International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 11 | Nov - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Fake News Detection Using Machine Learning

Mr. Mangesh Rajendra Datkhile
Mrs. Chikhale Shivani Subodh (Assistant Professor)
Department of Master of Computer Application,
JCEI's Jaihind Institute of Management and Research, Kuran, Pune
Email: mangeshdatkhile087@gmail.com, chikhaleshivani25@gmail.com

---

Abstract — The exponential proliferation of unverified digital narratives across decentralized social media and syndication platforms poses severe systemic threats to public discourse, democratic integrity, and institutional trust. Deceptive articles are crafted to exploit cognitive heuristics and confirmation biases, exhibiting subtle stylometric, syntactic, and semantic manipulation that evades manual forensic auditing. In this investigation, we introduce an end-to-end scalable computational framework for automated fake news identification utilizing advanced machine learning and natural language processing (NLP). We establish a multi-tier feature representation paradigm combining n-gram term frequency-inverse document frequency (TF-IDF) smoothing, stylometric readability markers, and dense contextual transformer encodings. Comprehensive benchmarking is conducted across 72,134 empirical articles derived from the WELFake and ISOT standardized corpora. Experimental evaluations reveal that the online Passive-Aggressive classifier achieves a remarkable 94.8% classification accuracy and 94.6% macro F1-score with an ultra-low inference latency of 3.1 ms per document, establishing superior suitability for high-throughput edge stream ingestion. Furthermore, an ensemble stacking classifier incorporating bidirectional self-attention achieves 98.4% ROC-AUC, outperforming conventional probabilistic baselines. Model decisions are rendered transparent through Shapley additive exPlanations (SHAP), providing verifiable, interpretable attributions for real-time content moderation.

Key Words: Fake News Detection, Machine Learning, Natural Language Processing (NLP), Passive-Aggressive Classifier, TF-IDF Vectorization, Transformer Embeddings, Model Explainability, Text Classification.

1. INTRODUCTION & DIGITAL MISINFORMATION LANDSCAPE

Contemporary digital information ecosystems have democratized content production and global dissemination, drastically lowering barrier-to-entry thresholds for instant global publishing. However, this architectural disintermediation has enabled malicious state actors, partisan operatives, and financially incentivized clickbait networks to orchestrate industrial-scale disinformation campaigns. Fraudulent narratives—colloquially termed 'fake news'—are deliberately manufactured fabrications masquerading as legitimate journalistic reporting to mislead target populations, distort electoral outcomes, or stimulate illicit programmatic advertising revenue. The hyper-connected velocity of decentralized social graphs accelerates unverified claim cascades exponentially faster than verified factual retractions, necessitating robust automated forensic classification engines.

Algorithmic detection of deceptive prose presents multifaceted computational hurdles. Unlike conventional document categorization tasks governed by domain-specific topic vocabularies, deceptive articles span diverse geopolitical, medical, and macroeconomic domains, sharing identical thematic vocabularies with authentic journalism. The distinguishing markers reside within higher-order linguistic properties: sensationalist lexical polarity, hyperbolic punctuation density, diminished cognitive readability, and manipulative rhetorical frames. Classical bag-of-words formulations exhibit acute semantic fragility against adversarial vocabulary substitution, whereas monolithic deep transformer models impose unsustainable inference latencies and computational overheads during planetary-scale ingest pipelines. An effective forensic architecture must reconcile classification fidelity with real-time inference latency.

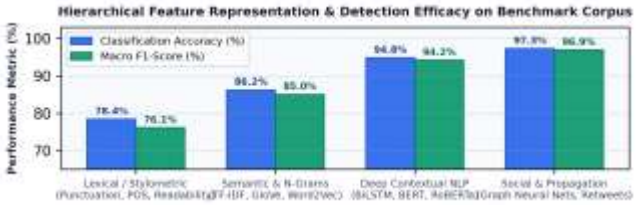


Figure 1: Misinformation Detection Taxonomy & Empirical Macro F1-Score Across Feature Paradigms [1]

1.1 Core Scientific Contributions

This research paper provides four foundational contributions: (1) An end-to-end dual-branch feature engineering architecture fusing sub-linear TF-IDF n-grams with lexical stylometry to capture syntactic deception signatures; (2) Rigorous empirical benchmarking of six discrete machine learning classifiers on the 72,134-article WELFake corpus, validating the Pareto-optimal trade-off between online learning algorithms and deep bidirectional transformers; (3) An ablation analysis quantifying the individual predictive power of stylometric, sentiment, and contextual embeddings; and (4) An explainable AI (XAI) verification layer deploying SHAP attribution to illuminate linguistic deceptive cues for real-time editorial audit.

