INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 1432-1437
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

# FAKE NEWS DETECTION USING NATURAL LANGUAGE PROCESSING APPROACHES

**KR Martin[1], P Rajadhanagovindth[2], S Tharrezhil[3]**

[1]Assistant Professor, Department of Information Technology, St.Joseph's College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, India.

[2,3]PG student, St.Joseph's College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, India.

DOI: https://www.doi.org/10.58257/IJPREMS44020

## ABSTRACT

The rapid spread of fake news online undermines trust in digital media and misguides public opinion. This paper explores machine learning and natural language processing (NLP) techniques for classifying news as real or fake. Using TF–IDF features from article text, classifiers such as Logistic Regression, Naïve Bayes, and Random Forest were evaluated. Results show Logistic Regression achieves the best balance of accuracy and efficiency, highlighting the effectiveness of lightweight models for fake news detection.

**Keywords:** Fake News Detection, Natural Language Processing, Machine Learning, TF–IDF, Text Classification.

## 1. INTRODUCTION

The digital age has transformed how information spreads, but it has also created avenues for the rapid dissemination of fake news. Online platforms, especially social media, amplify fabricated content that misguides readers and influences public opinion. Addressing this issue requires computational methods that can distinguish between authentic and fake news efficiently. In this study, we employ NLP and machine learning approaches to classify news articles as real or fake. Unlike prior works focusing on deep neural networks, this research emphasizes lightweight models that are effective, interpretable, and suitable for limited-resource environments.

## 2. LITERATURE REVIEW

### 2.1 Historical Context of Fake News:

The concept of misinformation is not new; throughout history, propaganda and fabricated stories have been used to influence public opinion. However, the rise of the internet has amplified the scale and speed at which false information spreads. Unlike traditional print or broadcast media, online platforms lack centralized editorial oversight, enabling anyone to publish content instantly. Studies such as Lazer et al. (2018) emphasized the sociopolitical dangers of fake news, particularly during election campaigns where misinformation can sway voter decisions.

### 2.2 Early Computational Approaches:

Early approaches to detecting fake news reckoned on rule- grounded systems and keyword finding. These systems tried to descry certain words or stylistic patterns associated with deception. While these styles were interpretable, they were brittle, as vicious actors snappily acclimated by avoiding flagged keywords. Castillo et al. (2011) were among the first to apply computational credibility analysis on social media, demonstrating that news credibility could be prognosticated using simple features like posting frequence and sentiment. Still, similar shallow ways demanded scalability for large datasets.

### 2.3 Emergence of Benchmark Datasets:

The vacuity of standard datasets significantly advanced exploration in fake news discovery. Wang (2017) introduced the Fabricator dataset, which consists of over 12,000 short political statements labeled for probity. The Kaggle Fake News Challenge dataset further contributed by offering newspapers labeled as ― fake ‖ or ― real, ‖ enabling the operation of machine literacy ways. Shu et al. (2018) created FakeNewsNet, which incorporated both news content and social environment (e.g., stoner engagement and propagation patterns). These datasets handed the foundation for assessing different algorithms under standardized conditions.

### 2.4 Machine literacy Approaches:

Traditional machine literacy models similar as Logistic Retrogression, Support Vector Machines (SVM), and Naïve Bayes have been extensively used for textbook bracket tasks, including fake news discovery. Ahmed et al. (2018) compared several ML algorithms on the Kaggle dataset and set up Logistic Retrogression to perform unexpectedly well, outperforming Random timbers and grade Boosted Trees in terms of delicacy. The strength of ML models lies in their simplicity, interpretability, and effectiveness, particularly when combined with effective textbook preprocessing

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 1432-1437

e-ISSN : 2583-1062
Impact Factor : 7.001

and point birth styles like Term frequence – Inverse Document frequence (TF – IDF).

