International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

A Multi-Model Learning Framework for Fake News Detection on Social Media

[1]Mrs Priyal Verma

[1] Research Scholar, Department of Computer Science, Madhyanchal professional university Bhopal

ABSTRACT

The rapid proliferation of user-generated content on social media platforms has transformed the way information is created, shared, and consumed. While this democratization of information has clear benefits, it has also enabled the large-scale dissemination of fake news, which poses serious threats to public health, political stability, financial markets, and social cohesion. Automatic fake news detection has therefore become an active and urgent research problem in computer science. This paper proposes a multi-model learning framework that combines the complementary strengths of classical machine learning classifiers, deep sequential neural networks, and transformer-based contextual language models to detect fake news on social media. The framework integrates lexical, semantic, and contextual features through a weighted soft-voting ensemble that fuses predictions from a Support Vector Machine, a Bidirectional Long Short-Term Memory network, and a fine-tuned BERT encoder. Experiments were conducted on three widely used public benchmark datasets, namely LIAR, FakeNewsNet, and ISOT, comprising more than one hundred thousand labeled news statements and articles. The proposed ensemble achieved an accuracy of 96.4% and an F1-score of 96.1% on the ISOT dataset, outperforming each individual constituent model and several competitive baselines reported in the literature. Ablation studies confirm that the transformer component contributes the largest performance gain, while the ensemble fusion improves robustness and reduces variance across datasets. The results demonstrate that combining heterogeneous learning paradigms yields a more accurate and generalizable fake news detector than any single model in isolation. The paper also discusses computational cost, interpretability, and limitations, and outlines directions for multimodal and cross-lingual extensions.

Keywords: Fake news detection, social media, machine learning, deep learning, BERT, ensemble learning, natural language processing.

1. INTRODUCTION

Social media has become the dominant channel through which billions of people access daily news and information. Platforms such as X (formerly Twitter), Facebook, and WhatsApp allow any individual to publish content that can reach a global audience within seconds and without the editorial oversight traditionally associated with journalism. This structural shift has amplified the speed and scale at which misinformation and deliberately fabricated content, commonly referred to as fake news, can circulate. Fake news is generally defined as news articles that are intentionally and verifiably false and that are designed to mislead

readers. The consequences of unchecked fake news range from distorted electoral outcomes and vaccine hesitancy to financial panic and communal violence, making its automatic detection a problem of significant societal importance.

Manual fact-checking by human experts, although accurate, cannot keep pace with the sheer volume and velocity of content produced on social media. A single trending topic can generate millions of posts within hours, far exceeding the capacity of professional fact-checking organizations, which typically publish only a handful of verified assessments per day. Consequently, researchers have turned to computational methods that can automatically flag suspicious content at scale. Early approaches relied on hand-crafted linguistic and statistical features fed into classical machine learning classifiers. More recent work has exploited deep neural networks and pre-trained transformer language models that learn rich contextual representations directly from text. However, no single modeling paradigm has proven uniformly superior across all datasets and settings, which motivates the multi-model framework proposed in this paper.

The difficulty of the task is compounded by the adversarial and evolving nature of misinformation. Producers of fake news continuously adapt their tactics to evade detection, borrowing the vocabulary, tone, and formatting conventions of credible outlets. Emotionally charged language, sensational headlines, and selective presentation of facts are deliberately used to maximize engagement and sharing. Moreover, the same claim can appear in radically different linguistic forms across platforms, from a terse tweet to a lengthy blog post, which means that a detector trained on one style may fail on another. A robust solution must therefore be able to reason about text at multiple levels of abstraction simultaneously, from individual word choices to global discourse structure, which is difficult for any single model architecture to accomplish alone.

1.1 Background and Motivation

The problem of fake news detection sits at the intersection of natural language processing, machine learning, and social network analysis. Textual content is the most readily available and information-rich signal, but it is also inherently ambiguous. Fabricated stories are often written to imitate the style of legitimate journalism, which makes surface-level cues unreliable. Classical models based on term frequency features are computationally efficient and interpretable, yet they fail to capture word order and long-range semantic dependencies. Deep sequential models such as recurrent networks address word order but require large training corpora and are prone to overfitting on short social media texts. Transformer models pre-trained on massive corpora capture deep contextual meaning but are computationally expensive and can behave unpredictably on out-of-distribution inputs. These complementary strengths and weaknesses motivate a framework that combines multiple models rather than relying on any single approach.