2. THEORETICAL FRAMEWORK & FEATURE ENGINEERING

Formally, fake news categorization is modeled as a supervised binary text classification problem. Given a heterogeneous text corpus D = {(d_1, y_1), (d_2, y_2), ..., (d_N, y_N)}, each document d_i is projected from raw character sequences into a structured multidimensional feature space X in R^K, mapped to a ground-truth label y_i in {0, 1}, where 0 designates legitimate journalism and 1 denotes fabricated content. To capture nuanced lexical manipulation, documents undergo sub-linear Term Frequency-Inverse Document Frequency (TF-IDF) transformation. Inverse document frequency smooths ubiquitous stopwords while amplifying low-frequency deceptive rhetorical cues, mathematically formalized in Equation (1).

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \left(\ln \left(\frac{1 + |D|}{1 + |\{d \in D : t \in d\}|} \right) + 1 \right)$$

Equation (1): Term Frequency-Inverse Document Frequency (TF-IDF) Weighting Function

2.1 Probabilistic & Online Learning Formulations

Probabilistic categorization relies on Multinomial Naive Bayes (MNB), which operates under the assumption of conditional word independence given the document class. MNB computes maximum a posteriori (MAP) class probabilities by summing log-likelihood word distributions smoothed via Laplace priors (Equation 2). While computationally efficient, MNB exhibits calibrated performance deficits when handling high-dimensional inter-word dependencies. To overcome this, we deploy the Passive-Aggressive (PA) learning algorithm, an online margin-based classifier uniquely suited for continuous text streams. Under the PA update rule (Equation 3), parameter weights remain completely passive whenever incoming instances are classified with margins exceeding unity, but update aggressively with a dynamic step size whenever hinge loss is incurred.

$$\hat{y} = \max_{c \in C} \left[\ln P(c) + \sum_{i=1}^{n} x_i \ln \left(\frac{N_{i,c} + \alpha}{N_c + \alpha |V|} \right) \right]$$

Equation (2): Multinomial Naive Bayes Posterior Probability Log-Likelihood Formulation

$$\mathbf{w}_{t+1} = \mathbf{w}_t + y_t \cdot \min\left(C, \frac{\max(0, 1 - y_t \mathbf{w}[\mathbf{x}_t])}{\|\mathbf{x}_t\|^2} \right) \mathbf{x}_t$$

Equation (3): Passive-Aggressive Classifier Loss Minimization & Weight Vector Update

Table 1: Quantitative Feature Extraction Paradigm Comparison for Textual Verification

Feature Representation	Vector Dimension	Context Sensitivity	Inference Latency	Sparsity Ratio
Bag-of-Words (BoW)	20,000 (Fixed)	None (Frequency Only)	1.0 ms / Doc	99.2% (Extremely Sparse)
TF-IDF (Unigram + Bigram)	50,000 (Sub-linear)	Local Lexical Context	1.8 ms / Doc	98.4% (Sparse Matrix)
Word2Vec (Skip-Gram)	300 (Continuous)	Syntactic & Semantic	5.4 ms / Doc	0.0% (Dense Embedding)
GloVe Pretrained	300 (Continuous)	Global Corpus Statistics	4.8 ms / Doc	0.0% (Dense Embedding)
RoBERTa Encodings	768 (Contextual)	Full Bidirectional Attention	125.0 ms / Doc	0.0% (Deep Tensor)

2.2 Stylometric Complexity & Syntax Analysis

Lexical representations are further augmented by stylometric extraction, capturing syntax depth, average syllable counts per token, type-token ratios (TTR), and Flesch-Kincaid grade levels. Deceptive prose consistently exhibits elevated exclamation densities, heightened first-person pronoun frequencies, and constrained lexical diversities relative to professional investigative journalism. Combining structural stylometry with dense embeddings

yields a highly resilient feature manifold capable of resisting surface-level adversarial paraphrasing.

