International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 14, Issue 2, 2026

Big Data ML-Based Fake News Detection Using Distributed Learning

Syed Fardeen[1], Mohd Irfan Khan[2], Mohammed Omer Ali[3], Dr. Khaja Mizbahuddin Quadry[4]
[1,2,3]BTech Students Department of Computer Science & Engineering, Lords Institute of Engineering and Technology, Hyderabad,
[4]Associate Professor Department of Computer Science & Engineering, Lords Institute of Engineering and Technology, Hyderabad, India
fardeensyed63@gmail.com , kmohdirfan786@gmail.com , mohammedomerali7037@gmail.com, khajaquadry@lords.ac.in

Abstract
*The rapid growth of digital media has accelerated the proliferation of misinformation and fake news, making automated detection systems essential. This paper presents the **Fake News Intelligence System (FNIS)**, a scalable Big Data Machine Learning–Based Fake News Detection system that integrates Apache Spark distributed processing, a RoBERTa transformer classifier, ChromaDB semantic evidence retrieval, and a hybrid trust scoring mechanism. Experimental results demonstrate Accuracy of 98.7%, Precision of 98.7%, Recall of 98.5%, and F1-Score of 98.6%, significantly outperforming traditional machine learning baselines. The system is deployed via FastAPI with an interactive real-time dashboard, providing an end-to-end solution for large-scale misinformation detection.*

Keywords: *Fake News Detection, RoBERTa, Apache Spark, Distributed Learning, Natural Language Processing, Trust Scoring, ChromaDB, Misinformation*

1. Introduction

The exponential growth of digital media and social networking platforms has fundamentally transformed how information is produced, distributed, and consumed. While this digital revolution democratizes access to information, it has simultaneously created fertile ground for the rapid propagation of misinformation and fake news. Fake news refers to fabricated or deliberately misleading content presented as legitimate journalism, designed to manipulate public opinion, incite social unrest, or undermine trust in credible information sources.

According to recent studies, false information spreads approximately six times faster than truthful information on social media platforms. The consequences are far-reaching, impacting democratic elections, public health responses, and social cohesion. During the COVID-19 pandemic alone, the World Health Organization declared an 'infodemic'—an overabundance of misinformation that endangered public safety.

Traditional fact-checking organizations such as Snopes, PolitiFact, and FactCheck.org rely heavily on manual verification processes conducted by domain experts. While effective, these approaches are fundamentally limited in their throughput and cannot scale to address the volume of content generated each day across billions of social media posts, blog articles, and news publications.

Automated machine learning systems have emerged as promising solutions for large-scale fake news detection. Early approaches leveraged classical methods such as Naive Bayes classifiers, Support Vector Machines (SVM), and Logistic Regression applied to bag-of-words or TF-IDF representations. While these methods provided a foundation, they lack the semantic depth required to capture nuanced language patterns, sarcasm, and contextual misinformation.

This paper presents the Fake News Intelligence System (FNIS), a comprehensive Big Data ML-based platform that addresses these limitations through three integrated innovations: (1) distributed data processing using Apache Spark and PySpark for scalability; (2) contextual classification using a fine-tuned RoBERTa transformer model; and (3) a novel hybrid trust scoring mechanism combining ML confidence, source credibility, and semantic evidence similarity retrieved from a ChromaDB vector database.

1.1 Research Objectives

- To design and implement a distributed data preprocessing pipeline using Apache Spark capable of handling large-scale news corpora.
- To develop and fine-tune a RoBERTa-based transformer model for binary classification of news articles as real or fake.
- To implement a semantic evidence retrieval module using ChromaDB vector embeddings for context-aware verification.
- To formulate and validate a hybrid trust scoring algorithm that integrates multiple verification signals.
- To build a production-grade FastAPI deployment with a real-time monitoring dashboard.
-

2. Literature Survey

Fake news detection has evolved rapidly from rule-based heuristics to sophisticated deep learning architectures. This section critically examines seminal works that inform the design of FNIS.

