IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14429

Comparative Analysis and Cross-Dataset Validation of Traditional Machine Learning Models for Fake News Detection

Pranjal L. Bhamre[1], Isha Y. Jain[2], Sunita N. Deore[3]

PG students, Department of Computer Science and Applications, K.T.H.M. College, Nashik[1,2]

Assistant Professor, MVP Samaj's S.V.K.T.College, Deolali Camp, Nashik[3]

Abstract: The viral spread of digital misinformation, often called an "infodemic," has moved beyond a technical nuisance to become a genuine threat to public health and social stability. While modern research favors Deep Learning models such as BERT, these architectures often demand hardware resources that are not practical for real-time, decentralized deployment. This study shifts the focus to computational efficiency by evaluating four traditional machine learning classifiers—Logistic Regression, Support Vector Machine (SVM), Multinomial Naïve Bayes, and the Passive-Aggressive Classifier (PAC)—on a corpus of 39,103 news articles. By combining a Regex-based preprocessing pipeline with optimized TF-IDF vectorization, the proposed framework achieved a peak in-domain accuracy of 0.995 using PAC. However, cross-dataset validation on the LIAR benchmark dataset revealed a performance decline to 0.474, primarily due to contextual sparsity in short-form political statements. These findings suggest that while traditional models are highly effective for long-form news classification, they require semantic enhancement to handle sparse social media content. Overall, this work supports a sustainable "Green AI" perspective that emphasizes computational efficiency while acknowledging cross-domain limitations.

Keywords—Fake News Detection; Traditional Machine Learning; Passive-Aggressive Classifier; TF- IDF; Cross-Dataset Validation; Domain Shift; Text Classification

I. INTRODUCTION

The advent of the World Wide Web and the widespread adoption of social media platforms have fundamentally transformed how information is created and disseminated. While this democratization of communication has enabled rapid information exchange, it has also facilitated the large-scale propagation of misleading and fabricated content commonly referred to as "fake news." Digital misinformation has emerged as a significant threat to democratic processes, public discourse, and public health, particularly during global crises such as the COVID-19 pandemic [1]. The rapid dissemination of misinformation across platforms such as X (formerly Twitter), Facebook, and WhatsApp limits the effectiveness of manual fact-checking mechanisms. Human verification processes are inherently resource-intensive and struggle to scale against the continuous generation of user content. Prior studies emphasize the necessity of automated, algorithmic approaches capable of operating efficiently in large-scale digital ecosystems [2]. In densely connected environments, misleading political and social narratives can amplify quickly, reinforcing the demand for real-time detection systems.

Research in fake news detection has broadly evolved along two methodological directions: Traditional Machine Learning (ML) and Deep Learning (DL). Deep learning architectures—including Convolutional Neural Networks (CNNs) and Transformer-based models such as BERT—demonstrate strong performance by capturing contextual and semantic relationships within text [5], [8]. However, these architectures often require substantial computational resources, including GPU acceleration and high memory capacity, which may restrict their deployment in resource-constrained environments. These limitations motivate renewed investigation into linear classifiers supported by effective feature engineering techniques. Term Frequency–Inverse Document Frequency (TF- IDF) representations remain a robust method for transforming unstructured text into discriminative high-dimensional feature spaces suitable for classification tasks. With rigorous preprocessing and optimized feature extraction, traditional linear models can potentially achieve competitive in-domain performance while maintaining significantly lower computational overhead.

An additional challenge in misinformation detection concerns cross-domain robustness. While models frequently achieve strong performance on long-form news articles, performance may degrade when evaluated on short-form political statements due to contextual sparsity and reduced lexical density. Accordingly, this study presents a systematic benchmarking of four traditional classifiers—Logistic Regression, Linear Support Vector Machine (SVM), Multinomial Naïve Bayes (MNB), and the Passive-Aggressive Classifier (PAC). The models are evaluated in- domain on the high-density ISOT dataset and subsequently assessed through cross-dataset validation using the LIAR benchmark to examine

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14429

generalization under domain shift. This dual evaluation framework provides empirical insight into the dataset dependency and cross-domain behavior of TF-IDF-based linear models for fake news detection.

