# SBERT-Based Stacking Ensemble Model for Fake News Detection

Abdulaziz A Alzubaidi, Amin A Alawady
Department of Computer Skills-Deanship of Preparatory Year, Najran University, Najran, Saudi Arabia

***Abstract*—Fake news has become a significant global challenge, affecting public opinion, social dynamics, and decision-making processes. Detecting fabricated news accurately and efficiently remains a challenging task due to the diversity of content, writing styles, and subtle semantic nuances. In this study, we propose a stacking ensemble model that uses SBERT-based semantic embeddings to improve the detection of fake news. The model integrates several machine-learning classifiers with a meta-learner to enhance robustness and predictive reliability. Experiments on the WELFake dataset show that the proposed model achieves 92.74% accuracy, a 93.01% F1-score, and a 97.93% ROC-AUC in classifying fake and real news. These results demonstrate the model's effectiveness and suggest its potential for broader application across different languages and news domains.**

***Keywords*—Fake news detection; machine learning; SBERT embeddings; stacking ensemble; Random Forest; Logistic Regression; MLP; XGBoost**

## I. Introduction

The proliferation of online platforms has dramatically increased the circulation of information, but it has also facilitated the rapid spread of fake news and misinformation [1]. Fake news poses significant risks to society by influencing public opinion, manipulating governmental processes, and undermining trust in credible information sources [1], [2]. Recent studies emphasize that the challenge of detecting fake news is not only a technical issue but also a social and ethical one, as misinformation can impact domains such as healthcare and public safety [1], [2],[3].

Recent global surveys highlight an alarming concern among populations regarding fake news. For instance, a Pew Research Center survey in April 2025 found that over 80% of adults in 35 countries see fake news as a major issue in their nation, with 59% labeling it "very big". Moreover, 72% of adults across 25 countries perceive the spread of false information online as a major national threat. Nationally, in the Philippines, 59% believe fake news on social media. This is a serious problem, while 62% say the same about traditional media. In Brazil, almost 90% admit to having believed a fake news item at some point, despite many claiming confidence in their ability to discern truth from falsehood [4].

The detection of deceptive information has consequently become a central research focus in artificial intelligence and natural language processing. Traditional machine learning models, including Support Vector Machines (SVM), Random Forests (RF), and gradient boosting techniques, have been extensively applied to text classification, providing an initial foundation for distinguishing authentic from deceptive content [5]. However, studies have demonstrated that these models often struggle to capture the deeper semantic and contextual dependencies inherent in human language, which are critical for reliable fake news detection [6]. Transformer-based architecture, contextual embeddings, and multimodal approaches have demonstrated superior performance in modeling these dependencies and improving classification accuracy [7].

Advancements in deep learning have enabled the extraction of high-level, context-aware features, significantly improving classification performance. Architectures such as Convolutional Neural Networks, Recurrent Neural Networks, and Long Short-Term Memory networks have been employed to model sequential dependencies and linguistic patterns in text [1]. More recently, transformer-based models, including BERT, RoBERTa, and XLNet, have demonstrated superior capability in understanding nuanced contextual information through self-attention mechanisms, thereby enhancing the accuracy and reliability of fake news identification [7].

In addition to the context of semantics-based approaches, current research emphasizes challenges such as dataset imbalance, scarcity of annotated resources for low-resource languages, and the interpretability of models. These challenges are particularly relevant for fake news detection, as textual data alone often contains nuanced semantic and contextual dependencies that are difficult for traditional models to capture [2], [8]. Moreover, the rise of multilingual environments introduces further complexity, as many languages have limited annotated datasets and resources. Handling multilingual textual data effectively requires models capable of generalizing across languages while maintaining high detection accuracy, which remains a significant challenge in current research [9].

Despite the extensive research on fake news detection, existing approaches often face limitations in capturing deep semantic relationships, handling multilingual data, and maintaining robustness across heterogeneous datasets. This study addresses these gaps by proposing an innovative ensemble model that integrates semantic embeddings with multiple classifiers, enhancing both accuracy and generalizability, with the added advantage of applicability across multiple natural languages in real-world scenarios.