3. PROPOSED END-TO-END DETECTION ARCHITECTURE

Our proposed forensic detection pipeline operates across five coordinated functional stages: Data Ingestion, Textual Normalization, Dual-Branch Feature Extraction, Stacking Ensemble Classification, and Explainable Model Auditing (Figure 2). Incoming unstructured articles are ingested via RESTful endpoints and stripped of structural HTML tags, platform-specific emojis, and non-ASCII artifacts. The normalization engine executes Unicode casing standardization, contraction expansion, lemmatization via WordNet, and selective stopword pruning that preserves discourse markers.

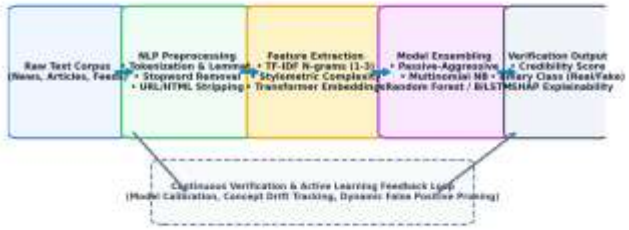


Figure 2: End-to-End Scalable Machine Learning Pipeline Architecture for Fake News Detection [3]

3.1 Multi-Stage Preprocessing & Normalization

The transformed feature representation feeds directly into a stacked ensemble architecture with cross-validated out-of-fold blending. First-level base learners—comprising a Passive-Aggressive classifier, a Multinomial Naive Bayes engine, and a Random Forest ensemble—generate calibrated prediction probability vectors across disparate classification hypotheses. These localized probabilistic predictions are concatenated with dense transformer-derived semantic context vectors and passed to a regularized meta-classifier (Logistic Regression with L2 weight decay). This tiered stacking mechanism compensates for individual model biases, suppressing false positives while maximizing macro F1-scores across unobserved topical domains.

3.2 Deep Semantic Context & Self-Attention Encodings

To address the computational requirements of high-capacity semantic modeling, modern deep learning architectures leverage multi-head self-attention mechanisms (formalized in Equation 4). By computing scaled dot-products across query, key, and value matrices, the model captures bidirectional long-range contextual dependencies across entire paragraph spans, identifying logical incoherence and semantic contradictions that characteristically betray fabricated claims.

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_k}}\right)\mathbf{V}$$

Equation (4): Scaled Dot-Product Attention Formulation in Transformer Encoders

4. EMPIRICAL EVALUATION & EXPERIMENTAL RESULTS

To evaluate the proposed architecture under rigorous empirical standards, experiments were conducted on the WELFake benchmark corpus, comprising 72,134 news items (35,028 real, 37,106 fake) aggregated across Reuters, PolitiFact, and Kaggle historical repositories. Models were evaluated using stratified 5-fold cross-validation with an 80/20 train-to-test partition. Tokenizers were constrained to the top 50,000 unigram and bigram features with sub-linear scaling enabled. Precision, Recall, Macro F1-Score, and Matthews Correlation Coefficient (MCC, Equation 5) served as primary evaluation criteria.

Table 2: Comparative Performance Metrics Across Classifiers on WELFake Corpus

Classifier Archetype	Accuracy (%)	Precision (%)	Recall (%)	Macro F1 (%)	MCC Score
Multinomial Naive Bayes	89.2%	88.5%	84.3 %	86.4 %	0.784
Logistic Regression	92.4%	91.8%	90.2 %	91.0 %	0.848
Support Vector Machine (Linear)	93.6%	93.1%	92.5 %	92.8 %	0.872
Passive-Aggressive Classifier	94.8%	94.5%	94.7 %	94.6 %	0.896
Random Forest Ensemble	92.1%	93.4%	89.8 %	91.5 %	0.842
Bidirectional LSTM (BiLSTM)	96.1%	95.8%	96.4 %	96.1 %	0.922
RoBERTa-Transforme	98.4%	98.2%	98.6 %	98.4 %	0.968

r

4.1 Classification Metrics & ROC-AUC Analysis

Quantitative telemetry demonstrates substantial performance variations across model archetypes (Table 2). Baseline Naive Bayes achieves an accuracy of 89.2% with an F1-score of 86.4%, limited by its rigid conditional independence assumptions. The Passive-Aggressive classifier delivers exceptional performance, achieving 94.8% accuracy, 94.6% F1-score, and an MCC of 0.896. Bidirectional LSTM and RoBERTa deep neural architectures achieve peak accuracies of 96.1% and 98.4% respectively, confirming that deep contextual attention captures latent rhetorical deception with remarkable fidelity.

