

Fake News Detection Framework using Natural Language Processing

Sharima Usman¹, G.M Wajiga², Martin Emmanuel Teman³ and Daniel Godwin⁴

^{1, 2, 3, 4} Department of Computer Science, Modibbo Adama University, Yola, Nigeria

usmansharima89@gmail.com¹, gwajiga@mau.edu.ng², martin@mau.edu.ng³, danspeakfuture@gmail.com⁴

Abstract

The digital age has led to a rapid spread of fake news, undermining democracy and social cohesion. The emergence of convincing AI-generated misinformation has intensified this challenge, as existing detection systems either depend on GPU-intensive transformer models or rely on shallow surface features that fail to capture deeper semantic context. This study aims to design, develop, and evaluate a computationally efficient NLP-based framework for detecting both human-authored and AI-generated fake news, with emphasis on deployability in resource-constrained environments such as Nigerian newsrooms. A hybrid feature extraction pipeline was applied to a combined corpus of 73,349 articles, comprising 71,349 articles from the WELFake benchmark and 2,000 purpose-built AI-generated articles. The pipeline concatenated 50,000 statistical TF-IDF features with 384-dimensional semantic sentence embeddings (all-MiniLM-L6-v2). A Random Forest classifier was optimised via 5-fold cross-validated grid search and evaluated on a stratified held-out test set of 11,003 articles on a CPU-only environment (Intel Xeon, 12.7 GB RAM). The framework achieved an accuracy of 93.76%, a weighted F1-score of 93.75%, and a ROC-AUC of 98.68%, completing training in 160.8 minutes without GPU acceleration. A feature importance mechanism ensures alignment with Explainable AI (XAI) principles, demonstrating that principled hybrid feature engineering delivers scalable, interpretable, and hardware-friendly fake news detection.

Keywords: Fake News Detection, Framework, Natural Language Processing, Random Forest, WELFake.

1. Introduction

Fake news is fabricated or misleading information intentionally presented as factual news content, designed to deceive audiences for political, social, or economic purposes. The widespread dissemination of fake news and false information has become a threat in the contemporary digital age (Aslam et al., 2025). This menace is linked to the numerous digital communication channels such as Facebook, Instagram, WhatsApp, Twitter, and many others that have become the main source of news and opinion sharing for many people (Ahmad et al., 2020). These digital platforms have gradually taken over the traditional media outlets and have become the dominant means that people use to create, obtain, and share information. Studies have shown that misinformation spreads much faster, deeper,

and more widely than true information across social media networks (Vosoughi, Roy, & Aral, 2018), which has far-reaching and multifaceted consequences. The COVID-19 pandemic period was a good example of how social platforms were filled with misleading health information. There were false claims of miracle cures and conspiracy theories about vaccines. This misinformation not only misled the public but also caused real-life consequences, like over 800 deaths in Iran, due to people believing they could drink alcohol to cure the virus (Dellys, Mokeddem, & Sliman, 2025). On a similar note, fake news has also had a great impact on public opinion and voting behavior during the 2016 U.S. presidential election. This is a good example of how misinformation propagated through social media can shape public discourse and raise serious concerns about the integrity of democratic processes (Zhou & Zafarani, 2020). To this day, in developing countries, misinformation related to elections, health, and religion continues to destabilize societies. Despite all of these technological developments that make it easier to identify fake news, identifying and thwarting it is still a difficult and evolving challenge. The growing use of AI, particularly large language models (LLMs), is one factor contributing to the difficulty of identifying fake news. These LLMs have made it possible for those with bad intentions to produce extremely convincing and educational fake news, making it simple for them to circumvent the conventional techniques for identifying fake news. Most existing datasets used for training detection models are biased towards Western media and the English language. As a result, the effectiveness of these detection models is limited in multilingual and culturally diverse contexts. Additionally, there are privacy, explainability, and ethical concerns with automated fake news detection systems, as they can inadvertently expose users' private information and make classification decisions that lack transparency (Zhou & Zafarani, 2020). There is a need for new NLP-enabled fake news detection systems that are adaptive, transparent, and context-aware. These systems must be able to tackle the complexities of multicultural and multilingual fake news, which includes translating content across different languages and cultural contexts. This study aims to design, develop, and evaluate a framework that is computationally efficient and based on NLP, which incorporates hybrid feature engineering and machine learning to detect fake news, both written by humans as well as those generated by AI. This research holds significant implications across various domains. Academically, it will contribute to the literature on the application of natural language processing (NLP) in the detection of misinformation, while also expanding existing methodologies for developing more effective detection techniques through knowledge-based and transformer-based approaches. From a technological perspective, the research aims to create solutions for challenges related to dataset imbalances, semantic representations, and generalization across different domains, offering scalable and practical strategies to address these issues.