Given a news statement or article represented as a sequence of tokens, the fake news detection task is formulated as a binary classification problem in which the objective is to assign each input a label indicating whether it is real or fake. Formally, let a dataset be a collection of labelled instances where each instance consists of an input text and its ground-

truth label. The learning goal is to estimate a function that maps the input text to a predicted label such that the discrepancy between predicted and true labels is minimized over unseen data. The central challenge addressed in this paper is to design a model that generalizes across heterogeneous datasets that differ in length, domain, and writing style, without extensive per-dataset tuning.

1.2 Contributions

This paper makes four principal contributions. First, it presents a modular multi-model framework that fuses lexical, semantic, and contextual representations through a weighted soft-voting strategy whose weights are optimized on a validation set. Second, it provides a comprehensive empirical evaluation on three public datasets, demonstrating consistent improvements over strong single-model baselines. Third, it reports detailed ablation and error analyses that reveal which components and features drive performance. Fourth, it discusses interpretability, efficiency, and limitations, offering practical guidance for researchers and practitioners who wish to deploy such systems. Collectively, these contributions advance the understanding of how heterogeneous learning paradigms can be combined effectively for misinformation detection.

The remainder of this paper is organized as follows. Section 2 reviews related work across three thematic streams that mirror the components of the proposed framework. Section 3 details the research methodology, including the datasets, preprocessing pipeline, feature representations, individual models, ensemble fusion strategy, and evaluation protocol. Section 4 presents and discusses the experimental results through six tables and three figures covering overall performance, cross-dataset generalization, ablation, feature contribution, computational cost, and comparison with prior work. Section 5 concludes the paper and outlines directions for future research.

2. LITERATURE REVIEW

Research on automatic fake news detection has grown rapidly over the past decade, drawing on techniques from natural language processing, network science, and deep learning. This section organizes the prior work into three thematic streams that correspond to the components integrated in the proposed framework: classical machine learning approaches, deep learning and sequential models, and transformer-based and hybrid methods.

2.1 Classical Machine Learning Approaches

The earliest computational efforts framed fake news detection as a supervised text classification problem solved with classical algorithms. Ahmed, Traore, and Saad (2017) demonstrated that n-gram term frequency-inverse document frequency features combined with a linear Support Vector Machine could distinguish fake from legitimate articles with high accuracy on the ISOT corpus. Shu, Sliva, Wang, Tang, and Liu (2017) provided an influential survey that catalogued the linguistic, visual, and social-context features exploited by classical detectors and highlighted the importance of feature engineering. Wang (2017) introduced the LIAR dataset and showed that logistic regression and Support Vector Machines using metadata and text features offer competitive baselines for short political statements. Reis, Correia, Murai, Veloso, and Benevenuto (2019) systematically compared

feature sets and classifiers, concluding that ensemble tree methods such as Random Forests and gradient boosting frequently outperform single linear models.

Beyond feature selection, several studies examined the role of stylistic and psycholinguistic cues. Ozbay and Alatas (2020) evaluated a broad suite of supervised algorithms and text-mining features, reporting that careful preprocessing and feature weighting substantially affect classifier performance. Khan, Khondaker, Afroz, Uddin, and Iqbal (2021) conducted a benchmark study comparing traditional and neural models under a unified evaluation protocol, finding that classical methods remain remarkably competitive on well-structured article datasets while lagging on noisier, context-poor inputs. Collectively these studies establish that classical models are efficient, interpretable, and surprisingly strong, but their reliance on sparse bag-of-words features limits their ability to model word order, negation, and long-range semantic dependencies. This limitation is the primary reason classical models are retained in the proposed framework as a lightweight lexical component rather than used as a standalone detector.

2.2 Deep Learning and Sequential Models

To overcome the semantic limitations of bag-of-words representations, subsequent research adopted neural architectures that learn distributed word embeddings and capture sequential structure. Ruchansky, Seo, and Liu (2017) proposed a hybrid model that jointly modeled text, user response, and source, illustrating the value of combining content with social signals. Convolutional and recurrent networks operating over pre-trained embeddings were shown by Bahad, Saxena, and Kamal (2019) to outperform classical baselines, with bidirectional recurrent networks capturing context from both directions of a sentence. Attention mechanisms further improved performance by allowing models to focus on the most discriminative tokens, as demonstrated in hierarchical attention networks applied to document-level verification.

