Fake News Detection Using Machine Learning

Group 2

Guide: Dr Geetha S

College of Engineering Chengannur

January 6, 2026

Introduction

- Fake news spreads misinformation across politics, health, and finance.
- Manual detection is slow, subjective, and unscalable.
- Machine Learning enables automated, real-time detection.
- Our project applies ML on the ISOT Fake News dataset.

Problem Statement

The rapid spread of fake news through social media platforms causes widespread misinformation across multiple domains, and manual detection methods are inefficient, subjective, and unscalable, creating a critical need for automated machine learning–based systems to accurately distinguish between fake and real news.

Literature Review

Authors	Year	Dataset	Key Contribution
Uma Sharma, Sidarth Saran, Shankar M. Patil	2021	LIAR (12.8k statements)	Compared ML models (Logistic Regression, Naive Bayes, Decision Tree, Random Forest); LR + TF-IDF performed best; .
Biplob Kumar Sutradhar et al.	2023	Custom dataset (1.8k news items)	Evaluated ML models (NB, LR) ; Naive Bayes achieved best accuracy (56%).
Naveed Sheikh et al.	2024	8.9k news items	Compared Naive Bayes and Logistic Regression; LR achieved 98% accuracy
Vyankatesh Rampurkar & Thirupurasundari D.R.	2024	ISOT dataset	Compared NB vs LR using TF-IDF; Logistic Regression effective for fake news classification.
Oni Oluwabunmi Ayankemi et al.	2024	Kaggle news dataset (45k items)	Compared LR, Decision Tree, and Random Forest; Decision Tree gave highest accuracy (99.64%).
Akshata Deshmukh et al.	2022	Public Kaggle dataset	Evaluated multiple vectorizers (TF-IDF, Count) and classifiers (NB, LR); provided comparative performance metrics.

Literature Review

Authors	Year	Dataset	Key Contribution
Mohammad Q. Alnabhan & Paula Branco	2024	ISOT, LIAR, FakeNewsNet, CoAID	Systematic review of deep learning models (CNN, LSTM, Transformers); summarizes datasets, metrics, and research gaps.
Omar Bashaddadh et al.	2025	90 peer-reviewed studies (2020–2024)	Review of ML & DL approaches; highlights transformer-based models; discusses dataset quality and deployment challenges.
Alaa Altheneyan & Aseela Alhadlaq	2023	FNC-1 (4 categories)	Proposed distributed learning using Apache Spark and stacked ensemble; achieved F1-score 92.45%.
Akanbi Caleb et al.	2025	LIAR dataset	Compared BERT, XGBoost, and hybrid models; XGBoost achieved 73% accuracy; explored hybrid feature engineering.
Anwar V. Mbaziira	2024	Fake news generated by trolls	Proposed explainable XGBoost approach using linguistic and psycholinguistic features.
S.A. Al-Obaidi	2024	FakeNewsNet dataset	Applied XGBoost; handled dataset imbalance and improved classification performance.

Objectives

- Detect and classify fake and real news using machine learning.
- Extract key linguistic and statistical features from text data.
- Evaluate suitable ML models for accurate classification.
- Compare different model performance in the operation

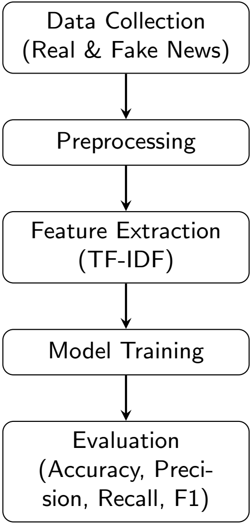
Methodology

- Adopt a supervised learning approach using labeled datasets.
- Apply NLP techniques for cleaning and transforming textual data.
- Use feature engineering to represent text in numerical form.
- Compare multiple ML algorithms to determine optimal performance.

System Architecture

1. **Data Collection** – Gather real and fake news articles.
2. **Preprocessing** – Clean text, remove stopwords, lowercase conversion.
3. **Feature Extraction** – Apply TF-IDF for vectorization.
4. **Model Training** – Logistic Regression classifier.
5. **Evaluation** – Accuracy, Precision, Recall, F1-score.