II. LITERATURE REVIEW

The field of automated fake news detection has evolved significantly, transitioning from rule-based linguistic heuristics to high-dimensional vector-space modeling and deep neural architectures. Although recent research has increasingly emphasized Deep Learning (DL) approaches, traditional Machine Learning (ML) models—particularly linear classifiers combined with effective feature engineering—continue to demonstrate strong performance in text classification tasks. This review synthesizes prior research across three thematic areas: feature extraction methodologies, comparative performance of linear and probabilistic classifiers, and the computational trade-offs between traditional ML and Deep Learning. The transformation of unstructured text into numerical representations is a foundational challenge in misinformation detection. Early approaches relied heavily on Bag-of-Words (BoW) and n-gram representations [1], which provided baseline performance but were limited by their inability to effectively down-weight high-frequency but semantically weak terms. To address this limitation, researchers increasingly adopted Term Frequency–Inverse Document Frequency (TF-IDF), which assigns importance to tokens based on corpus-level distinctiveness. Comparative studies have demonstrated that TF-IDF, when combined with linear classifiers, can outperform more complex embedding techniques such as Word2Vec in certain classification settings [2]. Similarly, implementations using TF-IDF on publicly available benchmark datasets have reported strong performance, exceeding 90% accuracy in several evaluations [3]. Additionally, prior work emphasizes the importance of rigorous preprocessing, including normalization and removal of non-alphabetic artifacts, to enhance feature quality and reduce noise [4]. These findings suggest that careful preprocessing plays a critical role in maximizing the linear separability of textual representations. Building upon this insight, the present study incorporates a Regex-based cleaning pipeline to strengthen the discriminative power of the TF-IDF feature space, particularly when evaluating performance across datasets of varying textual density.

The literature highlights a consistent performance gap between probabilistic classifiers, such as Multinomial Naïve Bayes (MNB), and geometric linear classifiers including Support Vector Machines (SVM) and Passive-Aggressive Classifiers (PAC). While Naïve Bayes serves as a computationally efficient baseline, its independence assumption may limit its ability to capture correlated linguistic patterns present in deceptive narratives. Empirical benchmarking studies report that linear SVMs frequently outperform Naïve Bayes across multiple datasets [5], [6]. These improvements are often attributed to margin- based optimization, which better accommodates high-dimensional sparse vectors typical of TF- IDF representations. The Passive-Aggressive Classifier (PAC), though less frequently emphasized in the literature, has been identified as particularly suitable for large-scale or streaming environments due to its online update mechanism [4], [7]. While ensemble-based methods such as XGBoost have achieved high reported accuracy in certain contexts [8], they introduce additional computational complexity and latency. This has motivated further exploration of whether a well-tuned linear model can achieve comparable in-domain performance while maintaining computational simplicity.

Recent advancements in fake news detection are strongly influenced by Deep Learning architectures, including Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Transformer-based models such as BERT [5], [8]. These approaches benefit from automated feature learning and contextual representation, often achieving high reported performance on benchmark datasets [10]. However, deep neural models typically require substantial computational resources, including high-memory GPUs and extended training times. Comparative evaluations have shown that in certain binary text classification scenarios, linear models such as SVM may achieve performance comparable to or exceeding that of recurrent neural architectures [11]. Similarly, Logistic Regression has demonstrated competitive results relative to more complex tree-based methods in specific studies [12]. These findings suggest that, for tasks where lexical separability is strong particularly in structured, long-form news articles traditional linear classifiers may offer an effective balance between accuracy and computational efficiency. Nevertheless, relatively fewer studies explicitly investigate how such models behave under cross-dataset or cross-domain conditions, especially when transitioning from long-form news content to short-form political statements. Addressing this gap forms a central motivation for the present study.

III. METHODOLOGY AND COMPUTATIONAL FRAMEWORK

The proposed methodology is structured as a streamlined text classification pipeline designed to evaluate performance under both in-domain and cross-domain settings. Unlike deep neural architectures that rely on dense embedding and multi-layer representations, this study employs sparse vector representations combined with linear classifiers.

A. Data Acquisition and Cross-Dataset Evaluation Design

To assess model behavior across heterogeneous textual environments, two datasets were utilized:

ISOT Dataset (High-Density News Articles)

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14429

A consolidated dataset consisting of 39,103 unique news articles, comprising 21,196 real and 17,907 fake news instances after duplicate removal. Although slightly imbalanced, the class distribution remains sufficiently balanced to avoid severe bias in accuracy-based evaluation [13].