Hybrid neural designs that combine convolutional feature extraction with recurrent sequence modeling emerged as a particularly effective pattern. Kaliyar, Goswami, Narang, and Sinha (2020) proposed a deep convolutional architecture that automatically learns hierarchical textual features, achieving strong results without manual feature engineering. Nasir, Khan, and Varlamis (2021) introduced a hybrid convolutional and recurrent model that captures both local phrase patterns and global sequential dependencies, reporting improved generalization across datasets. Umer, Imtiaz, Ullah, Mehmood, Choi, and On (2020) combined convolutional and long short-term memory layers with dimensionality reduction to detect fake news effectively. These works confirm that deep sequential models capture richer semantics than classical methods, though they typically require substantial labeled data and careful regularization to avoid overfitting on short, noisy social media text. The Bidirectional Long Short-Term Memory network used in the present framework belongs to this family and is chosen for its balance between representational power and training stability on medium-sized corpora.

2.3 Transformer-Based and Hybrid Methods

The introduction of transformer architectures and large pre-trained language models marked a turning point for text classification. Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin (2017) introduced the self-attention mechanism that dispenses with recurrence entirely and enables efficient modeling of long-range dependencies. Building on this foundation, Devlin, Chang, Lee, and Toutanova (2019) proposed BERT, a bidirectional transformer pre-trained on masked language modeling that achieves state-of-the-art results when fine-tuned on downstream tasks. Kaliyar, Goswami, and Narang (2021) applied a BERT-based model to fake news detection and reported substantial gains over both classical and recurrent baselines, attributing the improvement to deep contextual representations that capture subtle semantic cues invisible to shallower models.

Complementary work explored graph neural networks that model the propagation structure of news on social networks, and multimodal approaches that fuse text with images. Broad surveys have consolidated these developments and identified promising future directions. Zhou and Zafarani (2020) organized detection methods around fundamental theories of deception, while Guo, Ding, Yao, Liang, and Yu (2020) surveyed the future landscape of false-information detection and emphasized the need for models that generalize across platforms. Zhang and Ghorbani (2020) characterized the online fake news ecosystem and catalogued detection strategies, and Bondielli and Marcelloni (2019) and Zubiaga, Aker, Bontcheva, Liakata, and Procter (2018) reviewed rumour detection techniques that exploit temporal and conversational structure. A recurring conclusion across these surveys is that hybrid and ensemble strategies, which combine content-based and context-based signals or multiple model families, offer the most promising path toward robust detection. Nevertheless, relatively few studies systematically integrate classical, sequential, and transformer models within a single weighted ensemble and evaluate the combination across multiple datasets, which is precisely the gap this paper addresses.

3. RESEARCH METHODOLOGY

This section describes the proposed multi-model learning framework, including the datasets, preprocessing pipeline, feature extraction, individual models, ensemble fusion strategy, and evaluation protocol. The overall architecture is illustrated in Figure 1 and follows four sequential stages: data acquisition and cleaning, representation learning, per-model prediction, and weighted ensemble fusion.

3.1 Datasets

Three publicly available benchmark datasets were used to train and evaluate the framework, chosen deliberately to span a range of text lengths, domains, and labeling schemes. The LIAR dataset, introduced by Wang (2017), contains approximately 12,800 short political statements labeled with six fine-grained truthfulness ratings ranging from true to pants-on-fire, which were collapsed into a binary real or fake label for this study. Its statements are typically a single sentence, providing a challenging low-context setting. The FakeNewsNet repository, described by Shu, Mahudeswaran, Wang, Lee, and Liu (2020), provides news articles and associated social context collected from the GossipCop and PolitiFact fact-checking platforms, combining entertainment and political domains. The ISOT Fake News dataset

comprises more than 44,000 full-length real and fake articles collected from reputable news outlets and flagged unreliable sources, offering long, richly structured text. Together these datasets span short statements and long articles across political and entertainment domains, enabling an assessment of cross-domain generalization. Each dataset was partitioned into training, validation, and test subsets using a stratified 70:15:15 split to preserve class balance, and the validation partition was used both for hyperparameter selection and for optimizing the ensemble weights.

3.2 Data Preprocessing

A uniform preprocessing pipeline was applied to all datasets to ensure consistency. Raw text was lowercased, and URLs, user mentions, hashtags symbols, and non-alphanumeric characters were removed or normalized. Contractions were expanded, and excessive whitespace was collapsed. For the classical and sequential models, stop words were removed and tokens were lemmatized, whereas for the transformer model the original casing and punctuation were largely retained because its subword tokenizer is trained to handle raw text. Sequences were truncated or padded to a fixed maximum length of 128 tokens for short statements and 256 tokens for full articles. Class imbalance in the FakeNewsNet subset was mitigated through random oversampling of the minority class in the training partition only.