System Architecture


Data Collection (Real & Fake News)
Preprocessing
Feature Extraction (TF-IDF)
Model Training
Evaluation (Accuracy, Precision, Recall, F1)

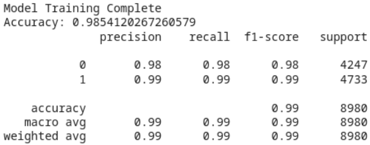
Implementation

- A supervised learning approach was used with labeled fake and real news data.
- Text data was cleaned and normalized using standard NLP preprocessing techniques.
- TF-IDF was applied to convert text into numerical feature vectors.
- The dataset was split into training and testing sets using an 80:20 ratio.
- ML models was used for efficient and interpretable classification.

Implementation Result

- The ISOT dataset was successfully processed and prepared using NLP preprocessing techniques.
- TF-IDF feature extraction effectively converted text data into numerical vectors.
- The Logistic Regression model was trained using an 80:20 train–test split.
- The trained model accurately classified fake and real news articles.
- High accuracy with balanced precision and recall was achieved on unseen test data.

Results

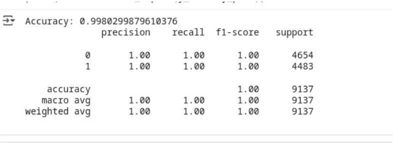


```
Model Training Complete
Accuracy: 0.9854120267260579
              precision    recall  f1-score   support

           0       0.98      0.98      0.98      4247
           1       0.99      0.99      0.99      4733

    accuracy                           0.99      8980
   macro avg       0.99      0.99      0.99      8980
weighted avg       0.99      0.99      0.99      8980
```

fig 1 Logistic Regression

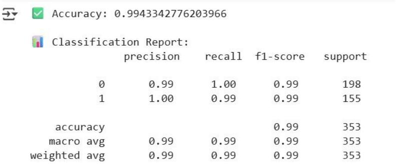


```
Accuracy: 0.9980299879610376
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      4654
           1       1.00      1.00      1.00      4483

    accuracy                           1.00      9137
   macro avg       1.00      1.00      1.00      9137
weighted avg       1.00      1.00      1.00      9137
```

fig 2 Random Forest



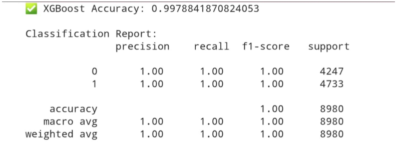
```
✅ Accuracy: 0.9943342776203966

📊 Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       198
           1       1.00      0.99      0.99       155

    accuracy                           0.99       353
   macro avg       0.99      0.99      0.99       353
weighted avg       0.99      0.99      0.99       353
```

fig 3 Support Vector Machine



```
✅ XGBoost Accuracy: 0.9978841870824053

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      4247
           1       1.00      1.00      1.00      4733

    accuracy                           1.00      8980
   macro avg       1.00      1.00      1.00      8980
weighted avg       1.00      1.00      1.00      8980
```

fig 4 XGBoost

Analysis

Table: **Model Performance Comparison for Fake News Detection**

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Test Samples**
XGBoost	0.998	1.00	1.00	1.00	8980
Random Forest (RF)	0.998	1.00	1.00	1.00	9137
SVM	0.994	0.99	0.99	0.99	353
Logistic Regression (LR)	0.985	0.99	0.99	0.99	8980

Conclusion

- All evaluated models achieved high accuracy (98% and above), demonstrating effective fake news detection.
- XGBoost and Random Forest performed the best, achieving near-perfect scores across Accuracy, Precision, Recall, and F1-Score.
- SVM and Logistic Regression also showed strong performance, making them suitable alternatives for lightweight implementations.
- Overall, ensemble-based models (XGBoost, RF) provide superior performance and robustness for real-world fake news classification tasks.

References

- U. Sharma, S. Saran, and S. M. Patil, “Fake news detection using machine learning,” *International Journal of Computer Applications*, vol. 176, no. 5, pp. 1–10, 2021.