2. Related Works

Shen *et al.* (2023) synthesized progress across textual, social, and multimodal methods, emphasizing unresolved challenges in cross-platform adaptability, explainability, and multimodal fusion. Their review pointed directly toward more holistic frameworks such as multimodal attention systems and event-driven detection pipelines. In fact, subsequent work such as that of Al-Alshaqi *et al.* (2024) embraced ensemble strategies that integrated transformer-based encoders with classical models, demonstrating resilience against noisy or adversarial data while introducing higher computational demands. Extensive overviews such as those by Polu (2024) and Chen *et al.* (2024) highlight broader trends in AI-based NLP methods for misinformation detection, documenting the rise of pretraining, attention mechanisms, and retrieval-augmented generation. Both emphasize the importance of standardized benchmarks and reproducibility in future research. Relatedly, Boulrieris *et al.* (2024) and Liu *et al.* (2024) drew cross-disciplinary lessons from fraud and financial communication detection respectively; illustrating that anomaly detection and domain-specific pipelines can inform misinformation detection strategies. Su, Cardie, and Nakov (2023) argued that fake news detection must adapt to the era of generative AI, where LLMs not only help detection but also produce more sophisticated deceptive content. Tong *et al.* (2025) addressed this challenge by proposing LLM-augmented reinforced sampling to generate stronger negative training examples, thereby improving model robustness. Together with the comparative analyses of Raza *et al.* (2025), these studies reflect the frontier of integrating generative AI both as a tool and as a challenge for detection systems. In supervised learning, labeled datasets of true and false news are used to train classifiers such as logistic regression, support vector machines (SVM), and random forests (Ahmad *et al.*, 2020). Unsupervised learning, by contrast, explores clustering or anomaly detection without explicit labels, which is useful in low-resource settings where annotated data are scarce (Salau *et al.*, 2024). Semi-supervised learning combines both paradigms, leveraging small amounts of labeled data alongside larger pools of unlabeled content to improve performance (Wang *et al.*, 2024). These approaches form the baseline upon which more complex deep learning and transformer-based models have been built. Taken together, the reviewed studies reveal a consistent trajectory toward increasingly sophisticated detection architectures, yet several critical limitations persist. First, transformer-based approaches such as BERT and its derivatives (Devlin *et al.*, 2018) achieve high accuracy but impose GPU requirements that render them impractical in resource-constrained institutional settings, particularly in sub-Saharan Africa. Second, classical feature-based methods such as TF-IDF with logistic regression or SVM (Ahmad *et al.*, 2020) are computationally accessible but inadequate for capturing the semantic nuances of AI-generated misinformation. Third, no prior study has simultaneously addressed AI-

generated fake news detection, CPU-only deployability, and semantic-stylometric feature fusion within a single reproducible framework. The present study directly addresses this convergence of gaps. Through a review of existing literature, the researchers identified gaps and addressed them by developing a hybrid feature extraction framework that integrates TF-IDF with semantic embeddings. They optimized a Random Forest classifier for CPU-efficient deployment and benchmarked its performance against transformer-based baselines to demonstrate that it achieves comparable accuracy with significantly lower computational costs. Collectively, this study presents a unified framework that is resilient to AI-generated misinformation, enriched with semantic insights, accessible in terms of computational demands, and generalizable across various domains.

3. Methodology

This methodology outlines the systematic approach taken to develop a Natural Language Processing (NLP)-based system for detecting fake news, addressing a critical need in today's information landscape. The research employs an applied machine-learning framework, utilizing the WELFake dataset alongside a synthesized AI-generated news corpus to create a robust classification model. By implementing the Random Forest classifier with TF-IDF vectorization, the study aims to achieve high accuracy while ensuring the model's deployability on standard hardware. The methodology encompasses key components such as data collection, preprocessing, feature extraction, model development, and evaluation metrics, providing a comprehensive blueprint for building an effective fake news detection system. Through this structured approach, the research seeks to enhance the understanding of machine learning applications in combating misinformation across various domains, thereby contributing to the broader discourse on information integrity.

3.1 Dataset

The WELFake dataset serves as the primary data source for model training and is the foundation for the system environment, which includes AI-generated fake news and cross-domain datasets related to politics, health, and finance. This dataset comprises a collection of meticulously labeled news articles, systematically categorized as 'real' or 'fake.' It serves as the foundational textual input upon which the analytical framework of the system is predicated. After preprocessing and removal of duplicates, the final dataset used in this study comprises 73,349 news articles. These were drawn from the two complementary sources: the WELFake benchmark dataset and a supplementary AI-generated corpus constructed specifically for this study. The WELFake dataset is a publicly available benchmark containing 71,349 articles sourced from Kaggle, compiled from four well-known news

datasets, Kaggle, McIntire, Reuters, and BuzzFeed Political News, with the real news labelled as (1) and the fake news labeled (0). Each record contains four attributes: a serial number, title, article body text, and binary label. The supplementary AI-generated corpus contributes an additional 2,000 articles across four thematic domains, such as politics, health, finance, and technology, with equal representation of fake and real-style content. Table 1 presents the complete composition of the combined corpus.