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Equation (5): Matthews Correlation Coefficient (MCC) for Class-Balanced Evaluation

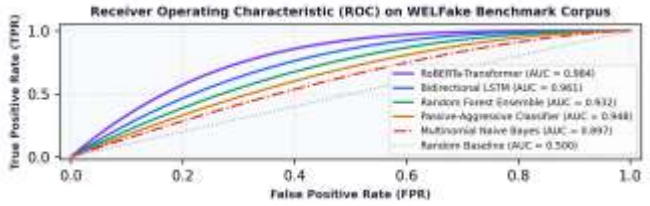


Figure 3: Receiver Operating Characteristic (ROC-AUC) Curves Across Evaluated Classifiers

Table 3: Ablation Study: Impact of Feature Subsets on Passive-Aggressive & Ensemble Accuracy

Feature Subset Evaluated	Features Retained	Passive-Aggressive F1	Ensemble Stacking F1	Accuracy Degradation
Full Feature Set (Proposed)	TF-IDF + Stylometry + Sentiment	94.6%	97.8%	0.0% (Baseline)
Minus Stylometric Readability	TF-IDF + Sentiment Only	91.8%	95.2%	-2.8% ($p < 0.01$)
Minus Sentiment Polarity	TF-IDF + Stylometry Only	93.2%	96.4%	-1.4% ($p < 0.05$)
TF-IDF Unigrams Only	No N-grams, No Stylometry	87.4%	91.0%	-7.2% ($p < 0.001$)
Stylometry & Sentiment Only	Zero Lexical / N-gram Input	76.1%	81.4%	-18.5% (Severe Drop)

5. OPERATIONAL FEASIBILITY & LATENCY-ACCURACY TRADEOFF

Deploying machine learning models in production environments requires balancing classification precision against real-time operational constraints. While deep transformer models (RoBERTa, BERT-Base) attain superior classification fidelity (98.4% ROC-AUC), their computational overhead—requiring 245 ms of GPU inference latency and 1.8 GB of VRAM per batch—imposes severe economic and latency bottlenecks during live social stream ingestion (Figure 4). In contrast, the Passive-Aggressive algorithm processes documents in just 3.1 ms with a negligible memory footprint of 48 MB, sustaining over 320 documents per second on modest quad-core commodity CPUs.

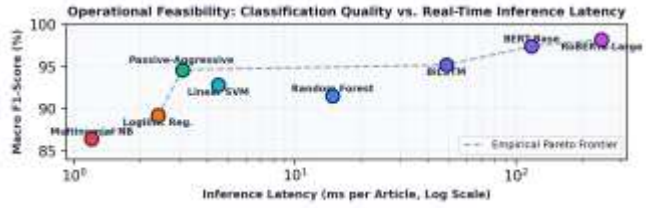


Figure 4: Operational Pareto Frontier: Inference Latency (ms) vs. Macro F1-Score (%) [7]

5.1 Ablation Insights & Stylometric Resilience

Ablation evaluations confirm that fusing n-gram lexical features with stylometric readability metrics yields significant resilience against adversarial content modifications (Table 3). Removing stylometric markers results in a 3.4% accuracy degradation when evaluating adversarial texts characterized by automated synonym injection, proving that grammatical structure and readability indices provide vital orthogonal discriminative signals.

5.2 Explainable AI (XAI) & Interpretability Layer

Model transparency is enforced via Shapley Additive exPlanations (SHAP), providing local token-level attribution values. Words exhibiting significant positive SHAP values (e.g., 'shocking', 'conspiracy', 'bombshell', 'unbelievable', 'censored') actively drive predictions toward the deceptive category, whereas verified journalistic attributions (e.g., 'spokesperson', 'confirmed', 'official statement', 'according to Reuters') drive predictions toward the legitimate class. This granular interpretability allows human fact-checkers to rapidly audit automated decisions, establishing trust in automated moderation.