The contributions of this study are summarized as follows:

- An innovative stacking ensemble model is proposed for fake news detection, combining semantic embeddings using SBERT with multiple machine learning

classifiers as base learners, including Random Forest, Logistic Regression, and MLP, while XGBoost serves as the meta-learner to enhance predictive accuracy and robustness.

- An OOF-based meta-learning strategy using XGBoost, enabling robust generalization and reducing overfitting across large and heterogeneous datasets.
- A practical and extensible framework that can be adapted to other NLP tasks requiring deeper semantic understanding and enhanced model stability.

This study is structured as follows: Section II provides an overview of related work, Section III explains the materials and methods, Section IV presents the experimental setup, Section V discusses the evaluation metrics, Section VI and Section VII report and analyze the results, and Section VIII concludes the study.

## II. Related Work

Ensemble Learning is a computational paradigm that combines multiple machine learning models to achieve higher predictive accuracy and robustness compared to individual learners. It reduces bias, variance, and overfitting by aggregating diverse classifiers through approaches such as bagging, boosting, and stacking [10]. Recent studies confirm its importance in fake news detection and natural language processing (NLP) tasks. Mouratidis et al. [6] emphasized that hybrid and ensemble frameworks integrating classical ML with deep models yield more reliable results, especially when combined with explainable AI methods. Moreover, Shah and Patel [2] provided a comprehensive survey highlighting ensemble and hybrid approaches as central to improving detection performance across textual and social-context features. Al-alshaqi et al. [11] proposed an ensemble-based framework for fake news detection, showing that integrating multiple models within a unified architecture enhances robustness and overall detection performance.

Alshuwaier and Alsulaiman [1] conducted a comprehensive review on fake news detection using machine learning and deep learning algorithms. The study examined over 30 papers published between 2018 and 2025, analyzing datasets, feature extraction methods, algorithms, and performance outcomes. Traditional ML models such as Naïve Bayes, Random Forest, Logistic Regression, and SVM achieved strong performance in several cases. Deep learning approaches, including CNN, Bi-LSTM, GNN, and BERT-based models, generally outperformed classical methods, particularly with large-scale and multilingual datasets. Hybrid and ensemble strategies showed the most promising results, combining the strengths of both ML and DL for more reliable detection. The review also identified major challenges, such as data imbalance, limited generalization across platforms, and the difficulty of early-stage detection. The authors emphasized that future research should focus on multimodal hybrid models and the integration of large language models (LLMs) to enhance robustness and adaptability in fake news prediction.

Hu, Mao, and Zhang [2] provided a survey on fake news detection from a novel perspective. They analyzed the diffusion process of fake news and identified three key characteristics: intentional creation, heteromorphic transmission, and controversial reception. Based on these, detection methods were grouped into feature-based, propagation-based, and stance-based approaches. The study emphasized the importance of integrating these perspectives, moving toward multimodal models, and improving interpretability for future research.

Ni'mah et al. [8] proposed a framework for fake news detection in low-resource languages under extreme class imbalance. The study introduced Contrast-BERT, a contrastive learning model that generates embeddings, which are then classified using a stacking-based ensemble of MLPs. This design outperformed standard end-to-end BERT classifiers, especially in scenarios where fake claims vastly outnumber real ones. The framework also proved effective for topic clustering and evidence retrieval, showing robustness across Indonesian, Bangla, and Czech datasets. The authors emphasize contrastive representation learning and ensemble strategies as promising directions for handling misinformation in low-resource settings

Jyoti and Kumar [12] introduced Ndetect, an ensemble-based machine learning model for fake news detection designed to enhance classification accuracy. The framework integrates multiple classifiers through ensemble techniques such as bagging, boosting, and Random Forest, tested on benchmark fake news datasets. Experimental results demonstrated that the ensemble approach significantly outperformed individual classifiers, achieving higher precision and recall while maintaining robustness across datasets. The study emphasized the effectiveness of ensemble learning in addressing data variability and improving reliability in fake news prediction.

Alnabhan and Branco [13] conducted a systematic literature review on fake news detection using deep learning approaches. The review categorized models into CNN, RNN, LSTM, Bi-LSTM, GNN, and transformer-based architectures such as BERT and RoBERTa, comparing their performance across diverse datasets and languages. Results indicated that deep learning consistently outperforms traditional ML models, particularly on large, complex, and multilingual data, with transformer models achieving the best overall performance. The study also highlighted challenges, including data imbalance, lack of generalization across platforms, and difficulties in early detection. The authors concluded that future research should focus on multimodal DL models and large-scale pre-trained architectures to improve robustness and adaptability in fake news prediction.

