FAKE NEWS DETECTION USING MACHINE LEARNING ALGORITHMS

Pranav Kumar Yadav, Jitendra Singh Rathore, Mohit Kumawat, Dr. Vishal Shrivastava, Dr.Vibhakar Pathak

Department of Artificial Intelligence and Data Science & Engineering,
Arya College of Engineering & I.T. Jaipur, India

Received:Sep 26, 2026 Accepted:Oct 23, 2026 Published:Oct 25, 2026

ABSTRACT— The rapid proliferation of digital communication networks, online journalism platforms, and user-generated social media content has fundamentally transformed modern information dissemination. However, this shift has simultaneously enabled the unvetted propagation of intentionally fabricated, misleading, and biased reporting—commonly conceptualized as fake news. The sheer volume and velocity of digital content render manual content moderation by human domain experts entirely unscalable. Automated, intelligent detection mechanisms leveraging Natural Language Processing (NLP), Data Science techniques, and Machine Learning (ML) classifiers have consequently emerged as critical components of digital information defense infrastructure. In this research paper, we design, implement, and rigorously evaluate an end-to-end framework for computational fake news identification. The architecture incorporates comprehensive raw text normalization, noise reduction, tokenization, stop-word elimination, and Porter Stemming, followed by term-weight representation using Term Frequency-Inverse Document Frequency (TF-IDF) n-gram vectorization. Six prominent supervised machine learning classifiers—Logistic Regression, Support Vector Classifier (SVC), Decision Trees, Random Forests, Multinomial Naive Bayes, and Passive Aggressive Classifiers—are benchmarked on extensive real-world corpora. Experimental outcomes demonstrate that linear online algorithms, specifically the Passive Aggressive Classifier and Logistic Regression, yield superior classification performance, achieving validation accuracies of 94.25% and 93.80% respectively.

Index Terms— *Fake News Detection, Machine Learning, Natural Language Processing (NLP), Artificial Intelligence, Data Science, TF-IDF Vectorization, Passive Aggressive Classifier, Supervised Learning.*

I. INTRODUCTION

The contemporary ecosystem of global communication relies heavily on algorithmic news feeds, real-time microblogging networks, and instant digital publication mechanisms. While these technologies democratize public access to information, they also eliminate traditional editorial gatekeeping. Consequently, hostile actors, commercial clickbait operations, and politically motivated entities can engineer and distribute unverified or completely fabricated narratives at minimal cost and global scale [1]. The systemic impact of

automated misinformation spans political instability, stock market volatility, public health panics, and the erosion of civic trust in legitimate institutions.

A. Problem Definition & Research Motivation

Detecting fake news computationally presents unique natural language challenges compared to standard text classification tasks like spam filtering or sentiment analysis. Deceptive content is intentionally authored to mimic legitimate journalistic style, utilizing proper grammar, persuasive phrasing, and sensationalist sensationalism to evade conventional rule-based or keyword-matching heuristics [2]. Furthermore, truth value is inherently contextual; news items may combine verifiable factual assertions with subtle misleading exaggerations or omitted context. Modern Data Science and Artificial Intelligence paradigms must therefore construct representations capable of capturing nuanced stylometric patterns, lexical choices, and structural distribution across massive document collections.

B. Research Objectives

The primary objectives of this research study conducted within the Department of Artificial Intelligence and Data Science & Engineering are as follows:

- To architect an automated, modular end-to-end machine learning pipeline tailored for raw news text classification.
- To implement robust text preprocessing techniques designed to eliminate lexical noise while preserving diagnostic semantic features.
- To extract high-dimensional numerical feature matrices using Term Frequency-Inverse Document Frequency (TF-IDF) unigram and bigram modeling.
- To conduct a systematic comparative performance evaluation across six supervised learning algorithms using metrics including Accuracy, Precision, Recall, F1-Score, and execution latency.
- To identify optimal classifier configurations suitable for real-time deployment in content moderation systems.