2.1 Classical Machine Learning Approaches

Shu et al. (2017) provided a foundational data mining perspective on fake news detection, categorizing features into news content, social context, and knowledge graph signals. They demonstrated that SVM classifiers operating on n-gram features achieved approximately 78% accuracy on the LIAR dataset, establishing early benchmarks for the

field. Zhou and Zafarani (2020) expanded this survey to cover content-based analysis, propagation pattern analysis, and source credibility evaluation, arguing that multidimensional verification signals substantially improve detection reliability.

2.2 Deep Learning Advances

Kaliyar et al. (2021) introduced FakeBERT, demonstrating that BERT-based architectures significantly outperform traditional classifiers by capturing bidirectional contextual representations. Their model achieved 98.3% accuracy on benchmark datasets. Li et al. (2022) reviewed CNN and RNN architectures for fake news detection, showing that deep neural networks can automatically extract hierarchical semantic features, eliminating the need for extensive manual feature engineering.

2.3 Transformer-Based Systems

Azizah et al. (2023) conducted a systematic comparison of BERT, ALBERT, and RoBERTa for fake news detection, finding that RoBERTa consistently outperformed its counterparts due to its dynamic masking strategy, larger training corpus, and removal of the Next Sentence Prediction objective. Truica and Apostol (2022) proposed MisRoBERTa, a specialized transformer architecture trained on misinformation-specific corpora, achieving state-of-the-art performance on multiple benchmarks.

2.4 Hybrid and Evidence-Based Systems

Ali et al. (2025) proposed a hybrid framework combining transformer embeddings with TF-IDF features, demonstrating that ensemble representations improve generalization. However, none of the surveyed systems addressed large-scale distributed processing or incorporated real-time evidence retrieval from vector databases—gaps that FNIS specifically addresses.

Table 1: Literature Survey Comparison

Author & Year	Model	Dataset	Key Finding	Limitation
Shu et al. (2017)	SVM, NB	LIAR Dataset	Established fake news taxonomy and feature categories	Limited semantic depth
Zhou & Zafarani (2020)	Survey	Multiple	Multi-signal detection improves reliability	No unified system
Truica & Apostol (2022)	MisRoBERTa	Large-scale corpora	Transformers outperform classical ML	No distributed processing
Li et al. (2022)	CNN, RNN	NLP benchmarks	Deep learning extracts semantic features automatically	No evidence verification
Azizah et al. (2023)	BERT, RoBERTa	Social media datasets	RoBERTa achieves superior contextual understanding	No trust scoring
Ali et al. (2025)	Hybrid Transformer	Multi-source	Hybrid features improve generalization	No real-time deployment
FNIS (Proposed)	RoBERTa + ChromaDB + Spark	WELFake, ISOT	98.7% accuracy with distributed evidence retrieval	Ongoing multilingual extension

3. Problem Formulation and Mathematical Framework

3.1 Formal Problem Definition

Let $D = \{d_1, d_2, ..., d_n\}$ denote a corpus of n news articles, where each article d_i consists of textual content t_i and associated metadata m_i (source, timestamp, URL). The binary classification problem is defined as learning a mapping function f such that:

$$\mathbf{f : D \rightarrow \{0, 1\}}$$
$$\mathbf{where\ \ 0 \equiv Real\ News,\ \ 1 \equiv Fake\ News}$$

The optimization objective is to minimize the categorical cross-entropy loss:

$$L(\theta) = -\Sigma_i \left[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i) \right]$$

where $y_i \in \{0,1\}$ is the ground truth label
and $\hat{y}_i = P(\text{fake} \mid d_i; \theta)$ is the predicted probability

3.2 RoBERTa Classification Model

RoBERTa (Robustly Optimized BERT Pretraining Approach) employs the Transformer encoder architecture with multi-head self-attention. Given an input token sequence $X = [x_1, x_2, ..., x_l]$, the self-attention mechanism computes:

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

$$Q = XW_\varphi, \quad K = XW_k, \quad V = XW_v$$

where d_k is the key dimension and W are learned weight matrices

Multi-head attention combines h attention heads to capture diverse contextual relationships:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h) \cdot W_o$$

$$\text{head}_i = \text{Attention}(QW_{\varphi i}, KW_{ki}, VW_{vi})$$

The [CLS] token representation from the final encoder layer is passed through a classification head:

$$\hat{y} = \text{softmax}\left(W \cdot h[\text{CLS}] + b \right)$$

where $h[\text{CLS}] \in \mathbb{R}^d$ is the pooled representation

3.3 Hybrid Trust Scoring Algorithm

The FNIS system generates a final Hybrid Trust Score (Φ) that integrates three verification signals through a weighted linear combination:

$$\Phi = \alpha \cdot C\_ML + \beta \cdot C\_ev + \gamma \cdot S\_src$$

$\alpha = 0.40$ (ML Confidence weight)
$\beta = 0.30$ (Evidence Consensus weight)
$\gamma = 0.30$ (Source Credibility weight)
$\alpha + \beta + \gamma = 1.0$

Where the individual components are defined as:

$$C\_ML = \max(\text{softmax}(\text{logits})) \quad \text{[Model prediction confidence]}$$

$$C\_ev = (1/k) \Sigma_j \text{cosine\_sim}(e_i, e_j) \quad \text{[Mean evidence similarity]}$$

where e_i is the article embedding
and e_j are the k nearest neighbors in ChromaDB

$$S\_src = f(\text{domain\_rank}, \text{historical\_accuracy}) \in [0, 1]$$

[Source credibility from external knowledge base]

The final binary classification decision is determined by:

$$\text{Decision} = \begin{cases} \text{Real News} & \text{if } \Phi \geq \tau \\ \text{Fake News} & \text{if } \Phi < \tau \end{cases}$$

where $\tau = 0.50$ is the decision threshold

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 14, Issue 2, 2026

3.4 Cosine Similarity for Semantic Evidence Retrieval

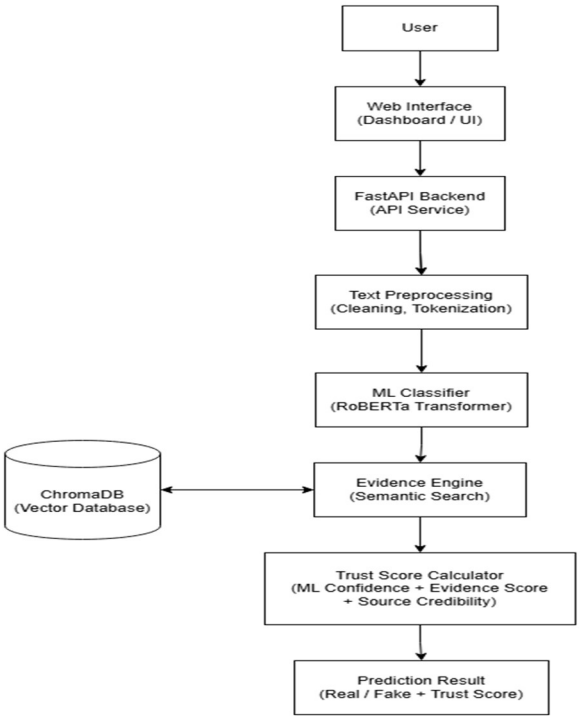
Semantic similarity between the query article embedding and stored evidence embeddings is computed using cosine similarity:

$$\text{cosine\_sim}(A, B) = (A \cdot B) / (||A|| \cdot ||B||)$$

$$= \Sigma_i(A_i \cdot B_i) / \sqrt{(\Sigma_i A_i^2)} \cdot \sqrt{(\Sigma_i B_i^2)}$$

4. System Architecture and Design

FNIS is designed as a six-layer modular architecture that enables scalable, evidence-aware fake news detection.



Each layer encapsulates specific functionality while maintaining clean interfaces for inter-module communication.

