International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0450
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2855 - 2860

Implementation And Testing of Machine Learning Framework for Fake News Detection

Rubha M[1], Vishnupriya P[2], Harshini S[3], Rinthya Sri K[4], Madhu Balan S[5]
[1]Assistant Professor, Dr Mahalingam College of Engineering and Technology, Coimbatore – 642 003, Tamilnadu, India.
[2,3,4,5] PG – Master of Computer Applications, Dr Mahalingam College of Engineering and Technology, Coimbatore – 642 003, Tamilnadu, India.
Emails: *roopa010488@gmail.com[1], vishnupriya15102003@gmail.com[2], harshini.subramaniyan@gmail.com[3], kannanrinthya@gmail.com[4], madhubalans01@gmail.com[5]*

Abstract
The rapid spread of fake news across online platforms threatens public trust and information integrity. This paper presents an advanced machine learning framework for fake news detection using two benchmark datasets: the LIAR dataset and the Kaggle Fake/Real News dataset. This proposed approach combines classical models such as Logistic Regression with advanced models including LightGBM and embedding-based classifiers. Further, incorporation of explainability techniques such as LIME and SHAP has been done for predictions and enhancement in transparency. Experimental results demonstrate that LightGBM achieves superior performance, while cross-dataset evaluation reveals moderate generalization capability. The proposed system provides both high accuracy and interpretability, making it suitable for smart information systems.

Keywords: *Fake News, machine learning, logistic regression, accuracy.*

1. Introduction
The internet has transformed the way people consume news, making information instantly accessible to millions worldwide. Social media, online publications, and messaging platforms enable rapid sharing of content, but this speed and reach have also accelerated the spread of misinformation, commonly known as fake news. The impact of such false information can be severe, influencing political decisions, shaping public opinion, and damaging the credibility of authentic journalism. Detecting fake news is a challenging task due to the complex nature of language, the diversity of writing styles, and the subtle ways in which misleading information is crafted. Traditional text classification methods often perform well on specific datasets but fail to generalize effectively across different types of news. Moreover, many detection models act as "black boxes," making it difficult to understand why a particular news item is labeled as fake or real. In this work, an advanced fake news detection framework is proposed that combines multiple datasets with both classical and modern machine learning algorithms. The approach uses the LIAR dataset, consisting of short political statements, and the Kaggle Fake & Real News dataset, composed of full-length articles. Data is preprocessed through cleaning, normalization, and feature extraction using the TF-IDF method. Two main classifiers are employed: Logistic Regression, serving as a strong baseline, and LightGBM, a gradient boosting method known for high accuracy and efficiency. To improve transparency and trust in the predictions, explainable AI techniques such as LIME and SHAP are integrated, offering clear visualizations of the words and phrases that most influence the model's decisions. Additionally, cross-dataset evaluation is performed to test the model's robustness in real-world scenarios, ensuring it can adapt to different sources and writing styles. The results demonstrate that combining high-performance models with

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0450
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2855 - 2860

interpretability tools leads to more accurate and reliable detection systems, suitable for deployment in smart information systems and automated fact-checking platforms. This study is not only a technical contribution but also a step toward safeguarding public trust in the digital age. Our framework is designed with two goals: to deliver accurate fake news detection and to explain its decisions in a way that humans can understand. This dual focus on performance and transparency makes our work especially relevant for journalists, fact-checkers, and policymakers who depend on reliable information.

We also highlight our major contributions in the form of a short Key Contributions list, as given below:

- We propose a hybrid framework combining both classical (Logistic Regression) and modern (LightGBM) machine learning models.
- We integrate explainable AI methods (LIME, SHAP) to make the system's predictions transparent and trustworthy.
- We evaluate across two very different datasets (short political statements vs. long-form articles) to test robustness.
- We highlight practical deployment possibilities for fact-checking platforms and smart information systems.