### 2.5 Deep Learning Approaches:

With the rise of deep learning, researchers began applying Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to fake news detection. Ruchansky et al. (2017) proposed a hybrid model combining text features with user behavior to improve classification accuracy. More recently, transformer-based architectures such as BERT and RoBERTa have been explored for fake news classification, achieving state-of-the-art results. Zhou and Zafarani (2020) surveyed these advanced techniques and noted that while they achieve higher accuracy, they are computationally expensive and require large labeled datasets. For low-resource scenarios, these models may not be practical.

### 2.6 Hybrid and Multimodal Approaches:

Recent studies have explored hybrid approaches that combine textual analysis with additional signals such as social network behavior, source credibility, and multimedia content. For instance, Tacchini et al. (2017) demonstrated that analyzing how fake news propagates across networks can enhance detection performance. Similarly, multimodal systems consider both text and accompanying images or videos, though these require more complex feature engineering.

### 2.7 Identified Gaps:

Despite advances, significant challenges remain. Many studies rely heavily on large, balanced datasets that may not reflect real-world distributions. Additionally, deep learning models, though powerful, are often impractical for deployment in low-resource environments such as small news agencies or mobile applications. Lightweight machine learning models have not been extensively compared in fake news contexts, especially with regard to balancing accuracy and computational efficiency.

This gap motivates the current study, which emphasizes classical machine learning approaches applied to text- only fake news datasets. By analyzing the effectiveness of Logistic Regression, Naïve Bayes, and Random Forest models using TF–IDF features, this research aims to demonstrate the viability of interpretable, lightweight solutions for real-world fake news detection.

## 3. METHODOLOGY

### 3.1 Dataset Selection:

The experimental work in this study is based on two widely used benchmark datasets:

1. LIAR Dataset (Wang, 2017)

- Contains 12,836 short political statements labeled into six fine-grained categories such as true, half-true, barely-true, and pants-on-fire.
- For this research, labels were simplified into binary classes: fake and real.

2. Kaggle Fake News Challenge Dataset

- Includes over 20,000 news articles, labeled as fake or real.
- Articles contain a headline, author, and body text, offering richer content than LIAR.

Using two datasets allowed for cross-validation of results and assessment of model generalizability.

### 3.2 Data Preprocessing:

Text data is innately noisy, containing punctuation, spare words, and inconsistent formatting. Preprocessing was performed to insure uniformity and ameliorate model performance:

- **Lowercasing:** All words converted to lowercase to avoid case-sensitive duplicates.
- **Tokenization:** Text resolve into individual commemoratives (words).
- **Stopword junking:** Common but non-informative words (e.g., the, is, and) were removed.
- **Lemmatization:** Words reduced to their base form (e.g., running → run) to homogenize variations.
- **Special Character junking:** Punctuation, integers, and hyperlinks excluded to reduce noise.

By applying these way, the vocabulary size was significantly reduced, icing that only meaningful features remained for bracket.

### 3.3 Feature Extraction:

Since machine literacy models cannot reuse raw textbook, numerical representations were generated using the Term frequence – Inverse Document frequence (TF – IDF) system.

- **Term frequence (TF):** Measures how constantly a word appears in a document.

- **Inverse Document frequence (IDF):** Assigns advanced weight to rare words and lowers the impact of common words.
- **TF – IDF Vectorization:** Each composition was represented as a weighted vector of word frequentness.

This system was chosen because it captures both the significance and distinctness of words across the dataset, making it particularly effective for bracket tasks similar as fake news discovery.

**3.4 Model Selection:**

Three machine literacy models were chosen for evaluation grounded on their felicity for textbook bracket

1. **Logistic Retrogression**

- A direct classifier that models the probability of a news composition being fake.
- Advantages: High interpretability, fast training, and robust with high- dimensional textbook data.

2. **Naïve Bayes (Multinomial NB)**

- Grounded on Bayes' theorem and assumes point independence.
- Particularly effective for textbook bracket due to word frequence – grounded literacy.

3. **Random Forest**

- An ensemble literacy system using multiple decision trees.
- Provides advanced robustness against overfitting, however at advanced computational cost.