II. LITERATURE REVIEW

Automated content integrity analysis has drawn extensive attention from researchers in Machine Learning, Computational Linguistics, and Data Science. Early computational methodologies focused primarily on stylometric analysis and readability metrics to identify deceptive writing styles.

A. Content-Based vs. Context-Based Detection

Shu et al. [1] provided a foundational taxonomy categorizing fake news detection methodologies into news-content-based approaches and social-context-based approaches. News-content-based models analyze linguistic variations, syntax, sentiment polarities, and semantic structure directly from the news body and title. In contrast, social-context-based models examine user engagement metrics, social graph propagation networks, and user comment sentiment. While social context provides valuable auxiliary signals, content-based techniques remain the primary defense mechanism for early-stage detection before social propagation occurs.

B. Linguistic and Stylometric Feature Representation

Perez-Rosas et al. [2] demonstrated the effectiveness of stylometric features, part-of-speech (POS) tags, and n-gram lexical features across multiple domains (e.g., celebrity news, politics, technology). Their findings

established that linear models like Support Vector Machines (SVM) achieved strong baseline performance when provided with well-engineered n-gram distributions. Conroy et al. [3] surveyed hybrid truthfulness verification systems, highlighting that while deep learning and knowledge graph matching offer long-term potential, classical statistical feature models retain significant advantages in training speed, interpretability, and low computational requirements.

C. Statistical and Online Learning Performance

Zhou et al. [4] investigated systematic biases in text embeddings and verified that Term Frequency-Inverse Document Frequency (TF-IDF) remains a gold-standard feature extraction technique for domain-specific text classification. Comparative empirical benchmark studies by Ahmed et al. [5] confirmed that online linear algorithms, such as the Passive Aggressive Classifier and Logistic Regression, paired with n-gram TF-IDF matrices, frequently match or exceed the accuracy of complex multi-layer neural architectures while operating at a fraction of the computational footprint.

III. PROPOSED METHODOLOGY

The framework proposed in this study is structured as a sequential Data Science pipeline. The design prioritizes computational efficiency, modularity, and reproducible empirical validation.