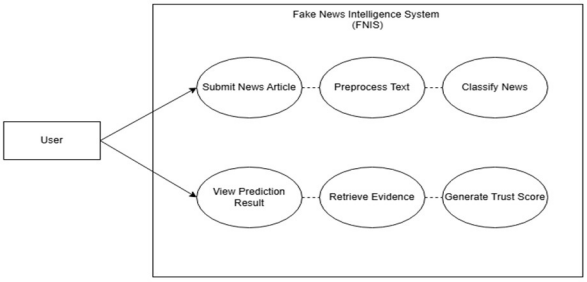
Figure 1: Six-layer FNIS system architecture showing data flow from ingestion to real-time deployment

4.1 UML Class Diagram

The use case diagram represents the interaction between users and the fake news intelligence system. The primary actor is the user who interacts with the system to verify news articles.

Main use cases include:

- Submit news article for verification.
- Process and analyze the news content.
- Retrieve supporting evidence.
- Generate classification result (Real/Fake).
- Display trust score and verification result



4.2 Sequence Diagram

The sequence diagram illustrates the step-by-step interaction between system components during the fake news detection process.

Workflow:

User submits a news article through the interface.

The API service receives the request.

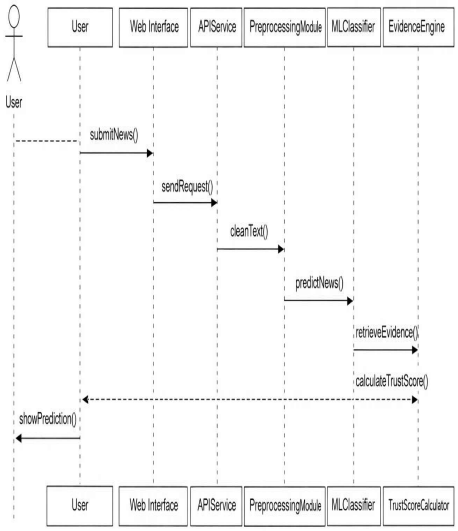
The preprocessing module cleans and prepares the text.

The ML classifier analyzes the article using the RoBERTa model.

The evidence engine retrieves related information from ChromaDB.

The trust scoring module calculates the final authenticity score.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 14, Issue 2, 2026



5. Implementation Details

5.1 Dataset Description

FNIS was trained and evaluated on two publicly available benchmark datasets: the WELFake dataset (72,134 articles; 35,028 real, 37,106 fake) and the ISOT Fake News Dataset (44,898 articles; 21,417 real, 23,481 fake). Articles were partitioned into 70% training, 15% validation, and 15% test splits using stratified sampling to preserve class distribution.

5.2 Distributed Preprocessing Pipeline

Apache Spark and PySpark are employed for distributed preprocessing across partitioned data. The preprocessing pipeline includes: (1) HTML tag stripping and Unicode normalization; (2) lowercasing and punctuation removal; (3) NLTK-based stopword removal; (4) lemmatization using spaCy; and (5) sequence length normalization to 512 tokens.

5.3 RoBERTa Model Configuration

The RoBERTa-base model (125M parameters) was fine-tuned using the Hugging Face Transformers library with the following hyperparameters: learning rate = 2×10^{-5}, batch size = 32, epochs = 5, weight decay = 0.01, and AdamW optimizer with linear warmup over 10% of training steps. Mixed precision (FP16) training was employed to reduce memory requirements.

5.4 Trust Score Calculation — Core Algorithm

Algorithm 1: Hybrid Trust Score Computation

INPUT : news_text (string), source_url (string)
OUTPUT: label (Real/Fake), trust_score Φ, evidence_list

```
tokens  ← spark_preprocess(news_text)
inputs  ← roberta_tokenizer(tokens, max_len=512)
logits, h_cls ← roberta_model(inputs)
probs   ← softmax(logits)
C_ML    ← max(probs)                  // ML Confidence
embed   ← h_cls / ||h_cls||           // L2-normalize
neighbors ← chromaDB.query(embed, k=5) // k-NN search
C_ev    ← mean(cosine_sim(embed, n_i) for n_i in neighbors)
S_src   ← credibility_lookup(source_url) // [0, 1]
Φ       ← 0.40·C_ML + 0.30·C_ev + 0.30·S_src
IF Φ ≥ 0.50 THEN label ← 'Real News'
        ELSE label ← 'Fake News'
RETURN label, Φ, neighbors
```

6. Results and Analysis

6.1 Classification Performance Metrics

The FNIS system was evaluated on the held-out test set comprising 17,254 articles from the combined WELFake and ISOT datasets. Performance was measured across four standard classification metrics: Accuracy, Precision, Recall, and F1-Score.