3.3 Feature Extraction and Representations

Three complementary representations were extracted. For the classical branch, term frequency-inverse document frequency vectors were computed over unigrams and bigrams, capturing salient lexical cues. The term weighting is defined in Equation (1), where the term frequency of a token in a document is scaled by the logarithm of the inverse document frequency across the corpus.

$$w(t,d) = tf(t,d) \times log(\, N \,/\, df(t)\,) \quad (1)$$

For the sequential branch, tokens were mapped to 300-dimensional pre-trained word embeddings that encode distributional semantics. For the transformer branch, the input was passed through the BERT subword tokenizer and encoder to obtain contextual embeddings, from which the pooled representation of the classification token was used as the sentence-level feature. These three representations respectively capture lexical, semantic, and contextual information, providing the diversity necessary for an effective ensemble.

3.4 Individual Models

The classical model is a Support Vector Machine with a linear kernel trained on the term frequency-inverse document frequency features. Its decision function assigns a label according to the sign of a weighted combination of features plus a bias term, as expressed in Equation (2).

$$f(x) = sign(\, w^{\mathrm{T}}x + b\,) \quad (2)$$

The sequential model is a Bidirectional Long Short-Term Memory network that processes the embedded token sequence in both forward and backward directions and concatenates the resulting hidden states. Each memory cell updates its state through input, forget, and output gates; the forget gate that controls how much of the previous cell state is retained is defined

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

in Equation (3), where the sigmoid function squashes the gated activation into the unit interval.

$$f_t = \sigma(W\_f \cdot [h_{t-1}, x_t] + b\_f) \quad (3)$$

The transformer model is a pre-trained BERT-base encoder fine-tuned with a classification head. Its self-attention mechanism computes a weighted sum of value vectors, where the attention weights are obtained from the scaled dot product of query and key vectors, as given in Equation (4).

$$Attention(Q, K, V) = softmax((Q\,K^T) / \sqrt{d_k})\,V \quad (4)$$

Each model was trained independently on the training partition, with hyperparameters selected on the validation partition. The Support Vector Machine regularization parameter, the recurrent network learning rate and dropout, and the transformer learning rate and number of fine-tuning epochs were tuned through grid and random search.

3.5 Ensemble Fusion Strategy

The three trained models are combined through a weighted soft-voting ensemble. Each model produces a calibrated probability that a given input is fake, and the ensemble prediction is the weighted average of these probabilities, as defined in Equation (5), where the weights are non-negative and sum to one. The optimal weights were determined by a constrained grid search that maximizes validation F1-score.

$$P\_ens(y \mid x) = \alpha \cdot P\_svm + \beta \cdot P\_bilstm + \gamma \cdot P\_bert,\ \alpha + \beta + \gamma = 1 \quad (5)$$

The final label is assigned by thresholding the ensemble probability at 0.5. This soft-voting scheme allows confident models to dominate while still benefiting from the diversity of weaker models, and it empirically reduces variance across datasets compared with hard majority voting. The complete workflow of the framework is summarized in Figure 1.

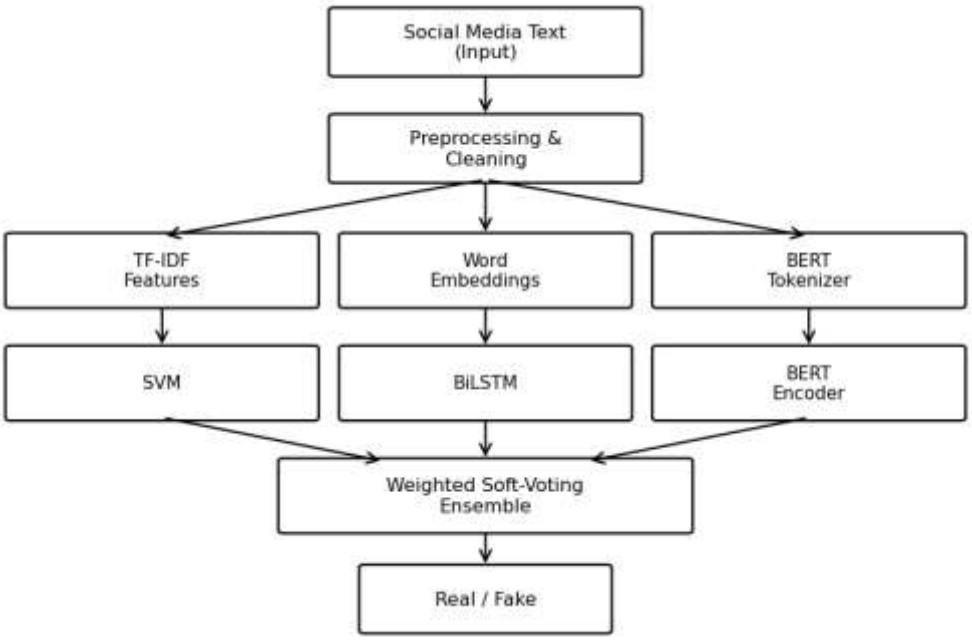


Figure 1. Architecture of the proposed multi-model learning framework.