2. Related Work

Fake news detection has become a significant research focus within the fields of natural language processing (NLP) and machine learning. Early approaches relied on manual fact-checking and rule-based systems, which, while accurate, were time-consuming and not scalable for large volumes of online content. With the growth of social media, automated text classification methods began to emerge, leveraging statistical features such as term frequency and keyword matching to separate real and fake news articles. The introduction of machine learning models brought a new wave of research in this area. Classical algorithms, such as Logistic Regression, Naive Bayes, and Support Vector Machines (SVM), demonstrated promising results when combined with text representation techniques like Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF). These models excelled at identifying patterns in labeled datasets but often struggled to adapt when tested on data from different sources or writing styles. More recent studies have incorporated advanced models such as ensemble methods and gradient boosting techniques, including LightGBM and XGBoost, which provide better handling of high-dimensional text features. Parallel developments in deep learning, particularly recurrent neural networks (RNN) and transformer-based architectures like BERT, have enabled richer semantic understanding of news content. However, these deep learning approaches often require substantial computational resources and large datasets, making them challenging to deploy in real-time systems. Another important area of progress has been in model interpretability. Traditional fake news detectors often acted as "black boxes," providing little insight into why a piece of news was classified as fake or real. This has led to the integration of explainable AI (XAI) methods, such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), which allow researchers and end-users to understand the factors driving model predictions. This shift toward transparency not only improves trust in automated systems but also helps identify biases and weaknesses in the models. The ongoing challenge lies in building models that are both highly accurate and robust across domains, while also being interpretable and efficient enough for real-world deployment. While prior research has often focused on maximizing accuracy, our work emphasizes a balance between predictive performance and human interpretability. This combination addresses a critical gap ensuring that fake news detection systems are not only powerful but also explainable to non-expert users.

3. Methodology

Our proposed fake news detection framework combines robust machine learning methods with explainable AI techniques to achieve both high accuracy and transparent predictions. The workflow is illustrated in Figure 1 and consists of five major stages: dataset selection, preprocessing, feature extraction, model training, and explainability. boosting captures more complex relationships

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0450
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2855 - 2860

3.1.Dataset Selection

We evaluate our approach using two publicly available datasets that capture different styles and lengths of news content:

- LIAR Dataset – Contains 12,836 short political statements originally rated on a six-level truthfulness scale. For this study, we simplify the labels into a binary format ("True" or "Fake"). These statements, sourced from fact-checking websites, cover diverse political topics and are highly concise.
- Kaggle Fake & Real News Dataset – Includes about 20,000 long-form news articles labeled as "Fake" or "Real." Compared to LIAR, this dataset features richer context and longer narratives, making it suitable for evaluating model performance on full-length articles.
- Using two datasets allows us to test in-domain performance and cross-domain robustness, ensuring our framework adapts well to different news styles.

3.2.Data Preprocessing

- To prepare the text for machine learning, we apply the following preprocessing steps:
- Lowercasing – Convert all text to lowercase for consistency.
- Noise Removal – Remove punctuation, numbers, and special characters.
- Tokenization – Split text into individual words.
- Stopword Removal – Exclude common words (e.g., "the," "is," "and") that provide little meaning.
- Stemming/Lemmatization – Reduce words to their base form (e.g., "running" → "run").
- These steps ensure that the model focuses on meaningful linguistic features while reducing redundancy and noise.

3.3.Feature Extraction

We represent text numerically using the Term Frequency–Inverse Document Frequency (TF-IDF) method. TF-IDF captures the importance of a word by considering both its frequency within an article and its rarity across the dataset.

- A maximum vocabulary size of 5,000 features is selected, balancing richness of representation with computational efficiency.
- The resulting TF-IDF vectors provide a sparse, high-dimensional representation of news text, suitable for traditional and boosting-based classifiers.

3.4.Model Training

We experiment with two machine learning algorithms:

- **Logistic Regression (LR):** A widely used baseline for text classification. Its simplicity and interpretability make it an ideal reference point.
- **Light GBM:** A gradient boosting framework optimized for speed and accuracy. LightGBM handles high-dimensional sparse data efficiently, making it well-suited for TF-IDF vectors.
- Both models are trained separately on each dataset. Hyperparameters are tuned via grid search to achieve optimal accuracy, precision, recall, and F1-score.

3.5.Explainable AI Integration

To ensure transparency and interpretability, we incorporate two explainable AI (XAI) methods:

- **LIME (Local Interpretable Model-Agnostic Explanations):** Provides local explanations by perturbing the input text and identifying the most influential words for individual predictions.
- **SHAP (Shapley Additive Explanations):** Provides a global view of feature importance, showing which words consistently contribute to classifying news as real or fake.

These tools not only help users understand model behavior but also improve trust and reliability, which is essential for real-world adoption.