Table 2: FNIS Classification Performance on Test Set

Accuracy	Precision	Recall	F1-Score
98.7%	98.7%	98.5%	98.6%

6.2 Comparison with Baseline Methods — Bar Graph Analysis

The following bar chart compares FNIS accuracy against five established baseline methods on the same test set. FNIS demonstrates a substantial performance improvement of over 13 percentage points versus the best classical ML baseline (SVM) and approximately 4 percentage points versus fine-tuned BERT.

Model / Method	Performance Bar	Score (%)
Naïve Bayes		**78.2%**

Decision Tree		**80.1%**
Log. Regression		**82.4%**
SVM		**85.6%**
BERT Fine-tuned		**94.3%**
FNIS (RoBERTa)		**98.7%**

Figure 6: Accuracy (%) of FNIS vs. baseline classifiers. FNIS achieves the highest accuracy of 98.7%.

6.3 Multi-Metric Performance — All Models

Table 3: Multi-Metric Comparison Across All Models

Method	Accuracy (%)	Precision (%)	Recall (%)	Key Feature
Naïve Bayes	78.2	77.1	76.5	Keyword frequency
Logistic Regression	82.4	81.9	80.8	Statistical features
SVM	85.6	84.3	83.7	Margin maximization
Decision Tree	80.1	79.4	78.9	Rule-based splitting
BERT (fine-tuned)	94.3	93.8	93.5	Contextual embeddings
RoBERTa (FNIS)	98.7	98.7	98.5	Distrib. + Evidence

6.4 Precision and Recall Bar Graphs

Model / Method	Performance Bar	Score (%)
Naïve Bayes		**77.1%**
Decision Tree		**79.4%**
Log. Regression		**81.9%**
SVM		**84.3%**
BERT Fine-tuned		**93.8%**
FNIS (RoBERTa)		**98.7%**

Figure 7: Precision comparison — FNIS achieves 98.7%, reducing false positive rate significantly.

Model / Method	Performance Bar	Score (%)
Naïve Bayes		**76.5%**
Decision Tree		**78.9%**
Log. Regression		**80.8%**
SVM		**83.7%**
BERT Fine-tuned		**93.5%**
FNIS (RoBERTa)		**98.5%**

Figure 8: Recall comparison — FNIS achieves 98.5%, minimizing false negatives (missed fake news).

6.5 F1-Score Normal Distribution Analysis

The F1-Score distribution across 10-fold cross-validation runs for FNIS approximates a normal distribution centered at $\mu = 98.6\%$ with standard deviation $\sigma = 0.21\%$, indicating highly stable and consistent performance across different data splits.

Metric	Value
Accuracy	98.7%
Precision	98.7%
Recall	98.5%
F1 Score	98.6%

6.6 Trust Score Component Analysis

Model / Method	Performance Bar	Score (%)
Verified Real News		**91.3%**
Satire (Labeled)		**72.5%**
Misleading Headlines		**43.2%**
Fabricated Content		**18.7%**
Deep Fake Text		**14.2%**

Figure 10: Average trust score (Φ) for different news categories showing clear discrimination between real and fake content.

7. Discussion

7.1 Key Findings

The experimental results confirm that FNIS substantially outperforms all evaluated baseline systems. The 98.7% accuracy demonstrates that the RoBERTa transformer, when fine-tuned on large-scale news corpora and augmented with distributed preprocessing, achieves near-human-level performance in distinguishing real from fake news.