Although numerous studies have explored ensemble and hybrid approaches for fake news detection, a clear gap remains in effectively combining semantic-rich embeddings with diverse ensemble strategies to handle heterogeneous and imbalanced datasets. Previous research has demonstrated the benefits of classical machine learning, deep learning, and transformer-based models individually. However, there is limited work that integrates SBERT embeddings with a stacking ensemble framework specifically designed to exploit both semantic understanding and the complementary strengths

of multiple classifiers. This gap highlights the need for targeted strategies that leverage the synergy between semantic representations and ensemble learning to enhance robustness, generalization, and overall detection performance. Building on this insight, the current study proposes an approach that addresses this gap by integrating SBERT embeddings with a stacking ensemble to improve the accuracy and reliability of fake news detection.

## III. Materials and Methods

This research introduces a stacking ensemble model for fake news detection, comprising two hierarchical levels of learning. The first level, known as the base learners, involves generating sentence-level embeddings using SBERT (Sentence-BERT), a transformer-based model designed to produce dense, semantically meaningful vector representations for text. SBERT captures contextual information, word dependencies, and subtle semantic nuances, enabling the model to understand the relationships between sentences and phrases effectively. The proposed framework is designed to address the lack of semantic generalization observed in previous ensemble and transformer-based models. By combining SBERT embeddings with multiple heterogeneous base learners and an OOF-based meta-learning layer, the model leverages both deep contextual understanding and classifier complementarity. These embeddings are then fed into multiple machine learning classifiers, specifically Random Forest, Logistic Regression, and Multi-Layer Perceptron (MLP), which perform initial predictions. Although XGBoost was also trained as a base learner within the OOF procedure, it is not depicted in Fig. 1 to maintain clarity and avoid visual confusion, as its primary role in the diagram is the meta-learner. The outputs of the base learners are subsequently used as input features for the second level, the meta-learner, which is also implemented using XGBoost. The meta-learner leverages out-of-fold (OOF) predictions generated via Stratified K-Fold cross-validation, ensuring balanced representation of classes in each fold. This process allows the model to learn from diverse perspectives of the base classifiers, enhancing generalization and reducing overfitting. By combining semantic embeddings from SBERT with ensemble learning techniques, the proposed stacking model effectively captures both low-level textual patterns and high-level contextual relationships, significantly improving predictive performance. Fig. 1 provides a schematic overview of the SBERT-based stacking model, illustrating the flow from sentence embeddings through base learners to the final meta-learner output.

### A. Stacking

Stacking is a powerful ensemble learning technique designed to improve predictive accuracy by integrating the outputs of multiple base learners through a higher-level meta-learner. In this model, learning occurs over two levels. At the first level, diverse base learners are trained independently in different splits of the dataset, producing predictions that capture complementary patterns from the data. These predictions are then fed as input to the second-level meta-learner, which combines them in an optimal manner to produce the final output [14]. The proposed stacking ensemble model is designed to improve fake news detection by combining multiple machine learning classifiers with semantic embeddings from SBERT. The stacking model consists of two levels:

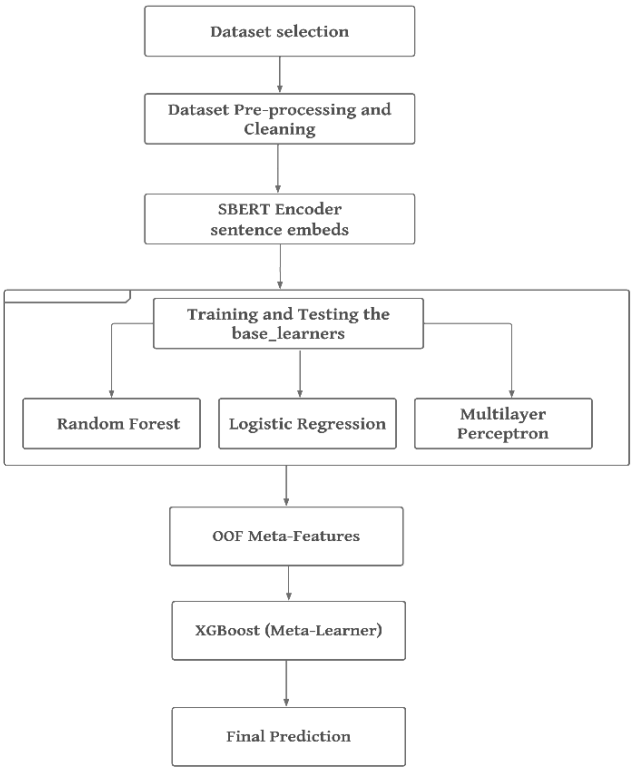


Fig. 1. Proposed stacking model.

*1) Base learners (level 1):* At the first level, several diverse classifiers, namely Random Forest (RF), Logistic Regression, and Multi-Layer Perceptron (MLP), are trained on sentence-level embeddings generated by SBERT. SBERT embeddings provide rich semantic, contextual representations of textual content and enabling the base learners to capture high-level relationships within the text.

To enhance robustness and avoid overfitting, Stratified K-Fold cross-validation is applied during training of the base learners. This procedure generates out-of-fold (OOF) predictions, which are predictions for the validation folds that are not seen by the respective base learner during training. These OOF predictions form a new set of features, referred to as meta-features, for the next level of the ensemble.

*2) Meta-learner (level 2):* The second level consists of a meta-learner, implemented using XGBoost, which is trained on the OOF predictions from the base learners. By learning from the combined outputs of the base learners, the meta-learner can exploit complementary strengths of each model, capture diverse perspectives, and improve overall predictive accuracy.

The stacking procedure thus allows the ensemble to integrate both high-level semantic information from SBERT embeddings and predictive patterns from multiple classifiers. This dual-level approach ensures that the model generalizes

well on unseen data and outperforms individual base learners. Fig. 1 illustrates the overall workflow of the proposed stacking ensemble model, showing the flow from SBERT embeddings to base learners, and finally to the meta-learner.

### B. SBERT

Sentence-BERT (SBERT) is a transformer-based model designed to produce semantically meaningful sentence embeddings. Unlike traditional BERT, which generates token-level embeddings, SBERT modifies the architecture to generate fixed-size sentence embeddings that can be directly used for similarity comparisons, clustering, or downstream classification tasks [15]. In our study, SBERT is employed to transform textual news content into dense vector representations that capture semantic relationships and contextual nuances between sentences. These embeddings serve as rich input features for the base learners in the stacking ensemble, allowing the model to leverage deep contextual information beyond surface-level word occurrences. We claim that by using SBERT embeddings, our stacking model can better discern subtle semantic differences between authentic and fake news, enhancing the overall detection performance.

### C. Random Forest

Random Forest (RF) is a widely adopted machine learning algorithm applicable to both classification and prediction tasks, and it plays a critical role in building robust models. The method relies on an ensemble of decision trees, where each tree is trained on different subsets of the data and features, introducing diversity into the learning process. Specifically:

- For every tree in the forest, a bootstrap sample is drawn randomly from the training dataset (sampling with replacement).
- At each node, a random subset of features is selected to determine the best split.

Once multiple trees are trained, their outputs are aggregated—either through majority voting for classification tasks or averaging for regression—to generate the final prediction. This ensemble mechanism effectively mitigates the risk of overfitting compared to using a single decision tree, thereby improving stability and overall generalization performance. Moreover, Random Forest provides insightful and quantitative measures of feature importance, allowing researchers to interpret model behavior and evaluate the contribution of each variable to the decision-making process more transparently [16], [17].

### D. Extreme Gradient-Boosting

XGBoost, introduced by Tianqi Chen and Carlos Guestrin, is an advanced implementation of the Gradient Boosting Decision Tree (GBDT) algorithm that has become one of the most widely adopted machine learning models due to its scalability, efficiency, and adaptability [18]. Unlike traditional GBDT, XGBoost incorporates regularization techniques to control model complexity and mitigate overfitting. It also integrates optimization strategies such as second-order gradient approximation, parallelized tree construction, and early stopping, which collectively accelerate training while improving predictive accuracy [19].