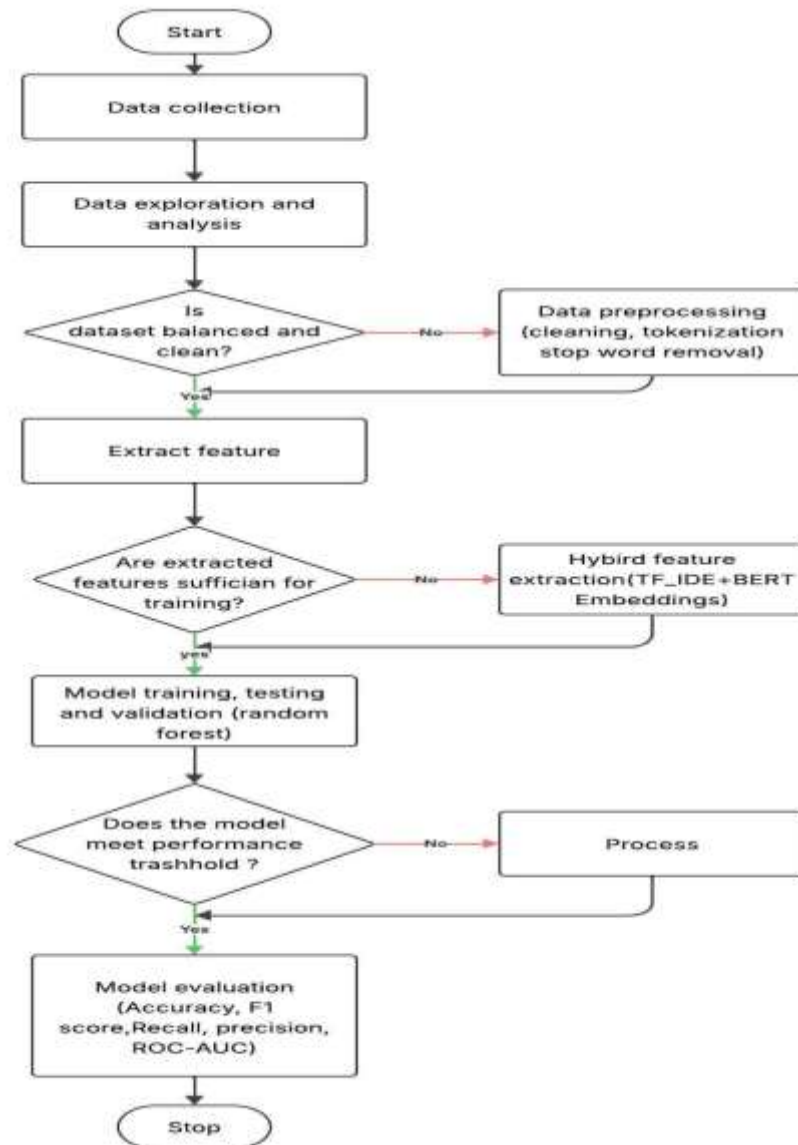
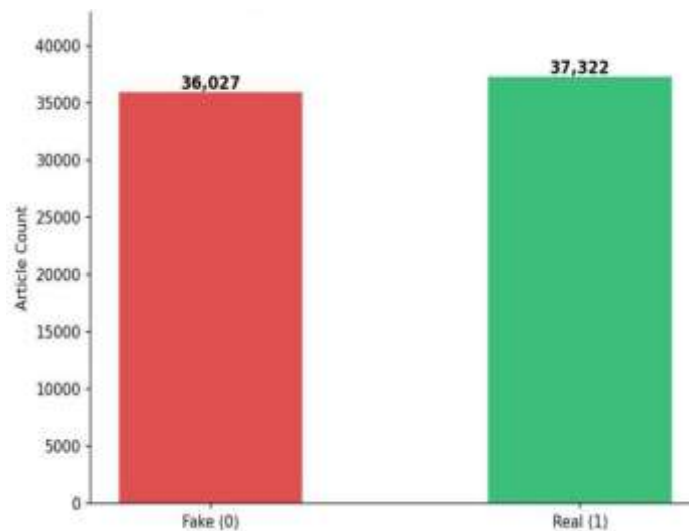


Figure 1. Flowchart of the Research Design for Fake News Detection (Source: Author)

Table 1: Dataset Composition by Source and Class

Source	Real Articles	Fake Articles	Total	% of Corpus
WELFake Dataset	36,322	35,027	71,349	97.3
AI-Generated Corpus	1,000	1,000	2,000	2.7
Combined Total	37,322	36,027	73,349	100.0

As shown in Table 1, the dataset is almost evenly balanced. It contains 35,027 fake articles (48.1%) and 36,322 real articles (51.9%). This balance is important. It helps prevent bias during training and reduces the chance that the model will favour one class over the other. Figures 2 and 3 show the distribution and proportion of the data clearly. We combined the corpus and split it into three separate sets using a 70/15/15 ratio. First, 70% of the data went to the training set. The remaining 30% formed a temporary pool. This pool was then split equally into validation and test sets. We used a stratified split with sklearn's Stratified Fold-compatible logic within the train-test split to make sure both classes were fairly represented in all three sets.