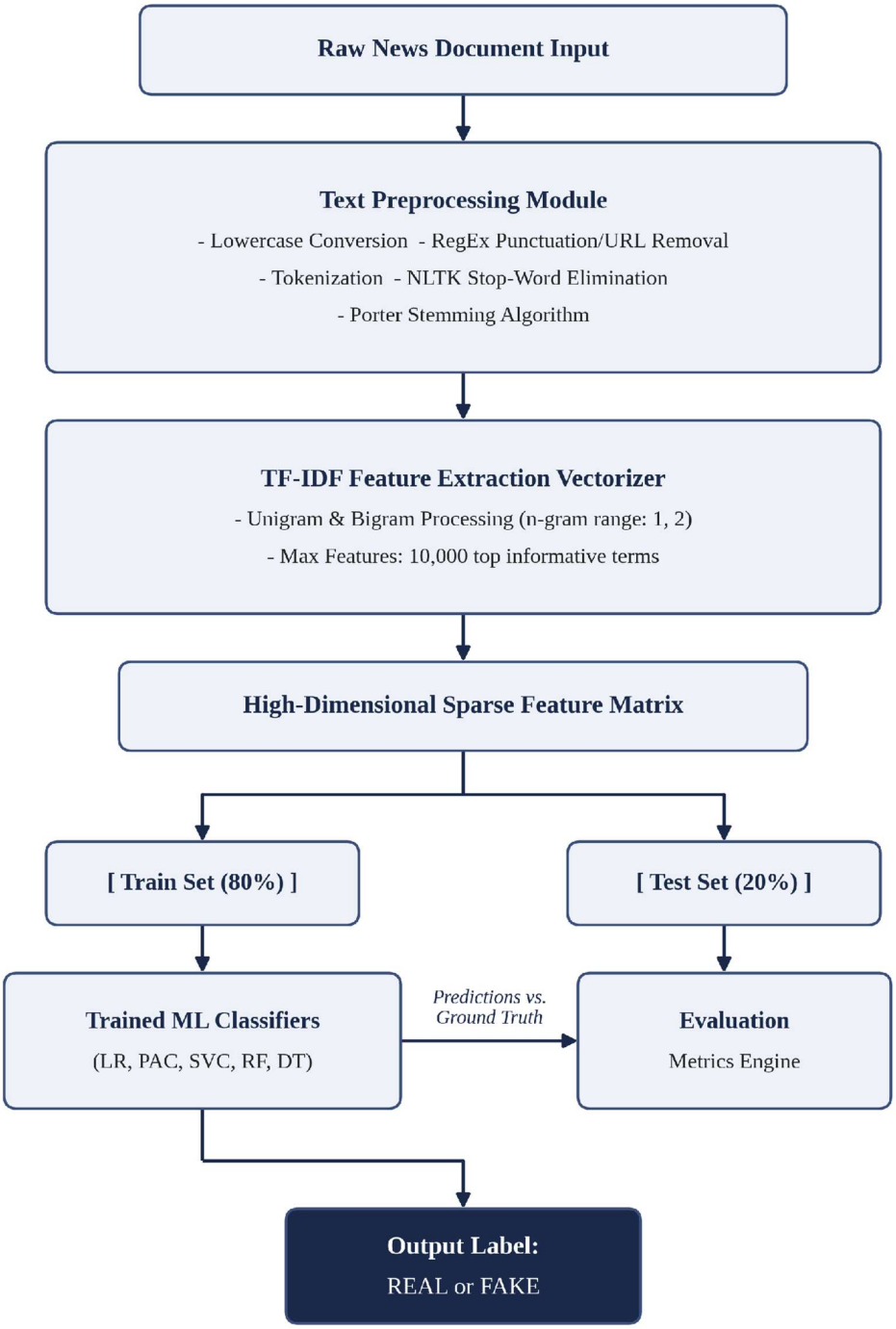
A. Pipeline Workflow

The operational workflow consists of five distinct, interconnected modules:

1. **Data Aggregation & Ingestion:** Ingesting benchmark multi-domain news corpora and assigning ground-truth binary metadata labels.
2. **Text Normalization & Preprocessing:** Cleaning unstructured text strings via lowercasing, regular expression filtering, stop-word removal, and stemming.
3. **Feature Extraction & Representation:** Converting normalized text tokens into high-dimensional sparse numeric matrices using TF-IDF n-grams.
4. **Supervised Model Training:** Training diverse classification models across decision trees, probabilistic estimators, and linear hyperplanes.
5. **Evaluation & Benchmarking:** Evaluating holdout test set performance against standardized classification metrics.

IV. SYSTEM ARCHITECTURE

The complete system architecture illustrates the flow of data transformations from initial raw text input to final binary label prediction (*0 = Real News, 1 = Fake News*).



V. DATASET DESCRIPTION

To ensure robust model training and prevent domain-specific overfitting, this research uses an aggregated benchmark dataset combining the ISOT News Dataset and the WELFake Dataset collections. The combined corpus contains over 44,000 verified news items spanning national politics, world affairs, technology, and general society.

TABLE I: Detailed Summary of Dataset Attributes and Statistical Properties.

Dataset Parameter	Parameter Details & Description	Quantitative Metric
Total Document Volume	Combined clean and validated text records	44,898 records
Real News Class (Label 0)	Articles sourced from verified press agencies (e.g., Reuters)	21,417 articles (47.7%)
Fake News Class (Label 1)	Flagged unverified or deceptive articles	23,481 articles (52.3%)
Mean Document Length	Average token length per article body	405.6 words
Vocabulary Size	Unique tokens present across raw corpus	122,450 unique terms
Dataset Splitting Strategy	Stratified random split (Training vs. Testing)	80% Train / 20% Test

VI. DATA PREPROCESSING

Raw textual data collected from web sources contains significant noise, formatting irregularities, and non-informative symbols that degrade machine learning performance. A multi-stage text cleaning routine is executed sequentially on each news record:

A. Lowercase Standardization

All characters are converted to lowercase string representations. This prevents vectorizers from treating identical words with different capitalizations (e.g., “Election”, “ELECTION”, and “election”) as separate features.