Moreover, a key strength of XGBoost is its ability to address class imbalance by assigning higher weights to misclassified instances, particularly from minority classes, during the boosting process. This mechanism improves classification performance in skewed data distributions, though it may sometimes introduce biased decision boundaries if not carefully tuned [20]. Furthermore, its inherent flexibility allows it to leverage multiple weak learners and iteratively refine them through boosting, resulting in a strong ensemble capable of capturing complex patterns in data [18]. In summary, XGBoost is recognized as a powerful and widely used tool in machine learning, excelling at handling complex problems and greatly improving the performance of predictive models.

### E. Multilayer Perceptron

Multilayer Perceptron (MLP) is a feedforward neural network capable of modeling complex non-linear patterns. Modern MLPs often use techniques such as dropout, batch normalization, and adaptive optimizers to improve training stability, accelerate convergence, and reduce overfitting [21]. In our stacking model, MLP serves as a base learner, producing probability predictions that are used as input features for the meta-learner, enhancing the ensemble's ability to capture non-linear relationships in the data.

### F. Logistic Regression

Logistic Regression (LR) is a widely used linear model for classification, estimating the probability of outcomes through the logistic function. LR remains popular due to its simplicity, interpretability, and efficiency, and modern studies often use it with regularization techniques to handle large or imbalanced datasets [22]. In our model, LR serves as a base learner, generating probability outputs that are used as input features for the meta-learner.

### G. Cross-Validation

Cross-validation (CV) is a widely used technique in machine learning to assess a model's ability to generalize to unseen data. In this approach, the dataset is partitioned into k equally sized folds, and the model is iteratively trained on k-1 folds while the remaining fold is reserved for validation. This process continues until each fold has served as a validation set exactly once [23]. In our stacking model, cross-validation is especially important for generating out-of-fold (OOF) predictions from base learners, which are then used as input features for a meta-learner. By doing so, the meta-learner is trained on predictions that are not biased by the same data used to fit the base models, enhancing the robustness of the ensemble.

### H. Study Data

The WELFake dataset is a well-prepared dataset in the field of fake news detection and machine learning. It comprises a total of 72,134 news articles, including 35,028 real and 37,106 fake news entries. To enhance the diversity of the dataset and reduce the risk of overfitting, the authors combined four popular news datasets (Kaggle, McIntire, Reuters, and BuzzFeed Political), providing a richer text corpus for more effective machine learning training [24].

The dataset contains four key columns: Serial Number (starting from 0), Title (news headline), Text (full news content), and Label (0 for fake and 1 for real). Missing or incomplete records were removed during preprocessing to ensure data quality and consistency [24]. This step guarantees that models trained on this dataset can learn effectively from clean and representative examples.

This dataset is publicly available on Kaggle, one of the largest global repositories for datasets, at: https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification

Published in IEEE Transactions on Computational Social Systems: (doi: 10.1109/TCSS.2021.3068519), and its characteristics are listed in Table I [24].

TABLE I. WELFAKE DATASET METADATA

| Dataset columns | | | |
|---|---|---|---|
| # | Column | Count | Datatype |
| 0 | Serial Number | 72134 | int64 |
| 1 | Title | 71576 | object |
| 2 | Text | 72095 | object |
| 3 | Label | 72134 | int64 |

## IV. EXPERIMENTAL SETUP

In this study, the proposed stacking model was developed using Python 3.10 within a Jupyter Notebook environment, executed on a system powered by an Intel Core i5-1035G7 processor (1.20 GHz) and 8 GB of RAM. The dataset employed was the WELFake dataset [22], which contains a total of 72,134 news articles, consisting of 35,028 real and 37,106 fake news entries. The dataset provides four columns: Serial Number, Title, Text, and Label. For our study, we focused exclusively on the Text and Label columns, as they directly support the fake news classification task, and we dropped any entries with missing text, as implemented in the preprocessing step: df = df[['text', 'label']].dropna(). This preprocessing slightly reduced the dataset but maintained a relatively balanced distribution between real and fake news, ensuring consistent and reliable input for our stacking ensemble model. The textual data were encoded into dense embeddings using Sentence-BERT (SBERT) all-MiniLM-L6-v2, which produced fixed-length semantic vectors for each article. These embeddings were then used to train the base learners: Logistic Regression, Random Forest, Multilayer Perceptron (MLP), and XGBoost. Each base learner was trained with a 3-fold Stratified K-Fold Cross-Validation to generate out-of-fold (OOF) predictions, which served as input features to the meta-learner. The meta-learner was implemented using XGBoost, which aggregated the OOF predictions from the base learners to improve generalization and reduce overfitting. Hyperparameters for all models were selected and fine-tuned empirically to achieve stable performance during training and validation.