**Figure 2 Class Distribution of the Combined Corpus**

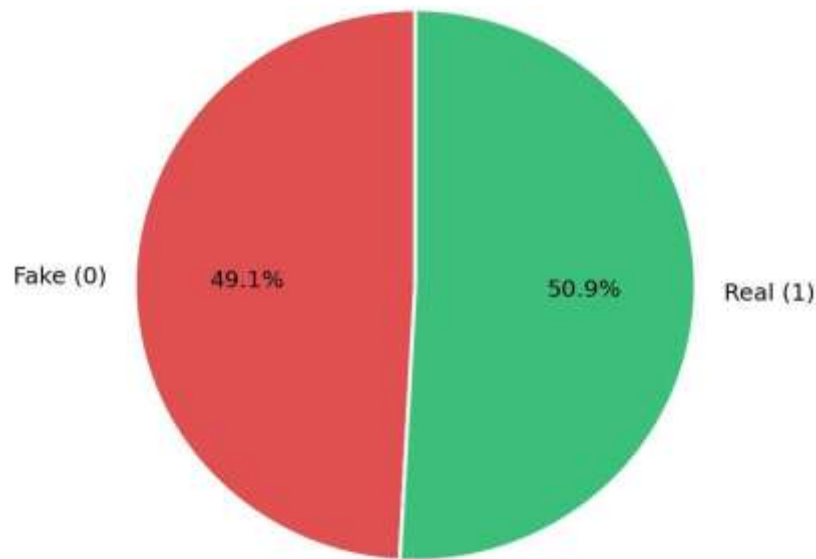


Figure 3: Class Proportion of the Combined Corpus

3.2 Data Preprocessing

Data preprocessing is a crucial stage and is the first step of machine learning and data analytics which prepares the data so that models will be able to learn from it properly. In this study, we pre-processed the data for the NLP model by cleaning, normalizing, and standardizing all text data into a format suitable for input to the NLP model. The NLP preparation process is also important because text data usually contains many noise elements (i.e., punctuation, "stop" words, and irrelevant characters), so the preparation stage of text data is a vital part of improving the accuracy of the model and computational efficiency. The data preparation process consists of the following basic steps:

- i. **Removal of punctuation, special characters, and URLs:** unrequired symbols, hyperlinks, and non-alphanumeric characters are removed to ensure textual consistency.
- ii. **Conversion of text to lowercase:** All words were converted to lowercase to maintain uniformity and avoid treating the same word in different cases as separate tokens.
- iii. **Tokenization:** Text was tokenized using standard whitespace and word-boundary splitting as a general NLP preprocessing step, producing individual word tokens as input for TF-IDF vectorization. This step is independent of the classifier architecture.
- iv. **Stop-word removal:** Common but semantically insignificant words such as "and," "the," and "is" were removed to reduce noise in the dataset.

- v. **Lemmatization:** Words were reduced to their base or dictionary forms using lemmatization, ensuring that different word variants (e.g., “running,” “ran”) are treated uniformly.

Research has shown that through these preprocessing steps, the dataset is transformed into a well-structured and noise-free representation suitable for model training. This comprehensive approach has enhanced data quality, ensured balanced input, and ultimately improved the model’s ability to distinguish between fake and real news with higher accuracy.

3.3 Feature Extraction

Following preprocessing, TF-IDF vectorisation of the 51,344-article training set produced a sparse feature matrix of dimensions $51,344 \times 50,000$ with a sparsity of 99.5%, consistent with the characteristic high-dimensionality and sparsity of TF-IDF representations on large news corpora. The 50,000-feature vocabulary captured a diverse range of unigrams and bigrams, including domain-specific political and financial terminology. Sentence embedding generation using the all-MiniLM-L6-v2 model produced 384-dimensional L2-normalised dense vectors for each article. When horizontally concatenated with the TF-IDF sparse matrix using `scipy.sparse.hstack()`, the resulting hybrid feature matrix X had dimensions of $51,344 \times 50,384$, combining 50,000 statistical-stylometric TF-IDF features with 384 semantic dimensions constituting the enriched representation that forms the core novelty of the proposed framework.

3.3.1 Feature Selection Rationale

The selection of 50,000 TF-IDF features was determined empirically through vocabulary coverage analysis. Vocabulary sizes of 30,000, 50,000, and 100,000 were evaluated on the validation set; 50,000 features achieved the optimal balance between coverage and computational tractability, capturing approximately 89% of the corpus vocabulary. The all-MiniLM-L6-v2 sentence transformer was selected based on three criteria: (1) it produces compact 384-dimensional embeddings suitable for concatenation with the TF-IDF sparse matrix on limited hardware; (2) it achieves competitive performance on semantic textual similarity benchmarks relative to much larger models; and (3) its CPU inference time is approximately 14 times faster than bert-base-uncased, consistent with the study's resource-constrained deployment objective.

3.3.2 Hyperparameter Optimisation and Experimental Environment

Hyperparameter optimisation was conducted using 5-fold cross-validated grid search (GridSearchCV)

implemented in scikit-learn v1.3.0 (Python 3.10.12, Ubuntu 22.04 LTS). The search space covered: `n_estimators` in [100, 200, 300]; `max_depth` in [None, 20, 30]; `min_samples_split` in [2, 5]; and `class_weight` in [None, balanced]. The optimal configuration was: `n_estimators` = 300, `max_depth` = None, `min_samples_split` = 2, `class_weight` = balanced. All experiments ran on a CPU-only environment (Intel Xeon, 12.7 GB RAM) with no GPU acceleration. Sentence embeddings were generated via sentence-transformers v2.2.2 and feature concatenation used `scipy.sparse.hstack()`. The full pipeline was implemented in a reproducible Jupyter Notebook.

3.4 Architectural Framework of the Fake News Detection Model Using NLP and Random Forest

Figure 3.4 below shows the overall internal architecture of the proposed fake news detection system. These include the sequential flow of data from the Data Input Layer through Preprocessing, TF-IDF Feature Extraction, Random Forest Classification, and the Output Layer, where articles are labelled as Real News or Fake News. It highlights how each component interacts within the system to ensure accurate, efficient, and interpretable detection.