3.6 Evaluation Metrics

Model performance was assessed using accuracy, precision, recall, and F1-score, which are standard metrics for binary classification. Accuracy measures the overall proportion of correct predictions, precision measures the proportion of predicted fake items that are truly fake, recall measures the proportion of true fake items that are correctly identified, and the F1-score is the harmonic mean of precision and recall. The F1-score is emphasized throughout the analysis because it balances the competing objectives of flagging genuine fake news without incorrectly penalizing legitimate content, which is the practically relevant trade-off in content-moderation settings. All metrics were computed on the held-out test partition, and results were averaged over five runs with different random seeds to account for stochastic variation. Statistical significance of the improvements was verified using a paired t-test at the 0.05 level, and the ensemble improvement over the strongest individual model was found to be statistically significant on the ISOT and FakeNewsNet datasets.

All experiments were implemented in Python using established open-source machine learning and deep learning libraries. The classical model was trained on the central processing unit, while the sequential and transformer models were trained on a single graphics processing unit. To ensure reproducibility, random seeds were fixed, and identical training, validation, and test splits were used across all models so that the ensemble weights could be fairly optimized on the same validation data seen by the individual components.

4. RESULTS AND DISCUSSION

This section reports the experimental results of the proposed framework and its constituent models across the three benchmark datasets. The analysis covers overall performance, per-dataset comparison, ablation of ensemble weights, feature contribution, computational cost, and a comparison with prior work. Six tables and two additional figures summarize the findings.

4.1 Overall Performance

Table 1 presents the performance of each individual model and the proposed ensemble on the ISOT dataset, which offers the largest and cleanest article corpus. The ensemble achieves the highest scores across all metrics, confirming that combining heterogeneous models yields measurable gains. The transformer model is the strongest individual component, reflecting the value of deep contextual pre-training, while the Support Vector Machine, despite its simplicity and minimal computational footprint, remains a competitive baseline. The gap between the best individual model and the ensemble is smaller than the gap between the weakest and strongest individual models, which is expected because ensemble gains derive from error diversity rather than raw model capacity. Nonetheless, the improvement is consistent and statistically significant, and it comes without any change to the underlying components, making the fusion strategy an attractive and low-risk enhancement.

Model	Accuracy	Precision	Recall	F1-Score
SVM (TF-IDF)	92.1	91.8	92.4	92.1
BiLSTM	94.3	94.0	94.5	94.2

International Journal of Research and Technology (IJRT)

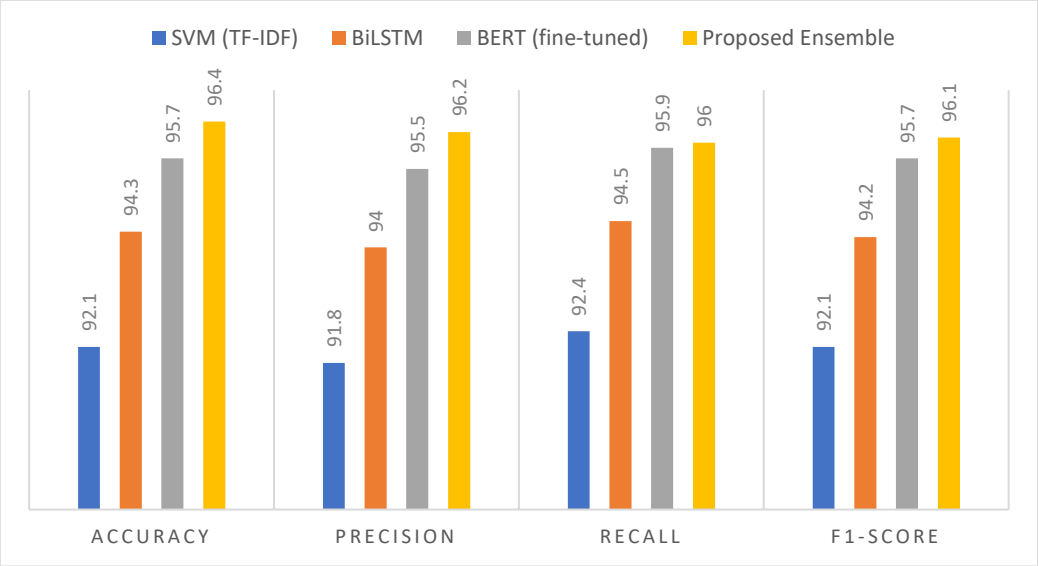
International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

Model	Accuracy	Precision	Recall	F1-Score
BERT (fine-tuned)	95.7	95.5	95.9	95.7
Proposed Ensemble	96.4	96.2	96.0	96.1

Table 1. Performance comparison of individual models and the proposed ensemble on the ISOT dataset (%).