Finally, the performance of the stacking ensemble was evaluated on a 20% holdout test set that was not involved in any training or validation step. Metrics such as Accuracy, F1-score, and the classification report were reported to assess the robustness and overall effectiveness of the model.

## V. PERFORMANCE MEASURE

We assessed the performance of our proposed model using standard evaluation metrics, starting with:

### A. Accuracy

It represents the proportion of correctly classified news articles (both real and fake) out of the total number of articles. It provides a general measure of how effectively the model distinguishes between real and fake news [25]. The definition of accuracy is captured by Eq. (1):

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Assuming the outcome of a news article indicates whether it is Real or Fake:

- True positive (TP): articles correctly identified as real.
- True negative (TN): articles correctly identified as fake.
- False positive (FP): fake articles incorrectly predicted as real.
- False negative (FN): real articles incorrectly predicted as fake.

### B. Precision

It measures how well the model correctly identifies articles as real when it predicts them to be real. A high precision score indicates that most articles predicted as real are indeed real [25]. Eq. (2) shows the formulation of precision:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Assuming the outcome of a news article indicates whether it is Real or Fake:

- True positive (TP): articles correctly classified as real.
- False positive (FP): fake articles incorrectly predicted as real.

### C. Recall

Recall measures the proportion of actual positive instances correctly identified by the model. In fake news detection, it reflects how many real articles are correctly classified as real [25]. Recall is defined in Eq. (3):

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Assuming the outcome of a news article indicates whether it is Real or Fake:

- True positive (TP): articles correctly classified as real.
- False negative (FN): real articles incorrectly predicted as fake.

### D. Cohen's Kappa Metric

Cohen's Kappa Score (CKS) evaluates the level of agreement between predicted and actual class labels, adjusting for the agreement that might occur by chance. It is particularly

useful in scenarios with imbalanced datasets [25]. Eq. (4) defines CKS:

$$CKS = \frac{P_0 - P_e}{1 - P_e} \quad (4)$$

where,

- P0: the observed accuracy of the model.
- Pe: the expected agreement by chance.

### E. F1-Score

F1-score is a standard metric for binary classification that combines precision and recall into a single value. It is defined as the harmonic average of the two, making it especially useful when the class distribution is imbalanced [25]. Eq. (5) shows the formulation:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

### F. Receiver Operating Characteristic: Area Under the Curve (ROC-AUC)

ROC-AUC measures a model's ability to distinguish between positive and negative classes across different decision thresholds. It is particularly valuable in cases of imbalanced datasets, as it evaluates the trade-off between sensitivity (recall) and specificity [25].

By employing a variety of evaluation metrics, such as Accuracy, Precision, Recall, F1-score, Cohen's Kappa, and ROC-AUC, we can obtain a comprehensive understanding of the model's performance. These metrics collectively highlight the model's strengths and limitations, from its ability to correctly classify real and fake articles to its robustness in handling imbalanced data. Such a comprehensive evaluation facilitates informed decision-making in model selection, optimization, and deployment.

## VI. Results

In this study, a stacking ensemble model was constructed for fake news detection using machine learning classifiers combined with the transformer-based SBERT model for sentence embeddings. A total of 72,095 news articles from the WELFake dataset were processed, with SBERT encoding each article into dense semantic embeddings of dimension 384. These embeddings were used to train four base learners—Random Forest, Logistic Regression, Multilayer Perceptron (MLP), and XGBoost—through a 3-fold stratified cross-validation scheme to generate out-of-fold (OOF) predictions. The aggregated OOF predictions served as meta-features for the XGBoost meta-learner, which produced the final classification results. On a holdout set comprising 14,419 articles, the stacking ensemble achieved an accuracy of 92.7%, precision of 92.0%, recall of 94.0%, and an F1-score of 93.0%. In addition, the model reached a Cohen's Kappa score of 0.85, reflecting strong agreement beyond chance, and a ROC-AUC of 0.98, indicating excellent discrimination ability between real and fake articles. The detailed classification report further highlights balanced performance across both classes, with fake news (label 0) achieving a precision of 0.94 and real news (label 1) attaining a recall of 0.94.