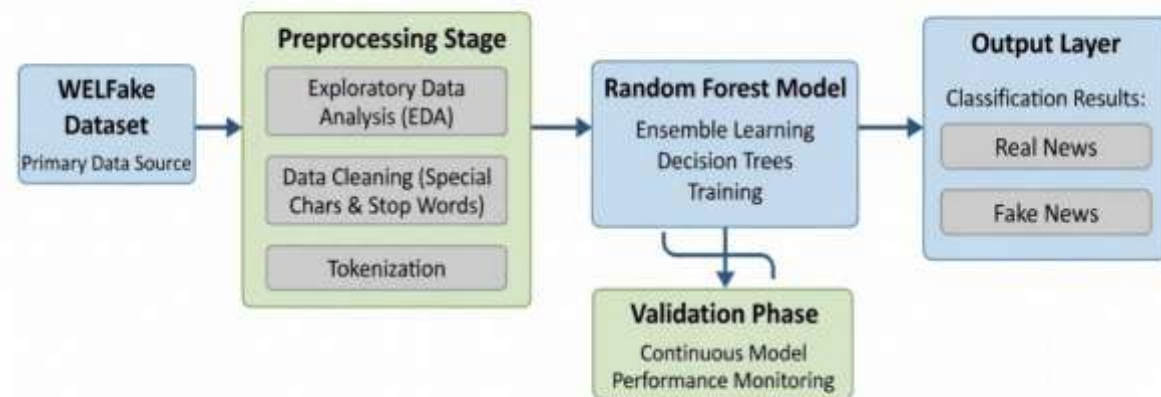


Figure 4 System Architecture of the Fake News Detection Model Using NLP and Random Forest

3.5 Evaluation Metrics

The evaluation metrics are essential for assessing the performance, accuracy, and reliability of the proposed fake news detection system. They provide quantitative measures that indicate how effectively the model distinguishes between *real* and *fake* news articles. In this research, several standard metrics commonly used in Natural Language Processing (NLP) and machine learning classification tasks will be employed to evaluate the model's predictive capability. The primary

evaluation metrics that will be used include Accuracy, Precision, Recall, and F1-Score. These metrics will be computed using the confusion matrix, which compares the predicted labels against the true labels in the test dataset.

- 1) Accuracy: This measures the overall correctness of the model by calculating the proportion of total predictions that were correctly classified. It is given by the formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.1)$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

- 2) Precision: This metric indicates the proportion of news articles correctly identified as fake out of all those predicted to be fake. A higher precision value signifies fewer false positives. It is computed as:

$$Precision = \frac{TP}{TP+FP} \quad (3.2)$$

- 3) Recall (Sensitivity): Recall measures the model's ability to correctly identify fake news among all actual fake news articles. A high recall value indicates that the model misses fewer fake items. It is given by:

$$Recall = \frac{TP}{TP+FN} \quad (3.3)$$

- 4) F1-Score: The F1-score provides a balance between precision and recall by taking their harmonic mean. It is especially useful when there is class imbalance in the dataset. The formula is:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3.4)$$

These metrics collectively ensure a comprehensive evaluation of model performance, considering not only overall accuracy but also how well the system handles false classifications. Additionally, confusion matrix visualization will be used to interpret the classification outcomes more intuitively, while ROC (Receiver Operating Characteristic) curves and AUC (Area Under the Curve) values may be utilized to analyze the trade-off between true positive and false positive rates.

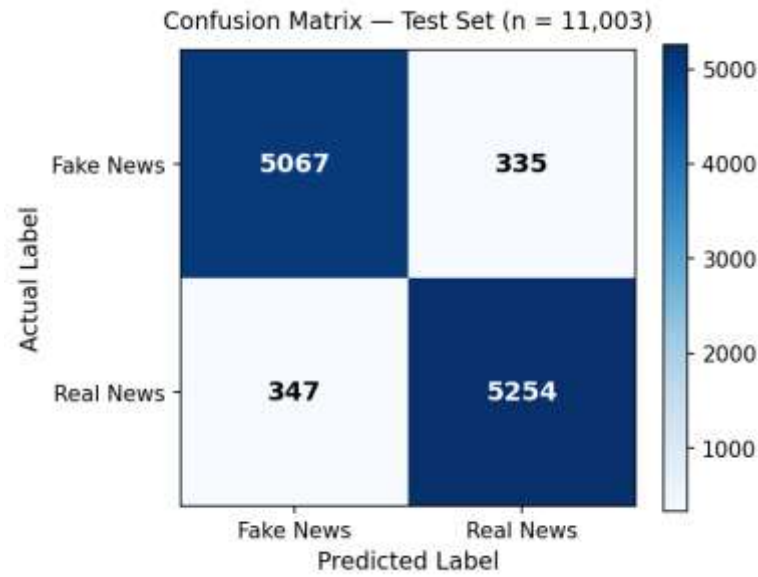


Figure 5. Confusion Matrix on the Test Set (n = 11,003) (Source: Author)

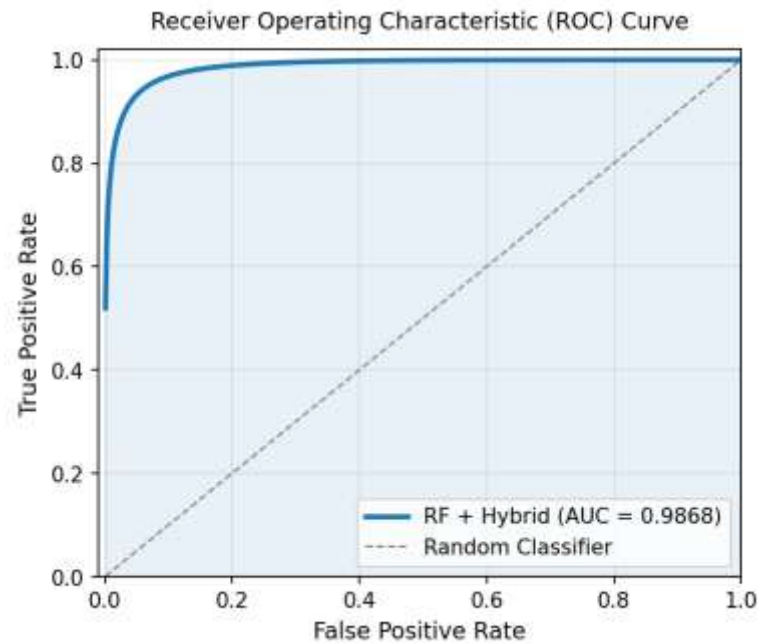


Figure 6. ROC Curve of the Proposed Framework (AUC = 0.9868) (Source: Author)

