest accuracy among classical approaches (0.868), while the Bi-LSTM model demonstrated strong performance even on limited data, suggesting scalability for larger corpora. The findings highlight the trade-offs between statistical IR pipelines and neural language models, offering insights for future transformer-based extensions.

1. Introduction

Fake news—misleading or intentionally false digital content—has become a significant societal threat, influencing public opinion, elections, and decision-making. Manual verification is slow and infeasible at scale, necessitating automated detection methods.

Information Retrieval (IR) provides the foundation for transforming raw text into searchable, analyzable structures. When combined with machine learning and deep learning models, IR creates a framework for detecting deceptive language patterns, linguistic cues, and semantic inconsistencies in news articles.

This report presents a comparative study of IR-driven machine learning and deep learning models applied to a large-scale fake news dataset.

2. Theoretical Framework

Fake news detection integrates three major domains:

2.1 Information Retrieval (IR)

IR focuses on converting unstructured text into structured representations. In this project, IR methods used include:

- Tokenization
- Stopword removal
- Stemming
- TF-IDF weighting
- N-gram modeling

These transformations enable mathematical representations of text that machine learning models can process.

2.2 Machine Learning for Text Classification

Supervised learning models such as **SVM, Decision Tree, Random Forest, Gradient Boosting, and XGBoost** were applied. These models learn patterns in term distributions and linguistic structures that differentiate fake from real news.

2.3 Deep Learning for Natural Language Understanding

A **Bidirectional LSTM (Bi-LSTM)** was implemented with pretrained **Word2Vec embeddings** to capture semantic relationships, context, tone, and writing style.

---

3. Dataset Description

Two publicly available datasets were combined:

- **WELFake Dataset**
- **Fake News Dataset (Kaggle)**

Each dataset contains labeled text samples. Labels were standardized:

Label	Meaning
0	Real news
1	Fake news

After merging and cleaning, the corpus contained a diverse set of writing styles and topics.

4. Methodology

4.1 Preprocessing

The following steps were applied:

Step	Purpose
Lowercasing	Normalize text
Tokenization	Split into words
Stopword removal	Remove high-frequency non-informative terms
Stemming	Reduce words to root form
Noise removal	Strip punctuation & non-alphabetic characters

4.2 Feature Extraction

Two representations were evaluated:

Method	Type	Dimension
TF-IDF (1–3 n-grams)	Sparse	~5000 features

Word2Vec embeddings	pretrained	Dense	300 features

TF-IDF emphasizes term importance statistically, whereas Word2Vec captures semantic similarity.

4.3 Model Training

The following models were trained:

- Linear SVM
- Decision Tree
- Random Forest
- Gradient Boosting
- XGBoost
- Word2Vec + SVM
- **Bi-LSTM with pretrained embeddings**

The Bi-LSTM was trained on a reduced dataset due to compute limitations.

---

5. Results

5.1 TF-IDF (Full Dataset)

Model	**Accuracy**	**F1 Score**
Linear SVM	0.8609	0.8654
Random Forest	0.8659	0.8714
XGBoost	**0.8683**	**0.8737**

5.2 Word Embeddings (10% Sample)

Model	**Accuracy**	**F1 Score**
Word2Vec + SVM	0.7775	0.7931
Random Forest	0.7487	0.7653
Bi-LSTM	**0.7949**	**0.8090**

6. Analysis and Discussion

TF-IDF with XGBoost achieved the strongest performance among classical approaches. The high dimensional sparse representation captured phrase-level patterns effectively, especially when combined with boosting algorithms.

The Bi-LSTM model showed **promising performance despite being trained on only 10% of the data**, demonstrating the strength of context-aware sequence models. With full-scale training and GPU optimization, Bi-LSTM or transformer-based models would likely surpass traditional pipelines.

7. Future Improvements

Future work may explore:

- Transformer architectures such as **BERT, RoBERTa, and DistilBERT**
- Explainability through **LDA or BERTopic**
- Deployment as a real-time web API or misinformation detection tool

8. Conclusion

This project demonstrates that Information Retrieval methods combined with machine learning can effectively detect misinformation. Classical approaches such as TF-IDF with XGBoost achieve strong performance with relatively low compute cost, while deep learning models like Bi-LSTM offer scalability and deeper contextual understanding.

The results reveal that the choice between approaches depends on available compute, dataset size, and deployment constraints.