LIAR Dataset (Low-Density Short Statements)

A benchmark dataset containing short-form political statements sourced from PolitiFact[14]. For consistency with binary classification settings, the original multi-class labels were mapped into two categories:

Real: True, Mostly-True

Fake: False, Barely-True, Pants-Fire

The ISOT dataset was partitioned using a stratified 80/20 split, yielding 31,282 training instances and 7,821 test instances. The LIAR dataset was not used for training; instead, it served exclusively as an out-of-distribution evaluation set to assess cross-dataset generalization.

B. Preprocessing and Text Normalization

Raw textual data contains artifacts such as URLs, punctuation, and inconsistent casing that may introduce noise into feature representations. A deterministic cleaning function C(s) was implemented using Regular Expressions (Regex) to standardize input text.

1) URL Removal

Hyperlinks were removed using the pattern r"http\S+|www\.\S+", ensuring the classifier learned from linguistic content rather than source identifiers.

2) Normalization and Noise Filtering

Text was converted to lowercase and non-alphabetic characters were removed using r"[^a-z\s]". This step reduces vocabulary sparsity and limits the feature space to alphabetic tokens.

The same preprocessing function was consistently applied to both datasets to maintain methodological uniformity.

C. Feature Extraction Using TF-IDF

Text was transformed into numerical vectors using Term Frequency–Inverse Document Frequency (TF-IDF). TF-IDF assigns weights to tokens based on their relative frequency within documents and across the corpus, thereby emphasizing discriminative terms while down-weighting ubiquitous words. For cross-dataset validation, the TF-IDF vectorizer was fitted exclusively on the ISOT training data and subsequently applied in transform mode to both the ISOT test set and the LIAR dataset. This strict separation ensures that no information from the validation dataset influenced vocabulary construction, thereby preventing data leakage.

D. Passive-Aggressive Classifier (PAC)

The Passive-Aggressive Classifier (PAC) is a margin-based online learning algorithm. Unlike probabilistic models, PAC's update mechanism is geometric, allowing it to shift the decision boundary significantly based on high-weight tokens.

1) Hinge Loss Function

The algorithm is governed by the hinge loss function, which ensures the model only learns from instances where the prediction is incorrect or the classification margin is violated:

$$L(t) = \max(0, 1 - y(t) * (w(t) \cdot x(t)))$$

Where:

L(t) is the hinge loss at time *t*.

y(t) is the true label of the document (+1 or -1).

w(t) · x(t) is the predicted score (dot product of weights and TF-IDF features).

2) Weight Update Rule

When the margin is violated (L(t) > 0), the model performs an aggressive corrective update to the weight vector **w**:

$$w(t+1) = w(t) + \tau(t) * y(t) * x(t)$$

Where the step size **τ(t)** is determined by the loss and the squared norm of the feature vector to ensure the smallest possible update that satisfies the margin:

$$\tau(t) = L(t) / \|x(t)\|^2$$

When predictions are correct with sufficient margin (L(t) = 0), the model remains passive; otherwise, it performs an aggressive corrective update to minimize the loss. This mechanism is particularly well-suited for sparse TF-IDF representations where specific keywords provide strong discriminative signals.

E. Computational Considerations

Compared to deep neural architectures that typically require large parameter counts and hardware acceleration for efficient training, linear classifiers operating on sparse matrices require substantially fewer computational resources. In our experimental setup conducted within the Kaggle CPU environment, model training completed within seconds. This highlights the practicality of linear TF-IDF-based approaches for environments where computational resources are limited.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14429

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental evaluation was conducted in two phases. First, the model was benchmarked against traditional classifiers using the high-density ISOT dataset. Second, a cross-dataset validation was performed using the LIAR benchmark to assess the model's generalizability in sparse textual environments.

A. Comparative Performance Analysis (ISOT Dataset)

We benchmarked the Passive-Aggressive Classifier (PAC) against three foundational baselines: Linear Support Vector Machines (SVM), Logistic Regression, and Multinomial Naïve Bayes (MNB). The performance metrics, calculated on a stratified 20% validation set, are detailed in Table I.