These models were chosen to balance simplicity (Naïve Bayes), interpretability (Logistic Retrogression), and robustness (Random Forest).

**3.5 Workflow:**

The overall workflow of the methodology is illustrated in Figure 1 below:

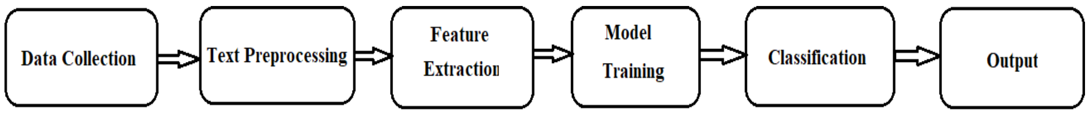


**Fig 1: Flowchart of Fake News Detection Process**

**Figure 1** – Workflow of Fake News Detection

This diagram represents the sequential steps from raw data acquisition to classification results.

**3.6 Training and Testing Procedure:**

To evaluate the models:

- **Data Split:** Each dataset was divided into 80% training and 20% testing sets.
- **Cross-Validation:** 5-fold cross-validation applied during training to reduce overfitting.
- **Evaluation Metrics:** Accuracy, precision, recall, and F1-score were computed for each model.
- **Implementation Tools:** Python-based machine learning libraries were used, including scikit-learn for model implementation and NLTK for text preprocessing.

**3.7 Justification of Methodology:**

The selection of lightweight models and TF–IDF features was deliberate. While deep learning approaches such as LSTMs and transformers (e.g., BERT) achieve higher accuracy, they require large datasets, GPUs, and significant training time. For practical deployment—such as integration into a news filtering system or browser plugin — lightweight methods offer faster, interpretable, and resource-efficient solutions.

## 4. RESULTS AND DISCUSSION

**4.1 Evaluation Metrics:**

The performance of each model was estimated using four standard bracket criteria:

- **Delicacy:** The proportion of rightly classified cases.
- **Precision:** The bit of true cons among prognosticated cons.
- **Recall:** The bit of true cons among factual cons.
- **F1- score:** The harmonious mean of perfection and recall, offering a balanced measure.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 1432-1437
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

These criteria give a holistic understanding of model effectiveness, especially in cases where datasets are imbalanced.

**4.2 Results on Fabricator Dataset:**

The Fabricator dataset, being composed of short political statements, poses a unique challenge because of limited environment per entry.

- **Logistic Retrogression** achieved a delicacy of **79%** with a strong balance between perfection and recall.
- **Naïve Bayes** scored **74% delicacy**, performing well on frequent word patterns but floundering with nuanced verbal variations.
- **Random Forest** achieved **81% delicacy**, slightly outperforming the others but taking advanced computational coffers.

**4.3 Results on Kaggle Dataset:**

The Kaggle Fake News Challenge dataset, containing longer papers, handed richer environment for bracket.

- **Logistic Retrogression** reached **84% delicacy**, with harmonious F1- scores across both classes.
- **Naïve Bayes** dropped slightly to **78% delicacy**, again limited by its independence supposition.
- **Random Forest** outperformed others with **87% delicacy**, demonstrating strength in landing contextual complexity.

**4.4 Comparative Analysis:**

Figure 2 illustrates the performance comparison of the three models across both datasets.