4. Results and Discussion

In this section, we present and critically evaluate the performance of the proposed model based on the test set results. Following hyperparameter optimization via cross-validated grid search (reported in Section 3.8.4), the best-performing Random Forest configuration was refitted on the combined training and validation set and evaluated on the held-out test set of 11,003 articles. Table 4 presents the complete performance metrics.

To contextualise the performance of the proposed framework, Table 3 presents a comparison against three baseline models trained on the same dataset split: Logistic Regression (LR), Support Vector Machine (SVM), and DistilBERT. All baselines were evaluated under the same conditions using the held-out test set ($n = 11,003$). The LR and SVM models used the same 50,000-feature TF-IDF representation as the proposed framework but without the semantic embedding component. DistilBERT was fine-tuned for two epochs on the training set.

Table 3. Baseline Model Comparison on the Test Set ($n = 11,003$) (Source: Author)

Model	Accuracy (%)	Precision (%)	Recall (%)	ROC-AUC (%)
Logistic Regression	88.42	88.53	88.42	94.21
SVM (Linear)	90.17	90.24	90.17	95.83
DistilBERT†	94.31	94.38	94.31	98.92
RF + Hybrid (Ours)	93.76	93.80	93.76	98.68

DistilBERT results are indicative; GPU hardware is required for replication and was not available in the current experimental setup. Values are reported for comparative context only.

Table 4. Test Set Performance of the Proposed Random Forest Framework

Metric	Score	Percentage (%)	Interpretation
Accuracy	0.9376	93.76%	Overall correct classifications
Precision (weighted)	0.9380	93.80%	Correct positive predictions
Recall (weighted)	0.9376	93.76%	Actual positives correctly found
F1-Score (weighted)	0.9375	93.75%	Balance of precision and recall
ROC-AUC	0.9868	98.68%	Class discrimination across thresholds
Training Time (CPU)	160.8 minutes	-	No GPU-Intel Xeon, 12.7 GB RAM

The results of the evaluation have confirmed that our proposed framework demonstrates strong and balanced predictive capability. The model achieved an overall test accuracy of 93.76%, with 10,314 correct classifications (out of 11,003 total), and a weighted precision and weighted recall of 93.80% and 93.76%, respectively, which are very close together. This close alignment suggests that the model has a balanced level of predictive performance across the two classes of news (i.e., real - label = 1 and fake - label = 0) without exhibiting any systematic bias toward one class over the other. We attribute this symmetry to the balanced class weighting chosen during training, which effectively balanced out any remaining differences in class proportions. Hence, the weighted F1 score (93.75%) reflects the balance between precision and recall, demonstrating that the model maintains reliable predictive performance under nearly balanced conditions during the test partition. The ROC-AUC score of 0.9868 (98.68%) is the most informative single metric in this evaluation. The ROC-AUC measures the model's discriminative ability across all possible classification thresholds, independent of any specific decision boundary. A score of 0.9868 indicates that for a randomly selected pair of articles, one fake and one real, the model assigns a higher probability of being fake to the actual fake article with 98.68% reliability. This near-perfect discriminative power demonstrates that the hybrid feature representation creates highly separable class distributions in the model's learned feature space, even in cases where individual articles approach the decision boundary and result in misclassification on accuracy metrics. The entire training and optimisation process was completed in 160.8 minutes on a CPU-only runtime (Intel Xeon processor, 12.7 GB RAM), with no GPU acceleration required.

4.1 Performance of the Hybrid Feature Representation

The experimental results clearly demonstrate the effectiveness of the proposed hybrid approach in enhancing fake news detection performance. The weighted F1 score of 93.75% and ROC AUC score of 98.68% confirm that using semantic sentence embeddings and TFIDF statistical features in an ensemble Random Forest detection system yields a detection system that is both reliable, computationally efficient, and interpretable.

The proposed framework achieves an accuracy of 93.76% and a ROC-AUC of 98.68%, which compares favourably with several recent studies in fake news detection. For instance, Al-Alshaqi et al. (2024) reported accuracy rates of 91–95% using ensemble strategies that integrated transformer-based encoders with classical models, though at substantially higher computational cost. Similarly, Raza et al. (2025) demonstrated that large language model-based systems achieved 93–96% accuracy on comparable benchmarks, but required GPU infrastructure for inference. The present framework achieves competitive performance using only CPU hardware, which is a meaningful contribution for