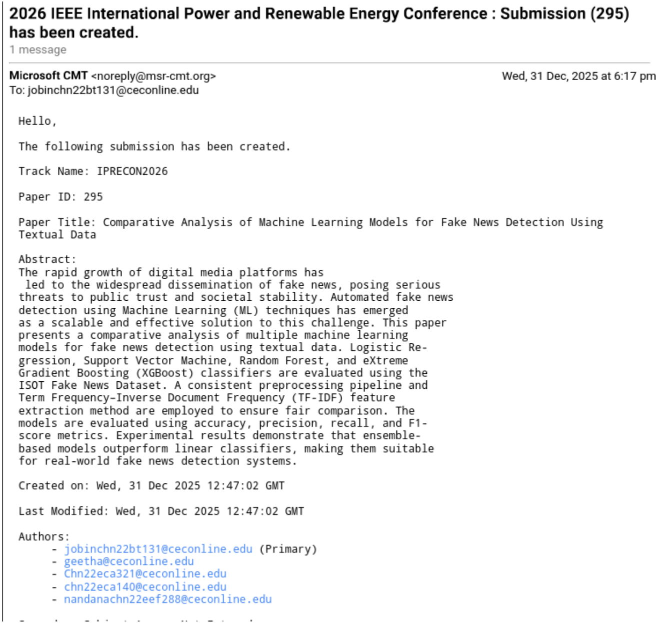
- B. K. Sutradhar, M. S. Alam, and R. Sinha, “Machine learning approach for fake news classification,” *Journal of Intelligent Systems*, vol. 32, pp. 123–134, 2023.

- N. Sheikh, M. Khan, and A. Rahman, “Comparative analysis of ML algorithms for fake news detection,” *International Conference on AI and Data Science*, pp. 45–52, 2024.

- V. Rampurkar and T. D. R., “TF-IDF and Logistic Regression for fake news detection,” *International Journal of Advanced Research*, vol. 12, no. 3, pp. 34–42, 2024.

- O. O. Ayankemi, A. L. Adetunji, and C. E. Okoro, “Evaluation of ML models for fake news classification,” *Kaggle Datasets*, 2024. [Online]. Available: `https://www.kaggle.com/datasets`

Research paper submission



2026 IEEE International Power and Renewable Energy Conference : Submission (295) has been created.

1 message

Microsoft CMT <noreply@msr-cmt.org>
Wed, 31 Dec, 2025 at 6:17 pm
To: jobinchn22bt131@ceconline.edu

Hello,

The following submission has been created.

Track Name: IPRECON2026

Paper ID: 295

Paper Title: Comparative Analysis of Machine Learning Models for Fake News Detection Using Textual Data

Abstract:
The rapid growth of digital media platforms has led to the widespread dissemination of fake news, posing serious threats to public trust and societal stability. Automated fake news detection using Machine Learning (ML) techniques has emerged as a scalable and effective solution to this challenge. This paper presents a comparative analysis of multiple machine learning models for fake news detection using textual data. Logistic Regression, Support Vector Machine, Random Forest, and eXtreme Gradient Boosting (XGBoost) classifiers are evaluated using the ISOT Fake News Dataset. A consistent preprocessing pipeline and Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction method are employed to ensure fair comparison. The models are evaluated using accuracy, precision, recall, and F1-score metrics. Experimental results demonstrate that ensemble-based models outperform linear classifiers, making them suitable for real-world fake news detection systems.

Created on: Wed, 31 Dec 2025 12:47:02 GMT

Last Modified: Wed, 31 Dec 2025 12:47:02 GMT

Authors:
- jobinchn22bt131@ceconline.edu (Primary)
- geetha@ceconline.edu
- Chn22eca321@ceconline.edu
- chn22eca140@ceconline.edu
- nandanachn22eef288@ceconline.edu

Figure: Research paper submission

External Resources

- Research paper: https://github.com/Jobn2/csd481_minor_project_CEC/tree/main
- Implementation Code: https://github.com/Jobn2/csd481_minor_project_CEC/blob/main/minor_v1.ipynb

Thank You