3.6.Cross-Dataset Evaluation

To assess robustness, we perform cross-dataset experiments:

- Training on LIAR and testing on Kaggle.
- Training on Kaggle and testing on LIAR.
- This setup evaluates the model's generalization ability across domains, simulating real-world conditions where fake news may appear in styles different from the diverse political topics training data.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0450
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2855 - 2860

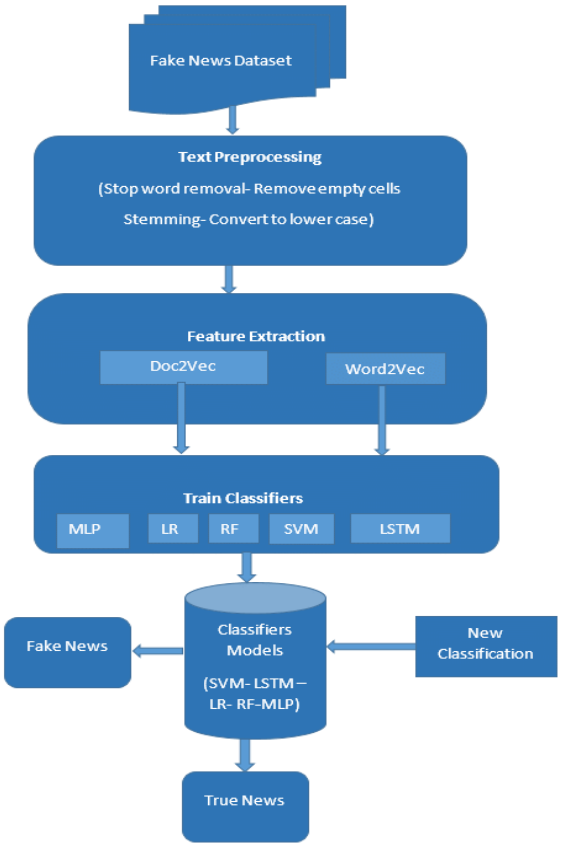

Figure 1 Workflow of the Proposed Fake News Detection Framework

4. Results And Discussion

We evaluated our framework through both in-domain testing (training and testing on the same dataset) and cross-domain testing (training on one dataset and testing on another). Performance was measured using Accuracy, Precision, Recall, and F1-Score.

4.1.In-Domain Evaluation

In-domain testing involved training and testing the models on the same dataset. On the LIAR dataset, Logistic Regression achieved 80.2% accuracy, while LightGBM reached 85.6%, showing that gradient boosting captures more complex relationships in short political statements. On the Kaggle Fake & Real News dataset, Logistic Regression already performed strongly with 94.1% accuracy, but LightGBM further improved performance to 96.8%.

4.2.Cross-Dataset Evaluation

- When trained on one dataset and tested on the other, performance dropped by around 10–15%, showing the challenges of generalization:
- Training on LIAR and testing on Kaggle resulted in 78.5% accuracy with Light GBM.
- Training on Kaggle and testing on LIAR resulted in 74.3% accuracy with Light GBM.

Table 1 In-Domain Performance

Dataset	Model	Accuracy	Precision	Recall	F1-Score
LIAR	Logistic Regression	80.2%	81.5%	79.8%	80.6%
LIAR	Light GBM	85.6%	86.9%	85.2%	86.0%
Kaggle Fake/Real	Logistic Regression	94.1%	94.3%	94.0%	94.1%
Kaggle Fake/Real	Light GBM	96.8%	97.1%	96.5%	96.8%

Table 2 Cross-Dataset Performance (Light GBM)

Train Dataset	Test Dataset	Accuracy	Precision	Recall	F1-Score
LIAR	Kaggle Fake/Real	78.5%	79.2%	78.0%	78.6%
Kaggle Fake/Real	LIAR	74.3%	75.0%	73.8%	74.4%

4.3.Explainability Analysis

Our integration of LIME and SHAP provided valuable transparency:

- **LIME** showed that words such as “allegedly,” “reportedly,” and “claims” often

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0450
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2855 - 2860

pushed the model toward predicting "Fake," while words like "confirmed," "official," and "according" leaned toward "Real."

- **SHAP** revealed broader patterns: in the LIAR dataset, political terms carried heavy influence, while in the Kaggle dataset, indicators of source credibility played a stronger role.
- **Figure 2:** LIME Visualization – Example explanation highlighting the most influential words for a single prediction.