Table I: Comparative Performance Metrics on ISOT Dataset

Model	Accuracy	Precision	Recall	F1-Score
Passive-Aggressive	0.995	0.995	0.995	0.995
Linear SVM	0.994	0.994	0.994	0.994
Logistic Regression	0.987	0.987	0.987	0.987
Multinomial Naïve Bayes	0.953	0.953	0.953	0.953

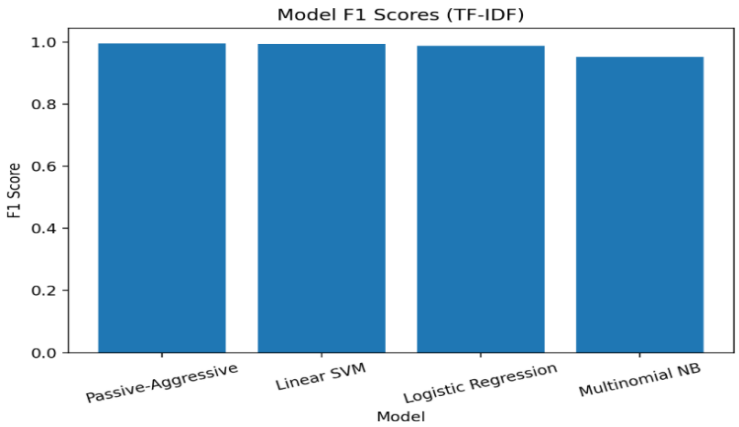

Fig. 1. Comparative Analysis of F1-Scores across PAC, Linear SVM, Logistic Regression, and Multinomial NB models

The PAC achieved the highest performance (0.995), indicating that the TF-IDF feature space is highly linearly separable for long-form news articles. In contrast, the lower performance of Naïve Bayes (0.952) aligns with the "independence assumption" which fails to account for complex linguistic dependencies found in deceptive news.

B. Cross-Dataset Generalization (LIAR Dataset)

To evaluate the model's robustness beyond its training domain, the PAC was deployed against the LIAR benchmark dataset. This dataset represents a significantly more challenging environment consisting of short-form political statements.

Table II: Cross-Dataset Validation Results

Dataset	Model	Accuracy	Macro F1- Score
ISOT(Original)	PAC	0.995	0.99
Liar(New Test)	PAC	0.474	0.43

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14429

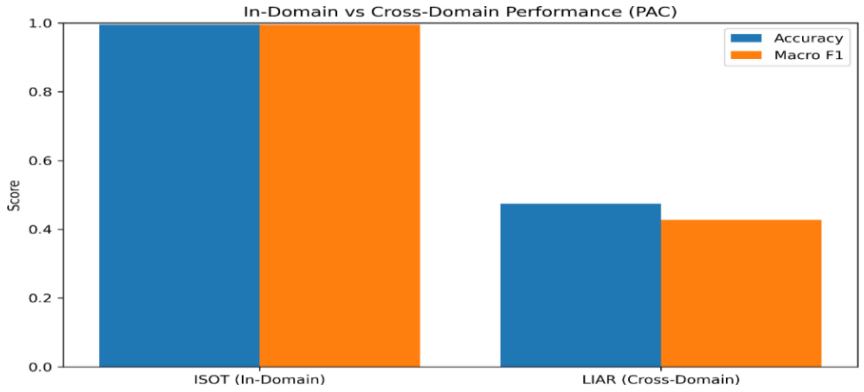

Fig. 2. Model Performance Delta: High-Density Articles (ISOT) vs. Sparse Statements (LIAR).

The results in Table II demonstrate a performance delta when shifting from high- density news articles to sparse statements. While the model maintained relatively high precision (0.87) for the 'Fake' category, it exhibited low recall for the 'Real' category, the overall accuracy drop to 0.474 reflects the inherent difficulty of classifying misinformation when textual evidence is limited.

C. Error Analysis and Confusion Matrix

The PAC misclassified only 39 articles out of 7,821 in the ISOT test set. The distribution is presented in Table III.

Table III: Confusion Matrix for Isot Test Set

	Predicted Real	**Predicted Fake**
Actual Real	4225	14
Actual Fake	25	3557

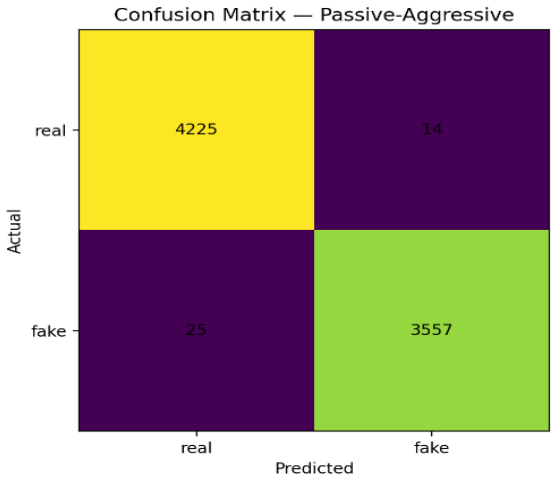

Fig. 3. Raw Confusion Matrix for PAC illustrating 3,557 True Positives and 4,225 True Negatives.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14429