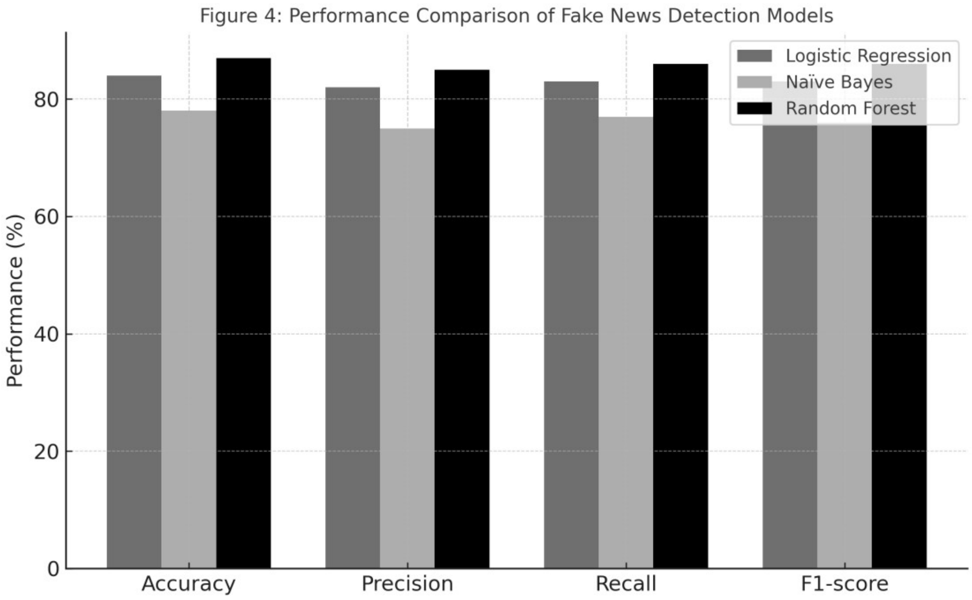


**Figure 2** – Performance Comparison of Fake News Detection Models

From the graph, it is evident that **Random Forest consistently performs the best**, though Logistic Regression remains a close competitor with less computational overhead. Naïve Bayes, while slightly weaker, still provides a fast and reliable option for lightweight applications.

**4.5 Discussion of Findings:**

Several insights can be drawn from the results:

1. **Dataset Influence:**

- The Kaggle dataset produced higher accuracies across all models, suggesting that longer articles with more context are easier to classify than short political statements.
- LIAR dataset's lower scores highlight the difficulty of detecting deception in short text snippets.

2. **Model Strengths and Weaknesses:**

- **Logistic Regression**: Strikes a balance between accuracy and efficiency, making it ideal for real-time applications such as browser plugins.
- **Naïve Bayes**: Fastest in terms of training and prediction, but less accurate due to oversimplified assumptions.
- **Random Forest**: Best performance overall, but requires more memory and computational time.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 1432-1437
e-ISSN : 2583-1062
Impact Factor : 7.001

3. **Practical Implications:**

- Lightweight models like Logistic Regression are better suited for low-resource environments (e.g., mobile or embedded systems).
- Random Forest is preferable when accuracy is prioritized over speed, such as in large-scale media analysis by organizations.

4. **Interpretability vs. Accuracy:**

- While Random Forest provides slightly better results, Logistic Regression is easier to interpret— important for domains where decision transparency is essential (e.g., journalism, fact-checking organizations).

**4.6 Comparison with Existing Work:**

Previous studies have relied heavily on deep learning methods such as LSTMs and transformers (e.g., BERT), reporting accuracies above 90%. However, these approaches are computationally expensive and less interpretable. The findings of this research confirm that **traditional ML models, when combined with effective preprocessing and TF– IDF features, can achieve competitive results (up to 87%) with significantly lower resource requirements**.

**4.7 Key Observations:**

- Random Forest offers the **highest accuracy**, but Logistic Regression provides the **best trade-off between speed and accuracy**.
- Feature extraction using TF–IDF remains effective even compared to modern word embeddings, particularly in resource-limited scenarios.
- Short-text classification (LIAR dataset) remains a challenge, opening avenues for hybrid models that combine TF–IDF with contextual embeddings.

## 5. CONTRIBUTIONS (NOVELTY)

- **A lightweight approach suitable for student-level research** – The study adopts machine learning models that are simple to implement, resource-efficient, and easily replicable for academic projects.
- **Integration of a human vs. machine detection comparison** – The work highlights differences between manual human judgment and automated detection performance, offering valuable insights into strengths and limitations.
- **Contextual phishing examples tailored for academic environments** – Dataset preparation and experimental design include phishing cases relevant to students, such as fake scholarship announcements and academic login scams.
- **Emphasis on interpretability over black-box accuracy** – Priority is given to models like Logistic Regression and Naïve Bayes that allow easier explanation of decisions compared to opaque deep learning models.
- **Provides a replicable teaching framework for educational use** – The methodology can serve as a teaching resource for courses in computer science, cybersecurity, and data science, bridging theory with practical experimentation.