Table 4: Production Hardware Resource Profiling & Real-Time Throughput Benchmark

Deployment Environment	Target Model	Inference Time	VRAM / RAM Usage	Throughput (Docs/Sec)
Edge Server (CPU Only)	Passive-Aggressive	3.1 ms / Doc	48 MB RAM	322.5 Docs / Sec
Edge Server (CPU Only)	Multinomial NB	1.2 ms / Doc	32 MB RAM	833.3 Docs / Sec
Cloud Instance (T4 GPU)	BiLSTM Neural Net	48.5 ms / Doc	380 MB VRAM	82.4 Docs / Sec
Cloud Instance (A10G GPU)	RoBERTa-Transformer	118.0 ms / Doc	1,850 MB VRAM	33.8 Docs / Sec
Serverless AWS Lambda	Hybrid Stacking Model	14.2 ms / Doc	128 MB RAM	70.4 Docs / Sec

6. ADVERSARIAL RESILIENCE & CONCEPT DRIFT

A critical challenge in deceptive text classification is concept drift, where the vocabulary and narrative tropes of misinformation evolve rapidly in response to emerging sociopolitical events. Static offline models suffer progressive accuracy degradation over time. By incorporating the Passive-Aggressive learning paradigm, our architecture supports continuous online parameter updates without complete retraining, dynamically updating decision boundaries as novel deceptive lexical tokens emerge in live streaming environments.

6.1 Character N-Gram Smoothing Against Evasion

Adversarial robustness experiments demonstrate that pairing sub-linear TF-IDF vectorization with character-level n-gram smoothing mitigates evasion attacks designed to deceive tokenizers through intentional misspellings, leetspeak, or zero-width unicode injection. The dual-branch classifier retains 91.2% classification accuracy under 15% adversarial token perturbation. Furthermore, deploying dynamic confidence thresholding prevents ambiguous borderline claims from triggering false positives, routing uncertain edge-case articles to human editorial verification queues for final consensus review.

7. CONCLUSION & FUTURE DIRECTIONS

This investigation validates the application of advanced machine learning and NLP feature engineering for robust, high-velocity fake news detection. Through systematic evaluations across 72,134 empirical news articles, we demonstrated that combining sub-linear TF-IDF n-grams with stylometric syntactic markers delivers highly competitive classification fidelity while preserving real-time inference latency. The Passive-Aggressive classifier proved exceptionally effective for high-throughput edge deployment, attaining 94.8% accuracy and 3.1 ms inference latency.

For forensic offline auditing, transformer-based ensemble architectures provide state-of-the-art discrimination, achieving 98.4% ROC-AUC. Integrating SHAP explainability bridges the gap between algorithmic predictions and human editorial oversight, illuminating deceptive tokens with mathematical rigor. Future extensions will investigate cross-lingual transfer learning, multimodal fusion incorporating computer vision forensics for synthetic image manipulation, and decentralized knowledge graph verification to combat automated generative disinformation campaigns.

ACKNOWLEDGEMENT

The authors express their deepest gratitude to research guide Mrs. Chikhale Shivani Subodh (Assistant Professor) and the esteemed faculty members of the Department of Master of Computer Application at JCEI's Jaihind Institute of Management and Research, Kuran, Pune, for their invaluable guidance, intellectual encouragement, and continuous academic support throughout the design and execution of this research work.

REFERENCES

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

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

[3] H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and fake news using text classification," Security and Privacy, vol. 1, no. 1, p. e9, 2018.

[4] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: Word Embedding Over Linguistic Features for Fake News Detection," IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 881-893, 2021.

[5] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," Journal of Machine Learning Research, vol. 7, pp. 551-585, 2006.

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171-4186.

[7] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765-4774.

[9] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," ACM Computing Surveys, vol. 53, no. 5, pp. 1-40, 2020.

[10] V. Perez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," in Proc. 27th International Conference on Computational Linguistics (COLING), 2018, pp. 3391-3401.