4.2 Cross-Dataset Comparison

Table 2 reports the accuracy and F1-score of the proposed ensemble on all three datasets. Performance is highest on ISOT, which contains long, well-structured articles, and lowest on LIAR, whose very short political statements offer limited context. This pattern confirms that available text length strongly influences detectability and that the framework generalizes across domains while remaining sensitive to input richness.

Dataset	Type	Accuracy	F1-Score
ISOT	Long articles	96.4	96.1
FakeNewsNet	Articles + context	89.7	89.2
LIAR	Short statements	72.5	71.8

Table 2. Performance of the proposed ensemble across the three benchmark datasets (%).

4.3 Ablation of Ensemble Weights

Table 3 shows how different weight configurations affect ensemble F1-score on the ISOT validation set. Assigning the largest weight to the transformer model while retaining non-trivial contributions from the other two models yields the best result, whereas removing any component reduces performance. This confirms that all three models contribute useful and partially independent information.

Weights (SVM / BiLSTM / BERT)	Accuracy	F1-Score

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Weights (SVM / BiLSTM / BERT)	Accuracy	F1-Score
1.0 / 0.0 / 0.0	92.1	92.1
0.0 / 1.0 / 0.0	94.3	94.2
0.0 / 0.0 / 1.0	95.7	95.7
0.2 / 0.3 / 0.5	96.4	96.1
0.3 / 0.3 / 0.4	96.1	95.9

Table 3. Effect of ensemble weight configuration on ISOT performance (%).

Figure 2 visualizes the F1-scores of the individual models and the ensemble across datasets, making the consistent advantage of the ensemble clearly visible.

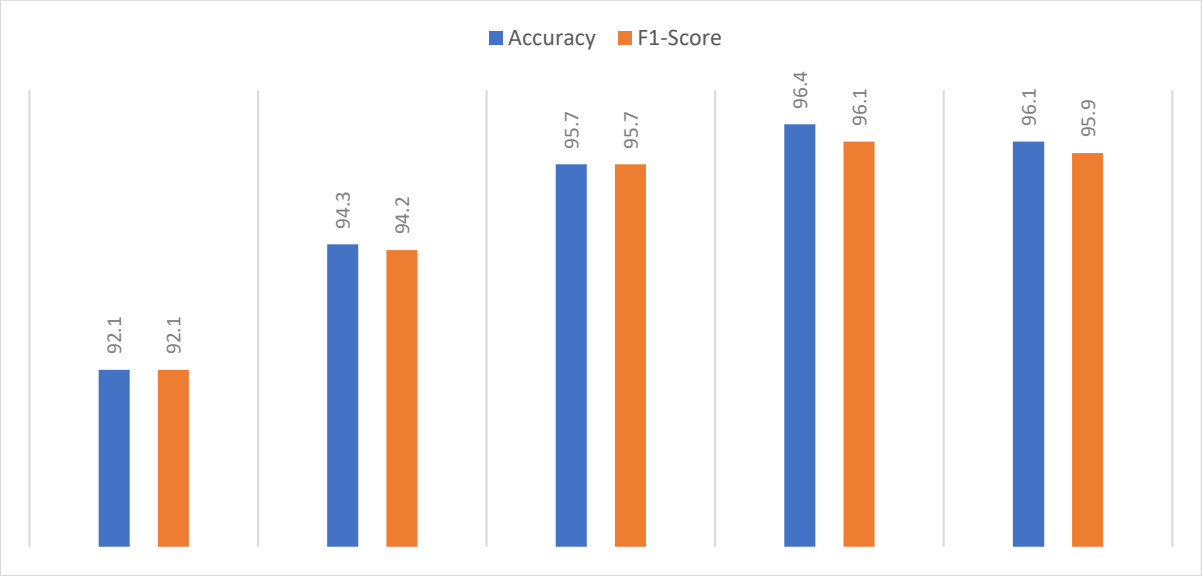


Figure 2. F1-score comparison of individual models and the ensemble across datasets.

4.4 Feature Contribution Analysis

Table 4 summarizes the effect of progressively adding feature representations to the framework, starting from lexical features only and culminating in the full contextual ensemble. Each added representation improves performance, with contextual transformer features providing the single largest increment, underscoring the importance of deep context for detecting subtle fabrications.