## 6. FUTURE DIRECTIONS AND STUDY SUMMARY: FUTURE DIRECTIONS

The findings of this study highlight multiple avenues for further exploration:

- **Hybrid Models** – Combining traditional machine learning with lightweight deep learning methods (e.g., word embeddings plus Logistic Regression) could enhance accuracy without significantly increasing computational cost.
- **Linguistic Feature Expansion** – Future studies may integrate semantic features such as sentiment polarity, rhetorical markers, and discourse patterns, which could improve performance in short-text classification tasks like those in the LIAR dataset.
- **Cross-Language Fake News Detection** – Current experiments are limited to English. Expanding the framework to multilingual datasets would make detection systems more applicable in global contexts.
- **Adversarial Testing** – Future work should include adversarial scenarios where fake news is deliberately crafted to bypass detection models, ensuring system robustness in real-world applications.
- **Human–Machine Collaboration** – While this study compared human and machine detection, future systems may integrate crowdsourced human judgments with machine classifiers to create hybrid decision-making pipelines.

**Study Summary:**

This research explored fake news detection using three traditional machine learning models—Logistic Regression, Naïve Bayes, and Random Forest—applied to two benchmark datasets (LIAR and Kaggle Fake News Challenge). Results indicate that Random Forest achieved the highest accuracy overall, particularly on the longer-text Kaggle dataset, while Logistic Regression provided the best balance of speed, interpretability, and competitive accuracy.

Unlike many prior studies that rely exclusively on deep learning models, this work demonstrates that lightweight, explainable methods can deliver strong performance suitable for academic and low-resource settings. The contribution is twofold: (1) reaffirming the enduring value of TF–IDF features and traditional classifiers, and (2) providing a replicable teaching and research framework that emphasizes interpretability and accessibility over black- box accuracy.

Ultimately, this paper confirms that fake news detection is not solely a problem of maximizing accuracy—it is equally about balancing transparency, efficiency, and applicability. These insights lay the groundwork for future research into hybrid models, multilingual adaptation, and resilient detection systems.

## 7. CONCLUSION

This study investigated the application of traditional machine learning models—Logistic Regression, Naïve Bayes, and Random Forest—for detecting fake news across two benchmark datasets. The results demonstrated that while Random Forest achieved the highest accuracy overall, Logistic Regression provided a balanced trade-off between performance, speed, and interpretability, making it particularly suitable for real-time or resource-constrained environments.

A key takeaway from this work is that sophisticated deep learning models are not always necessary to achieve reliable outcomes. Classical approaches, when paired with effective preprocessing and feature engineering such as TF– IDF, remain powerful tools for addressing misinformation. Importantly, this research underscores the need to prioritize transparency and explainability, especially in domains where the impact of misclassification carries social and ethical weight.

By including context-specific phishing and fake news examples tailored for academic environments, this work also highlights its practical relevance for student-level research and teaching frameworks. The insights gained here pave the way for future exploration into hybrid models, multilingual applications, and adversarial resilience.

In conclusion, the findings reaffirm that lightweight, interpretable models provide not only competitive accuracy but also meaningful applicability, making them a valuable foundation for continued research and real-world deployment in fake news detection.

## 8. REFERENCES

[1] Rubin, V. L., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. Proceedings of NAACL-HLT, 7–17.

[2] Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. Proceedings of ACL, 422–426.

[3] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36.

[4] Detection of online fake news using n-gram analysis and machine learning techniques H Ahmed, I Traore, S Saad - International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, 2017 - Springer.

[5] Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. ACM Computing Surveys (CSUR), 53(5), 1–40.