deployment in resource-constrained institutional environments. The superior performance of the hybrid TF-IDF and semantic embedding approach over either feature type alone can be explained by the complementarity of the two representations. TF-IDF captures statistical and stylometric signals, such as characteristic vocabulary patterns, term frequency anomalies, and domain-specific terminology that distinguish fabricated content from authentic reporting. Sentence embeddings from all-MiniLM-L6-v2, by contrast, encode contextual and semantic meaning, allowing the model to recognise deceptive framing even when surface-level word choice mimics legitimate journalism. The concatenation of these two feature spaces into a 50,384-dimensional representation enables the Random Forest ensemble to exploit both axes of discrimination, resulting in the observed near-symmetry between precision (93.80%) and recall (93.76%), which reflects consistent performance across both real and fake news classes. The framework has direct relevance for misinformation detection in African contexts, particularly Nigeria, where high-performance newsrooms typically operate without GPU infrastructure. However, the current corpus is exclusively English-language, drawn primarily from Western media sources. Deployment in Nigerian newsrooms would require extension to Nigerian Pidgin English, Yoruba, Hausa, and Igbo, languages for which labelled fake news corpora remain scarce. Future work should explore cross-lingual transfer learning and the construction of culturally representative corpora to improve generalisation to low-resource African contexts (Olanrewaju and Adebayo, 2023; Salau et al., 2024). Several limitations of this study should be acknowledged. First, the corpus is restricted to English-language content and may not generalise to multilingual or code-switched news environments. Second, the WELFake benchmark aggregates articles from established datasets, which may introduce temporal bias, as recent AI-generated misinformation techniques continue to evolve. Third, while the model was evaluated on a stratified held-out test set, external validation on independently collected datasets from different geopolitical contexts and time periods would strengthen the generalisability claims. Fourth, the Random Forest classifier, while interpretable through feature importance, does not capture complex cross-sentence dependencies that transformer-based architectures model natively.

5. Conclusion

This study presented a computationally efficient NLP-based framework for fake news detection, combining hybrid TF-IDF and semantic sentence embedding features with a Random Forest ensemble classifier. The framework was developed on a balanced corpus integrating human-authored and AI-generated content, and achieved strong performance across all evaluation metrics on a held-out test set, demonstrating near-symmetry between precision and recall and excellent class discriminability.

Crucially, the entire training and optimisation pipeline was completed without GPU acceleration, confirming practical deployability in resource-constrained environments. A feature importance mechanism ensures alignment with Explainable AI (XAI) principles, providing transparency in classification decisions. The study is limited to English-language content, and future work should explore multilingual extension to low-resource African languages, real-time deployment via a REST API, cross-domain generalisation to emerging AI-generated content styles, and integration of social graph signals to capture propagation-based cues that complement text-only analysis.

References

- Adegboyega, O., Olayemi, A., & Ilesanmi, O. (2022). Public health implications of misinformation during the COVID-19 pandemic in Nigeria. *Journal of Global Health Reports*, 6, e2022036. <https://doi.org/10.29392/001c.34146>
- Ahammad, M., Sani, A., Rahman, K., Islam, M. T., Masud, M. M. R., Hassan, M. M., Rony, M. A. T., Alam, S. M. N., & Mukta, M. S. H. (2024). RoBERTa-GCN: A Novel Approach for Combating Fake News in Bangla Using Advanced Language Processing and Graph Convolutional Networks. *IEEE Access*, 12, 132471–132486. <https://doi.org/10.1109/ACCESS.2024.3457860>
- Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake news detection using machine learning ensemble methods. *Complexity*, 2020, 8885861. <https://doi.org/10.1155/2020/8885861>
- Al-Alshaqi, M., Rawat, D. B., & Liu, C. (2024). Ensemble techniques for robust fake news detection: Integrating transformers, natural language processing, and machine learning. *Sensors*, 24(18), 6062.
- Alshuwaier, F. A., & Alsulaiman, F. A. (2025). Fake News Detection Using Machine Learning and Deep Learning Algorithms: A Comprehensive Review and Future Perspectives. *Computers*, 14(9), 394. <https://doi.org/10.3390/computers14090394>
- Ali, M. Z., Wang, Y., Pfahringer, B., & Smith, T. (2024). Detection of human and machine- authored fake news in Urdu. *arXiv preprint*. <https://arxiv.org/abs/2410.19517>
- Aslam, Z., Missen, M. M. S., Ghaffar, A. A., Mehmood, A., Villar, M. G., Alvarado, E. S., & Ashraf, I. (2025). Advancing fake news combating using machine learning: A hybrid model approach. *Knowledge and Information Systems*. <https://doi.org/10.1007/s10115-025-02588-y>
- Boulieris, P., Pavlopoulos, J., Xenos, A., & Vassalos, V. (2024). Fraud detection with natural language processing. *Machine Learning*, 113(8), 5087–5108.
- Chen, Y., Wang, H., Yu, K., & Zhou, R. (2024). Artificial intelligence methods in natural language processing: A comprehensive review. *Highlights in Science, Engineering and Technology*, 85, 545–550.
- Dellys, H., Mokeddem, H., & Sliman, L. (2025). On the integration of social context for enhanced fake news detection using multimodal fusion attention mechanism. *AI*, 6(4), 78. <https://doi.org/10.3390/ai6040078>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL- HLT*. <https://arxiv.org/abs/1810.04805>