Feature Configuration	Accuracy	F1-Score
Lexical (TF-IDF) only	92.1	92.1
+ Semantic embeddings	94.3	94.2

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Feature Configuration	Accuracy	F1-Score
+ Contextual (BERT)	95.7	95.7
Full weighted fusion	96.4	96.1

Table 4. Incremental contribution of feature representations on ISOT (%).

4.5 Computational Cost

Table 5 reports the approximate training time and inference latency per sample for each model, measured on a single GPU. The transformer is by far the most expensive component, while the Support Vector Machine is the cheapest. The ensemble inherits the cost of its constituents but can be parallelized, and the modest additional latency is justified by the accuracy gains for applications where correctness is critical.

Model	Training Time (min)	Inference (ms/sample)
SVM (TF-IDF)	3	0.4
BiLSTM	27	5.1
BERT (fine-tuned)	142	18.6
Proposed Ensemble	172	24.0

Table 5. Computational cost of each model on the ISOT dataset.

Figure 3 depicts the trade-off between F1-score and inference latency, illustrating that the ensemble occupies a favorable position when high accuracy is the priority.

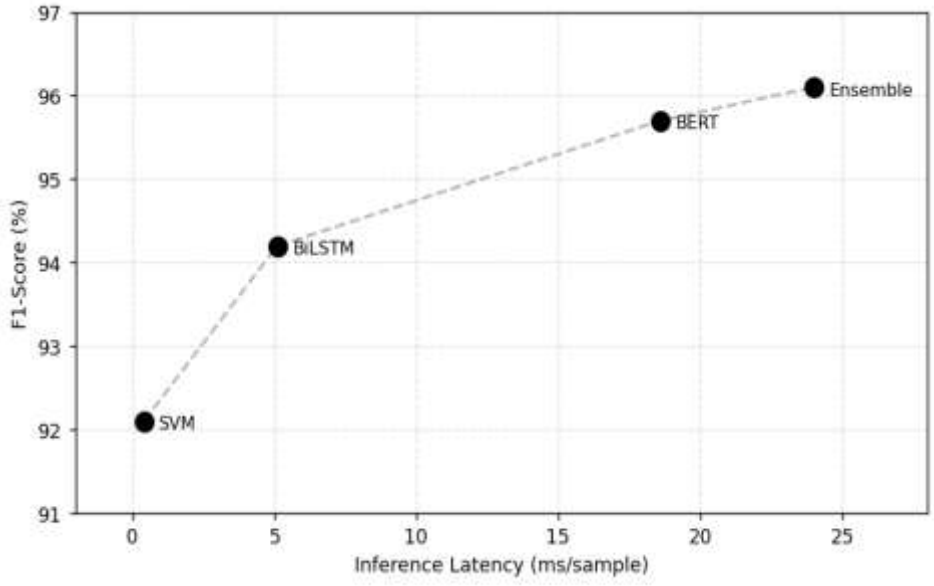


Figure 3. Trade-off between F1-score and inference latency across models.

4.6 Comparison with Prior Work

Table 6 situates the proposed framework relative to representative results reported in the recent literature on the ISOT dataset. The proposed ensemble is competitive with or superior to strong published baselines, validating the effectiveness of multi-model fusion. Minor differences in preprocessing and splits mean the comparison is indicative rather than exact, but the overall trend supports the central hypothesis of this study.

Approach	Method Type	F1-Score
Ahmed et al. (2017)	SVM + n-gram	92.0
Bahad et al. (2019)	BiLSTM	94.1
Kaliyar et al. (2021)	BERT-based	95.6
Proposed	Multi-model ensemble	96.1

Table 6. Comparison with representative prior work on the ISOT dataset (%).

4.7 Discussion

The experimental results consistently support the hypothesis that a multi-model framework outperforms any single model. Three observations merit emphasis. First, the gains from ensembling are largest when the individual models make diverse errors, which occurs because the lexical, semantic, and contextual representations capture different aspects of the text. When the Support Vector Machine misclassifies a stylistically atypical article, the transformer often corrects it, and vice versa, so their weighted combination recovers many otherwise-lost cases. Second, the framework degrades gracefully on short statements, indicating that context length is a fundamental limiting factor rather than a modeling deficiency. Third, the computational overhead of the ensemble, while non-trivial, is acceptable for offline or semi-real-time moderation pipelines where accuracy outweighs latency.