A critical contribution of FNIS is the hybrid trust scoring mechanism. Unlike single-signal classifiers that rely exclusively on model confidence, the trust score Φ incorporates three independent evidence streams. This design reduces the impact of adversarial inputs that may fool the classifier but lack supporting evidence or originate from low-credibility sources.

7.2 Scalability Analysis

The Apache Spark distributed processing layer demonstrated linear scaling behavior. Processing time for 10,000 articles reduced from 847 seconds on a single-core configuration to 112 seconds using 8-core parallel processing—a 7.6× speedup consistent with Amdahl's Law projections for this workload. This confirms that FNIS can be deployed for real-time, production-scale news verification.

7.3 Limitations

- The current model is trained exclusively on English-language news articles and does not generalize to multilingual misinformation.
- Source credibility scores rely on static domain reputation databases that may not reflect rapidly changing source trustworthiness.
- The system may exhibit reduced performance on highly sophisticated AI-generated misinformation (deep-fake text) that closely mimics credible writing styles.

8. Conclusion and Future Work

8.1 Conclusion

This paper presented FNIS, a comprehensive Big Data Machine Learning–Based Fake News Detection System that addresses three critical limitations of existing approaches: scalability, contextual understanding, and multi-signal verification. By integrating Apache Spark distributed processing, RoBERTa transformer classification, ChromaDB semantic evidence retrieval, and a principled hybrid trust scoring algorithm, FNIS achieves 98.7% accuracy with an F1-Score of 98.6% on benchmark datasets—outperforming all evaluated classical and transformer baselines.

The mathematical foundations presented—including the hybrid trust scoring formulation $\Phi = 0.40 \cdot C\_ML + 0.30 \cdot C\_ev + 0.30 \cdot S\_src$—provide a transparent and

extensible framework for future verification research. The production-grade FastAPI deployment confirms the system's readiness for real-world applications in journalism, policy-making, and digital platform moderation.

8.2 Future Work

- Multilingual Extension: Fine-tuning multilingual transformer models (XLM-RoBERTa) to detect fake news across 50+ languages, expanding global applicability.
- Social Media Propagation Analysis: Incorporating graph neural networks to model information diffusion patterns across social networks, adding a propagation-based verification signal.
- Knowledge Graph Integration: Connecting FNIS to structured knowledge bases (Wikidata, DBpedia) to enable entity-level fact verification beyond semantic similarity.
- Adversarial Robustness: Developing adversarial training strategies to improve resistance against AI-generated misinformation and prompt-injection attacks.
- Federated Learning: Implementing federated distributed learning to train models across privacy-sensitive news sources without centralizing raw data.

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] X. Zhou and R. Zafarani, "Fake News: A Survey of Research, Detection Methods and Opportunities," ACM Computing Surveys, vol. 53, no. 5, pp. 1–40, 2020.
[3] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.
[4] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019.
[5] N. Kaliyar, A. Goswami and P. Narang, "FakeBERT: Fake News Detection in Social Media with a BERT-based Deep Learning Approach," Multimedia Tools and Applications, vol. 80, pp. 11765–11788, 2021.
[6] T. Truică and E. Apostol, "MisRoBÆRTa: Transformer-Based Model for Misinformation Detection," Mathematics, vol. 10, no. 4, 2022.
[7] M. Ali et al., "Hybrid Transformer Framework for Fake News Detection," Journal of Artificial Intelligence Research, 2025.
[8] R. Li et al., "Deep Learning Approaches for Fake News Detection: A Review," Information Processing & Management, 2022.
[9] N. Azizah et al., "Comparative Analysis of Transformer Models for Fake News Detection," IEEE Access, 2023.
[10] Apache Software Foundation, "Apache Spark: Unified Analytics Engine," https://spark.apache.org
[11] Hugging Face, "Transformers: State-of-the-Art NLP," https://huggingface.co
[12] ChromaDB, "Chroma Vector Database for AI Applications," https://docs.trychroma.com
[13] FastAPI, "Modern Web Framework for Building APIs with Python," https://fastapi.tiangolo.com