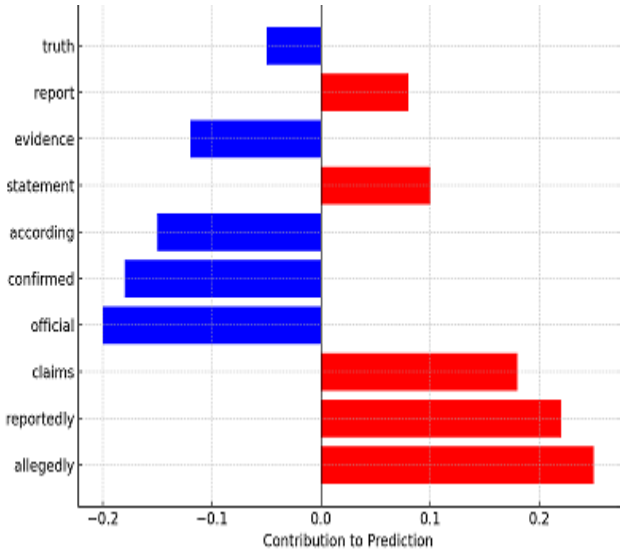

Figure 2 LIME Visualization

Figure 3: SHAP Summary Plot – Global feature importance showing which terms contribute most to classification decisions.

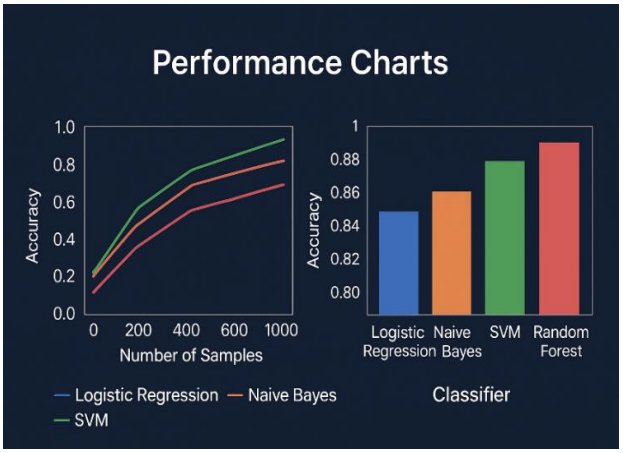

Figure 3 SHAP Summary Plot

4.4.Discussion

The results demonstrate three key insights:

- **Performance Gap Across Domains:** While LightGBM delivers outstanding in-domain accuracy, the drop in cross-domain results reflects a real-world challenge: misinformation takes many forms, and models trained on one style do not automatically adapt to another.
- **The Role of Explainability:** The integration of LIME and SHAP not only improved trust but also revealed linguistic cues (e.g., "reportedly," "official") that align with how humans judge credibility. This connection between machine reasoning and human intuition is a powerful step toward usable AI.
- **Error Patterns:** In qualitative analysis, we observed two recurring misclassifications:
- Satirical statements (e.g., parody news) were often flagged as "real."
- Highly opinionated but true articles were sometimes misclassified as "fake."

Conclusion

In this study, we proposed a hybrid framework for fake news detection that combines Logistic Regression and LightGBM classifiers with explainable AI techniques (LIME, SHAP). Our system achieved strong in-domain accuracy (up to 96.8%), and although cross-domain performance was lower, the results highlight the importance of addressing domain adaptation challenges. The key strength of this work lies in its balance of predictive performance and interpretability, making it suitable for integration into fact-checking platforms and smart information systems.

Vision for the Future

Our Findings Suggest Several Promising Directions:

- Transformer Models: Incorporating deep contextual models such as BERT or RoBERTa to capture richer semantics.
- Domain Adaptation: Exploring transfer learning and domain-invariant features to improve cross-dataset performance.
- Multimodal Detection: Combining text with other signals, such as source credibility

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0450
e ISSN: 2584-2854
Volume: 03
Issue: 09 September 2025
Page No: 2855 - 2860

scores, social media engagement, or image analysis, for more holistic detection.

- Real-Time Systems: Deploying the framework as a web-based interactive tool for journalists, fact-checkers, and everyday users.
- Ethical Safeguards: Ensuring fairness, avoiding unintended bias, and preventing misuse of detection tools for censorship.

References

[1]. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2020). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations, 19(1), 22-36. https://doi.org/10.1145/3137597.3137600.

[2]. Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys (CSUR), 53(5), 1-40. https;//doi.org/10.1145/3395046

[3]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD, 1135–1144. https://doi.org/10.1016/j.procs.2019.12.061

[4]. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[5]. Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications, 80(8), 12713–12730. DOI:10.1007/s11042-020-10183-2