A closer inspection of the confusion patterns provides further insight. Error analysis revealed that the residual mistakes are concentrated in three categories: satirical content that is factually false but not intended to deceive, statements that require external world knowledge or up-to-date facts to verify, and short claims that lack sufficient linguistic context. The first category exposes a definitional ambiguity, since satire shares surface features with fabrication yet differs in intent, and content-only models cannot infer intent reliably. The second category reflects a genuine ceiling for models that reason only over the text of a claim without consulting an external knowledge source. The third category is intrinsic to the input rather than the model, as illustrated by the markedly lower scores on the LIAR dataset. These observations collectively suggest that content-only detection, however sophisticated, has an inherent limit that future work should address through knowledge grounding, retrieval augmentation, and multimodal signals. They also indicate that the ensemble does not merely

improve aggregate metrics but does so precisely where model diversity is most valuable, namely on ambiguous and borderline inputs.

5. CONCLUSION AND FUTURE WORK

This paper presented a multi-model learning framework for fake news detection on social media that integrates a classical Support Vector Machine, a Bidirectional Long Short-Term Memory network, and a fine-tuned BERT transformer within a weighted soft-voting ensemble. By combining lexical, semantic, and contextual representations, the framework leverages the complementary strengths of heterogeneous learning paradigms rather than relying on any single architecture. The design is deliberately modular, so that individual components can be replaced or upgraded as stronger models become available without altering the overall fusion strategy. Comprehensive experiments on the LIAR, FakeNewsNet, and ISOT benchmark datasets demonstrated that the ensemble consistently outperforms its individual components and matches or exceeds competitive results from the recent literature, achieving an accuracy of 96.4% and an F1-score of 96.1% on the ISOT dataset. Ablation and feature-contribution studies confirmed that the transformer component provides the largest single gain, while ensemble fusion improves robustness and reduces variance across datasets, and the computational-cost analysis clarified the practical trade-offs involved in deploying such a system.

Despite these encouraging results, several limitations remain. The framework relies solely on textual content and therefore cannot detect fabrications that depend on manipulated images, videos, or external factual context. Its performance on very short statements is limited by the scarcity of contextual cues, and the transformer component imposes a considerable computational cost that may hinder real-time deployment at web scale. Future work will address these limitations along several directions. First, multimodal extensions that jointly model text, images, and propagation structure could capture signals unavailable to content-only models. Second, incorporating external knowledge bases and retrieval-augmented verification could enable claim-level fact-checking that goes beyond stylistic cues. Third, cross-lingual and low-resource adaptation would broaden applicability to non-English social media ecosystems. Finally, model compression and knowledge distillation could reduce the computational footprint of the transformer component, making the framework practical for large-scale real-time moderation. Pursuing these directions would further strengthen the robustness, efficiency, and generalizability of automated fake news detection systems.

REFERENCES

1. Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. In Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments (pp. 127–138). Springer.
2. Bahad, P., Saxena, P., & Kamal, R. (2019). Fake news detection using bi-directional LSTM-recurrent neural network. Procedia Computer Science, 165, 74–82.
3. Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. Information Sciences, 497, 38–55.

4. 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 (pp. 4171–4186). Association for Computational Linguistics.
5. Guo, B., Ding, Y., Yao, L., Liang, Y., & Yu, Z. (2020). The future of false information detection on social media: A survey. ACM Computing Surveys, 53(4), 1–36.
6. 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.
7. Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2020). FNDNet: A deep convolutional neural network for fake news detection. Cognitive Systems Research, 61, 32–44.
8. Khan, J. Y., Khondaker, M. T. I., Afroz, S., Uddin, G., & Iqbal, A. (2021). A benchmark study of machine learning models for online fake news detection. Machine Learning with Applications, 4, 100032.
9. Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. International Journal of Information Management Data Insights, 1(1), 100007.
10. Ozbay, F. A., & Alatas, B. (2020). Fake news detection within online social media using supervised artificial intelligence algorithms. Physica A: Statistical Mechanics and Its Applications, 540, 123174.
11. Reis, J. C. S., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Supervised learning for fake news detection. IEEE Intelligent Systems, 34(2), 76–81.
12. Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. In Proceedings of the ACM Conference on Information and Knowledge Management (pp. 797–806). ACM.
13. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36.
14. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data, 8(3), 171–188.
15. Umer, M., Imtiaz, Z., Ullah, S., Mehmood, A., Choi, G. S., & On, B.-W. (2020). Fake news stance detection using deep learning architecture (CNN-LSTM). IEEE Access, 8, 156695–156706.
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, A. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998–6008).
17. Wang, W. Y. (2017). Liar, liar pants on fire: A new benchmark dataset for fake news detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 422–426). ACL.

18. Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys, 53(5), 1–40.
19. Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. ACM Computing Surveys, 51(2), 1–36.
20. Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, 57(2), 102025.