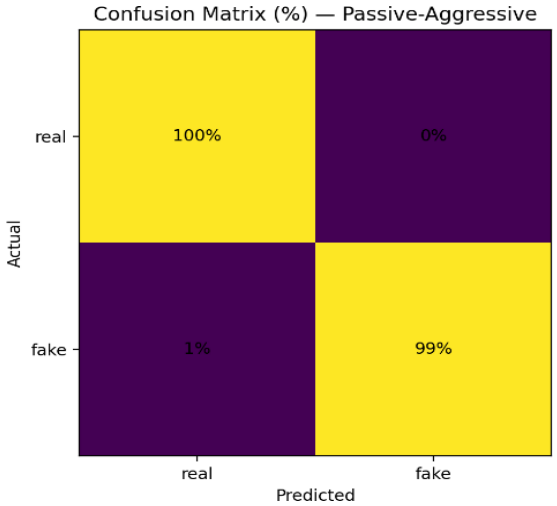

Fig. 4. Normalized Confusion Matrix (%) demonstrating near-perfect sensitivity in real news detection.

Analysis of the misclassifications across both datasets identifies two primary sources of error:

1) Contextual Sparsity

Short-form statements in the LIAR dataset often lack the keyword density required for TF-IDF to generate a distinct vector. This explains why the PAC achieved 99% on long articles but faced challenges with the short quotes in LIAR.

2) Sarcasm and Satire

Linear models struggle when formal, "credible" vocabulary is used ironically, as the statistical features align with "Real" news despite deceptive intent.

D. Benchmarking Against Deep Learning

While state-of-the-art literature has reported accuracies of 0.997 using CNNs, the 0.2% performance difference observed in this study is numerically small. Given that our PAC-based pipeline requires significantly fewer computational cycles and operates on standard CPUs, it represents a competitive accuracy-to-resource trade-off for deployment in resource-constrained environments within the Indian digital ecosystem.

V. DISCUSSION AND ANALYSIS

A. The "No Free Lunch" Trade-off

While state-of-the-art benchmarks achieved 0.997 with Deep Learning architectures, our Passive-Aggressive model reached 0.995—a numerically small difference. However, the computational cost differs substantially. While a Transformer-based model requires millions of parameters and significant VRAM, the PAC operates on a sparse matrix and completes training in seconds. These findings suggest that deep learning may not always be necessary for achieving high in-domain accuracy in structured text classification tasks.

B. Cross-Domain Generalization and Contextual Sparsity

A significant limitation identified in the literature is performance degradation when shifting datasets. Our validation on the LIAR dataset (0.474 accuracy) confirms this challenge. Unlike the high-density articles in ISOT, LIAR consists of short statements where TF-IDF vectors become "sparse," meaning there are too few unique tokens to create a clear decision boundary. Although PAC supports online updates through its hinge-loss-based mechanism, the current evaluation was conducted in batch mode. Therefore, the observed performance degradation on LIAR reflects the limitations of static training under domain shift rather than real-time adaptive behavior.

C. Explainability through Feature Attribution

To facilitate trust among human fact-checkers, such as those at the Press Information Bureau (PIB) India, "black-box" predictions are insufficient. Future work may integrate interpretability frameworks such as LIME. By perturbing the input text and observing changes in the PAC's prediction, LIME can generate a heatmap of "suspicious" tokens, providing transparency to the classification process.

VI. FUTURE SCOPE AND SOCIETAL IMPACT

The lightweight computational profile of PAC suggests potential suitability for deployment in resource-constrained environments. This is critical for the Indian ecosystem, where misinformation often spreads via encrypted platforms.

DOI: 10.17148/IJIREEICE.2026.14429

Future research will focus on Hybrid Embedding for Short-Form News to improve the 0.474 accuracy observed on the LIAR dataset, future work will integrate Word2Vec or GloVe embedding with TF-IDF to capture semantic meaning in sparse, single-sentence statements. Multilingual Support to extend the pipeline to handle Indo-Aryan morphology (Hindi and Marathi) using specialized stop-word filtration. Multi-modal Detection to integrate CNNs to analyze the relationship between text headlines and their accompanying images/videos. XAI Integration to implementing SHAP (SHapley Additive exPlanations) to provide global model interpretability.