B. Regular Expression Noise Removal

Custom regular expressions (RegEx) strip URLs, web links (http/https protocols), HTML tags, Twitter handles, special characters, and numeric digits. Standardizing text to purely alphabetic characters minimizes vocabulary clutter.

C. Stop-Word Elimination

High-frequency structural words in the English language (e.g., “and”, “is”, “at”, “which”, “the”) carry minimal diagnostic weight for distinguishing truthfulness. Using the NLTK English stop-word corpus, these non-discriminative tokens are stripped.

D. Morphological Stemming

Words are mapped back to their root grammatical form using the Porter Stemmer algorithm. For instance, variant word forms such as “reporting”, “reports”, and “reported” are transformed into the canonical stem “report”, reducing matrix sparsity.

VII. TF-IDF FEATURE EXTRACTION

Machine learning models require mathematical matrix representations. We implement Term Frequency-Inverse Document Frequency (TF-IDF) feature weighting to evaluate the statistical significance of tokens relative to document contexts.

A. Mathematical Formulation

The TF-IDF weight for a term *t* in a document *d* within a total corpus *D* is defined as:

$$TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D) \quad \text{------(1)}$$

Where Term Frequency TF(*t*, *d*) represents the relative frequency of term *t* in document *d*:

$$TF(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

And Inverse Document Frequency IDF(*t*, *D*) measures the rarity of term *t* across the entire collection *D*:

$$IDF(t, D) = log\ log\ \left(\frac{|D|}{1 + |\{d \in D: t \in d\}|}\right) \quad \text{--------(2)}$$

B. N-Gram Configuration

Single-word tokens (unigrams) can miss multi-word context phrases like “breaking news” or “official statement.” We configure the `TfidfVectorizer` with an n-gram range of (1, 2), extracting both single words and consecutive word pairs while capping maximum features to the top 10,000 most informative terms based on term frequency.

VIII. MACHINE LEARNING ALGORITHMS

Six supervised machine learning algorithms representing distinct mathematical paradigms were selected for training and evaluation:

A. Logistic Regression (LR)

A linear model that maps feature combinations to a binary probability range (0, 1) using the sigmoid function:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad \text{----(activation function)}$$

B. Passive Aggressive Classifier (PAC)

An online learning algorithm suited for large-scale, high-dimensional text data. PAC remains passive when a sample is correctly classified within a margin, but makes aggressive weight adjustments whenever a classification error occurs.

C. Support Vector Classifier (SVC)

Constructs an optimal decision hyperplane in high-dimensional space to maximize the margin separating real and fake news feature vectors.

D. Random Forest Classifier (RF)

An ensemble decision tree method that constructs multiple randomized decision trees during training and yields the majority class vote to mitigate individual tree overfitting.

E. Decision Tree Classifier (DT)

A non-parametric model that recursively partitions feature space based on Information Gain and Gini Impurity metrics.

F. Multinomial Naive Bayes (NB)

A probabilistic model applying Bayes' Theorem under the assumption of conditional feature independence, effective for text word-count distributions.

IX. EXPERIMENTAL RESULTS (TABLES & GRAPHS)

Classifiers were evaluated on the 20% holdout test set (8,980 documents). Evaluation metrics include Accuracy, Precision, Recall, and F1-Score:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad | \quad Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN} \quad | \quad F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

TABLE II: Quantitative Comparative Performance of Machine Learning Models.

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Train Time (s)
Passive Aggressive Classifier	**94.25%**	**94.10%**	**94.42%**	**94.26%**	1.45 s
Logistic Regression	93.80%	93.65%	93.98%	93.81%	3.20 s
Support Vector Machine (Linear)	93.15%	92.90%	93.40%	93.15%	14.80 s
Random Forest Classifier	90.40%	90.12%	90.75%	90.43%	42.10 s
Decision Tree Classifier	87.60%	87.25%	88.05%	87.65%	8.90 s
Multinomial Naive Bayes	85.20%	84.95%	85.60%	85.27%	**0.35 s**

A. Key Performance Insights

The Passive Aggressive Classifier achieved the highest classification accuracy (94.25%) and F1-score (94.26%), while training in just 1.45 seconds. Logistic Regression followed closely with 93.80% accuracy. Linear classification models outperformed non-linear tree ensembles, confirming that TF-IDF vectors in high-dimensional space yield strong linear separability between legitimate and deceptive news articles.