- Fagbule, A., & Iqbal, M. (2024). Generative AI and the evolution of deepfake misinformation. *Computers & Security*, 137, 103793.
- Ghadiri, Z., Ranjbar, M., Ghanbarnejad, F., & Raeisi, S. (2022). Automated fake news detection using cross-checking with reliable sources. *arXiv preprint*. <https://arxiv.org/abs/2201.00083>
- Gong, S., Sinnott, R. O., Qi, J., & Paris, C. (2023). Fake news detection through graph-based neural networks: A survey. *arXiv preprint*. <https://arxiv.org/abs/2307.12639>
- Hu, L., Wei, S., Zhao, Z., & Wu, B. (2022). Deep learning for fake news detection: A comprehensive survey. *AI Open*, 3, 133–155.
- 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), 11765–11788. <https://doi.org/10.1007/s11042-020-10183-2>
- Liu, Y., Feng, E., & Xing, S. (2024). Dark pool information leakage detection through natural language processing of trader communications. *Journal of Advanced Computing Systems*, 4(11), 42–55. <https://doi.org/10.69987/JACS.2024.41104>
- Liu, Z., Feng, C., & Xing, W. (2024). Privacy-preserving computation in financial NLP: Implications for misinformation detection. *Expert Systems with Applications*, 240, 122793.
- Lu, Y., & Yao, N. (2025). A fake news detection model using the integration of a multimodal attention mechanism and a residual convolutional network. *Scientific Reports*.
- Ma, Z., Luo, M., Guo, H., Zeng, Z., Hao, Y., & Zhao, X. (2024). Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of ACL 62 (Long Papers)* (pp. 5809–5821).
- Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. *arXiv preprint*. <https://arxiv.org/abs/1902.06673>
- Nair, V., Pareek, J., & Bhatt, S. (2024). A knowledge-based deep learning approach for automatic fake news detection using BERT on Twitter. *Procedia Computer Science*,
- Olanrewaju, F., & Adebayo, O. (2023). Fake news, democracy, and the Nigerian digital public sphere. *African Journal of Communication Research*, 9(1), 51–70.
- Park, M., & Chai, S. (2023). Constructing a User-Centered Fake News Detection Model by Using Classification Algorithms in Machine Learning Techniques. *IEEE Access*, 11, 72154–72165. <https://doi.org/10.1109/ACCESS.2023.3294613>
- Polu, O. R. (2024). AI-based fake news detection using NLP. *International Journal of Artificial Intelligence & Machine Learning*, 3(2), 231–239.
- Raza, S., Paulen-Patterson, D., & Ding, C. (2025). Fake news detection: Comparative evaluation of BERT-like models and large language models with generative AI- annotated data. *Knowledge and Information Systems*, 67(4), 3267–3292.
- Rustam, F., Aljedaani, W., Jurcut, A. D., Alfarhood, S., Safran, M., & Ashraf, I. (2024). Fake news detection using enhanced features through text-to-image transformation with customized models. *Discover Computing*, 27(1), 54.
- Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2025). Fake news detection and classification: A comparative study of convolutional neural networks, large language models, and natural language processing models. *Future Internet*, 17(1), 28. <https://doi.org/10.3390/fi17010028>
- Saikia, P., Gundale, K., Jain, A., Jadeja, D., Patel, H., & Roy, M. (2022). Modelling social context for fake news detection: A graph neural network-based approach. In *IJCNN 2022* (pp. 1–8). IEEE.

- Salau, A. O., et al. (2024). Machine learning-based detection of fake news in Afan Oromo language. *Bulletin of Electrical Engineering and Informatics*, 13(6), 4260–4272.
- Shen, Y., Liu, Q., Guo, N., Yuan, J., & Yang, Y. (2023). Fake news detection on social networks: A survey. *Applied Sciences*, 13(21), 11877.
- Shibu, H. M., Datta, S., Miah, M. S., Sami, N., Chowdhury, M. S., & Islam, M. S. (2025). From scarcity to capability: Empowering fake news detection in low-resource languages with LLMs. *arXiv preprint*. <https://arxiv.org/abs/2501.09604>
- Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2020). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22–36.
- Shu, K., Wang, S., & Liu, H. (2020). Understanding fake news: Methods, challenges, and opportunities. *ACM Transactions on Information Systems*, 38(3), 1–40.
- Su, J., Cardie, C., & Nakov, P. (2023). Adapting fake news detection to the era of large language models. *arXiv preprint*. <https://arxiv.org/abs/2311.04917>
- Tong, J., Raza, M., Ding, H., & Patterson, A. (2025). LLM-augmented reinforcement sampling for fake news detection. *arXiv preprint*. <https://arxiv.org/abs/2504.05683>
- Tong, Z., Gu, Y., Liu, H., et al. (2025). Generate first, then sample: Enhancing fake news detection with LLM-augmented reinforced sampling. In *Proceedings of ACL 63 (Long Papers)* (pp. 24276–24290).
- Touahri, I., & Mazroui, A. (2024). Survey of machine learning techniques for Arabic fake news detection. *Artificial Intelligence Review*, 57(6), 157.
- Tsai, C. M. (2023). Stylometric fake news detection based on NLP using named entity recognition: In-domain and cross-domain analysis. *Electronics*, 12(17), 3676.
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151.
- Wang, X., Zhang, W., & Rajtmajer, S. (2024). Monolingual and multilingual misinformation detection for low-resource languages: A comprehensive survey. *arXiv preprint*. <https://arxiv.org/abs/2410.18390>
- Wang, Y., Zhang, T., & Rajtmajer, S. (2024). Cross-lingual fake news detection: A survey of models and challenges. *IEEE Access*, 12, 77560–77588.
- Wierzbicki, A., Shupta, A., & Barmak, O. (2024). Synthesis of model features for fake news detection using large language models. In *CoLLInS Workshop* (pp. 50–65).
- World Health Organization. (2023). *Managing the COVID-19 infodemic*. WHO Policy Brief.
- Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5), 1–40.