VII. CONCLUSION

This research presented a rigorous evaluation of machine learning architectures for misinformation detection. Through optimized Regex-based preprocessing and TF-IDF vectorization, we demonstrated that the Passive-Aggressive Classifier (PAC) is a highly efficient solution, achieving 0.995 accuracy on high-density news articles while maintaining minimal hardware requirements. By matching the benchmarks of resource-heavy neural networks with a linear model, this study offers a scalable and decentralized solution to the global infodemic. However, the cross-dataset validation on the LIAR benchmark revealed a significant performance delta, with accuracy shifting to 0.474 due to the **contextual sparsity** inherent in short-form political statements. This finding provides a critical reality check for the deployment of automated systems: while linear models excel in feature-rich environments, they require hybrid semantic embedding to handle the brevity of social media misinformation. Ultimately, our findings provide a viable path for real-time, low-resource detection in linguistically diverse environments like India, provided that future iterations address the semantic challenges of short-form text.

REFERENCES

[1] U. Sharma, S. Saran, and S. M. Patil, “Fake News Detection using Machine Learning Algorithms,” *Int. J. Eng. Res.*, vol. 9, no. 3, 2020.

[2] J. Shaikh and R. Patil, “Fake News Detection using Machine Learning,” in *2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, Gunupur Odisha, India: IEEE, Dec. 2020, pp. 1–5. doi: 10.1109/iSSSC50941.2020.9358890.

[3] K. A, “A Comparative Study of Machine Learning Algorithms for Fake News Detection Using NLP Techniques,” *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 14, no. 1, pp. 847–849, Jan. 2026, doi: 10.22214/ijraset.2026.76966.

[4] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, “Fake News Detection Using Machine Learning Ensemble Methods,” *Complexity*, vol. 2020, pp. 1–11, Oct. 2020, doi: 10.1155/2020/8885861.

[5] M. F. Mridha, A. J. Keya, Md. A. Hamid, M. M. Monowar, and Md. S. Rahman, “A Comprehensive Review on Fake News Detection With Deep Learning,” *IEEE Access*, vol. 9, pp. 156151– 156170, 2021, doi: 10.1109/ACCESS.2021.3129329.

[6] M. Sudhakar and K. P. Kaliyamurthie, “A Machine Learning Framework for Automatic Fake News Detection in Indian News,” Nov. 18, 2022, *In Review*. doi: 10.21203/rs.3.rs-2268597/v1.

[7] R. Kumar, “Fake News Detection using Passive Aggressive and TF-IDF Vectorizer,” vol. 07, no.12, 2020.

[8] D. Mouratidis, A. Kanavos, and K. Kermanidis, “From Misinformation to Insight: Machine Learning Strategies for Fake News Detection,” *Information*, vol. 16, no. 3, p. 189, Feb. 2025, doi: 10.3390/info16030189.

[9] V. Agarwal, H. P. Sultana, S. Malhotra, and A. Sarkar, “Analysis of Classifiers for Fake News Detection,” *Procedia Comput. Sci.*, vol. 165, pp. 377–383, 2019, doi: 10.1016/j.procs.2020.01.035.

[10] K. Nath, P. Soni, Anjum, A. Ahuja, and R. Katarya, “Study of Fake News Detection using Machine Learning and Deep Learning Classification Methods,” in *2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, Bangalore, India: IEEE, Aug. 2021, pp. 434–438. doi: 10.1109/RTEICT52294.2021.9573583.

[11] J. Janssen, “Comparative Analysis of Machine Learning Algorithms for Fake News Detection,” Bachelor’s thesis, Dept. of Cognitive Science & Artificial Intelligence, Tilburg University, Tilburg, The Netherlands, May 2024.

[12] S. Pandey, S. Prabhakaran, N. V. Subba Reddy, and D. Acharya, “Fake News Detection from Online media using Machine learning Classifiers,” *J. Phys. Conf. Ser.*, vol. 2161, no. 1, p. 012027, Jan. 2022, doi: 10.1088/1742-6596/2161/1/012027.

[13] H. Ahmed, I. Traore, and S. Saad, "ISOT Fake News Dataset," (2017). Distributed by University of Victoria, ISOT ResearchLab. [Online].Available: https://onlineacademiccommunity.uvic.ca/isot/datasets/

[14] W. Y. Wang, "“Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 422–426. [Online]. Available: https://aclanthology.org/P17-2067/