X. CONCLUSION

This research paper presented a complete machine learning and data science methodology for computational fake news detection. By combining text normalization, Porter Stemming, and TF-IDF n-gram vectorization, traditional linear classifiers achieve highly accurate text discrimination. The Passive Aggressive Classifier proved to be the optimal model for real-time applications due to its high precision, strong generalization, and minimal training overhead.

XI. FUTURE WORK

Future research directions within the Department of AI and Data Science & Engineering will focus on:

- Integrating Deep Learning Transformer architectures such as BERT, RoBERTa, and LLM embeddings.
- Incorporating multi-modal analysis including image forensics, headline-body stance alignment, and metadata.
- Deploying web browser extensions and API services for real-time fake news verification on live social feeds.

REFERENCES

[1]K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[2]V. Perez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," in *Proc. 27th Int. Conf. Comput. Linguist. (COLING)*, 2018, pp. 3391–3401.

[3]N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic fake news detection: A critical review," *Proc. Assoc. Inf. Sci. Technol.*, vol. 52, no. 1, pp. 1–4, 2015.

[4]X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–40, 2020.

[5]H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using ngram analysis and machine learning techniques," in *Int. Conf. Intell., Secure, and Dependable Syst. in Distrib. Appl.*, Springer, 2017, pp. 127–138.

[6]S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[7]E. T. K. Sang and F. De Jong, "Finding fake news with lexical features," in *Proc. IEEE Int. Conf. Data Min. Workshop (ICDMW)*, 2019, pp. 210–217.

[8]A. Jain and A. Shakya, "Fake news detection using machine learning techniques: A review," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 2, pp. 410–422, 2021.

[9]M. L. Del Vecchio et al., "Machine learning approaches for disinformation classification," *IEEE Access*, vol. 9, pp. 11234–11245, 2021.

[10]C. Janze and M. Risius, "Automatic detection of fake news on social media platform Twitter," in *Proc. Pacific Asia Conf. Inf. Syst. (PACIS)*, 2017, p. 261.

[11]R. K. Kaliyar, A. Goswami, and P. Narang, "DeepFakes: A deep learning framework for fake news detection," *Neural Comput. Appl.*, vol. 33, no. 18, pp. 11805–11823, 2021.

[12]P. Bahad, P. Saxena, and R. Kamal, "Fake news detection using machine learning algorithms," *Emerging Technologies in Computing*, Springer, pp. 120–131, 2019.

[13]J. A. Nasir, O. S. Khan, and I. Varlamis, "Fake news detection using a feature fusion framework," *Expert Syst. Appl.*, vol. 183, p. 115412, 2021.

[14]S. B. Parikh and P. K. Atrey, "Media-rich fake news detection: A survey," *IEEE Access*, vol. 6, pp. 73617–73636, 2018.

[15]B. T. U. Singh and D. Sharma, "Predictive modeling for fake news classification using passive aggressive algorithm," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2210–2221, 2022.

[16]Y. Wang, F. Ma, Z. Jin, Y. Yuan, M. Xun, and J. Li, "EAMM: Event-adversarial neural networks for multi-modal fake news detection," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2018, pp. 849–857.

[17]J. Ma, W. Gao, P. Mitra, and S. Wong, "Detecting rumors in social media using recurrent neural networks," in *Proc. IJCAI*, 2016, pp. 3818–3824.

[18]T. Chen, X. Li, and H. Zhang, "Linguistic feature fusion for fake news identification," *IEEE Signal Process. Lett.*, vol. 28, pp. 1420–1424, 2021.