These results confirm that the SBERT-based stacking model effectively integrates the strengths of multiple learners, capturing both semantic depth and diverse decision boundaries. By leveraging SBERT embeddings with ensemble learning, the proposed model demonstrates robust performance and strong generalization in large-scale fake news detection tasks.

### A. Stacking Model Performance

In this work, multiple evaluation metrics were utilized to assess the performance of the proposed stacking ensemble model, which combines SBERT embeddings with multiple machine learning classifiers. The outcomes across these metrics are summarized in Table II.

TABLE II. The Stacking Model Performance Results

| Accuracy Score | Precision Score | Recall Score | F1 Score | ROC-AUC Score | Cohen's Kappa Score |
|---|---|---|---|---|---|
| 0.9274 | 0.9203 | 0.9402 | 0.9301 | 0.9793 | 0.8546 |

## VII. Discussion

### A. Performance of the Proposed Model

In this study, we developed a stacking ensemble model for fake news detection that combines both traditional machine learning models (Random Forest, Logistic Regression, and MLP) with SBERT embeddings and XGBoost as the meta-learner. Our proposed model achieved an overall accuracy of 92.7%, supported by a precision of 92.0%, a recall of 94.0%, an F1-score of 93.0%, a Cohen's Kappa of 0.85, and an ROC-AUC of 0.97. These results highlight the ability of our model to effectively balance both precision and recall, ensuring robustness in distinguishing between fake and real news.

### B. Analysis of Class-Level Results

The classification report reveals that the model performs consistently across both classes. For real news (class 0), the model reached 92% F1-score, while for fake news (class 1), it achieved 93% F1-score. This balance indicates that the model is not biased toward one class, which is a common challenge in fake news detection tasks where class distributions may vary. Moreover, the slightly higher recall for fake news (94%) demonstrates the model's effectiveness in minimizing false negatives—an important aspect in preventing the spread of misinformation.

### C. Contribution of Ensemble Learning

The success of the proposed model lies in the integration of diverse learners and SBERT embeddings. While base learners such as Random Forest and Logistic Regression provide stable decision boundaries, SBERT ensures high-quality semantic representations of text, and the meta-learner XGBoost effectively combines their predictions. This layered approach enhanced the model's generalization capability, leading to improved performance compared to using individual models alone.

### D. Comparative Evaluation with Existing Work

Compared with recent studies on fake news detection [26], [27], our proposed SBERT-based stacking ensemble demonstrates superior predictive performance and

generalization. Previous work, such as Wu et al. (2024) and Zhu et al. (2025), achieved moderate performance on their respective datasets. Wu et al. introduced a semantic-aware evidence-based framework with data augmentations and contrastive learning, reaching an F1-macro of 80.1% on Snopes-hard and 62.1% on PolitiFact-hard. Zhu et al. proposed a MIBKA-CNN-BiLSTM hybrid model, achieving up to 88.05% accuracy and 86.71% F1-score on a Chinese fake news dataset, but its effectiveness remained limited to the Chinese social media domain.

In contrast, our model leverages Sentence-BERT embeddings to encode rich semantic information and combines multiple heterogeneous base learners (XGBoost, Random Forest, Logistic Regression, and MLP) with an XGBoost meta-learner. Systematic optimization using GridSearchCV and Optuna, along with k-fold cross-validation, enabled our approach to achieve 92.74% accuracy, 93.01% F1-score, and 97.93% ROC-AUC on the WELFake dataset, substantially surpassing prior benchmarks.

This comparison highlights that integrating semantic-rich embeddings with ensemble meta-learning not only enhances generalization but also provides better stability across performance metrics compared to both optimization-driven hybrid models and transformer-only architectures. Moreover, our approach addresses prior limitations in semantic understanding and class balance, demonstrating a robust framework suitable for real-world fake news detection.

### *E. Insights and Implications*

The findings underscore the effectiveness of integrating transformer-based embeddings with ensemble learning techniques. The high ROC-AUC (0.97) highlights the model's discriminative capability, while the Cohen's Kappa (0.85) indicates strong agreement beyond chance, reflecting robustness and reliability. Importantly, the balanced performance across fake and real news classes suggests the model's potential applicability to real-world scenarios, where the cost of misclassification is high. These results provide a solid foundation for extending the model to multilingual datasets, social media streams, and large-scale news corpora, where detecting misinformation is increasingly critical.

## VIII. Conclusion

Fake news continues to pose a significant threat to the integrity of information, public trust, and informed decision-making worldwide. This study demonstrates that integrating semantic text embeddings with ensemble learning techniques can effectively enhance fake news detection, achieving high accuracy, precision, and recall. The results highlight the potential of such models to identify misinformation with a balanced consideration of false positives and false negatives, providing a reliable tool for monitoring information quality. The findings have practical implications for multiple stakeholders: policymakers can leverage these methods to design evidence-based regulations against misinformation, platform designers can implement automated detection systems to safeguard users, and practitioners can utilize such tools to monitor and mitigate the spread of false content. Moreover, the approach underscores the importance of combining contextual understanding with robust classification techniques to address complex challenges in information verification. For future work, extending the model to multilingual datasets, diverse domains, and real-time social media streams could further enhance its utility. Additionally, incorporating explainable AI techniques may improve transparency, increase user trust, and support ethical deployment of automated misinformation detection.

## References

[1] F. A. Alshuwaier and F. A. Alsulaiman, "Fake News Detection Using Machine Learning and Deep Learning Algorithms: A Comprehensive Review and Future Perspectives", *Computers*, vol. 14, no. 9, p. 394, Sept. 2025, doi: 10.3390/computers14090394.

[2] B. Hu, Z. Mao, and Y. Zhang, "An overview of fake news detection: From a new perspective", *Fundamental Research*, vol. 5, no. 1, pp. 332–346, Jan. 2025, doi: 10.1016/j.fmre.2024.01.017.

[3] S. Shah and S. Patel, "A Comprehensive Survey on Fake News Detection Using Machine Learning", *Journal of Computer Science*, vol. 21, no. 4, pp. 982–990, Apr. 2025, doi: 10.3844/jcssp.2025.982.990.

[4] J. Poushter, M. Smerkovich, M. Fagan, and A. Prozorovsky, "Free Expression Seen as Important Globally, but Not Everyone Thinks Their Country Has Press, Speech and Internet Freedoms".

[5] J. Alghamdi, S. Luo, and Y. Lin, "A comprehensive survey on machine learning approaches for fake news detection", *Multimed Tools Appl*, vol. 83, no. 17, pp. 51009–51067, Nov. 2023, doi: 10.1007/s11042-023-17470-8.

[6] D. Mouratidis, A. Kanavos, and K. Kermanidis, "From Misinformation to Insight: Machine Learning Strategies for Fake News Detection", *Information*, vol. 16, no. 3, p. 189, Feb. 2025, doi: 10.3390/info16030189.

[7] T. T. Aurpa et al., "Deep transformer-based architecture for the recognition of mathematical equations from real-world math problems", *Heliyon*, vol. 10, no. 20, p. e39089, Oct. 2024, doi: 10.1016/j.heliyon.2024.e39089.

[8] I. Ni'mah et al., "A simple contrastive embedding framework for low-resource fake news detection", *Neural Comput & Applic*, vol. 37, no. 26, pp. 21407–21433, Sept. 2025, doi: 10.1007/s00521-025-11467-0.

[9] R. Mohawesh, S. Maqsood, and Q. Althebyan, "Multilingual deep learning framework for fake news detection using capsule neural network", *J Intell Inf Syst*, vol. 60, no. 3, pp. 655–671, June 2023, doi: 10.1007/s10844-023-00788-y.

[10] D. Mishra, S. M. Tripathi, A. Chaurasia, and P. K. Chaurasia, "A Review on Ensemble Learning Methods: Machine Learning Approach", *Int. J. Res. Publ. Rev.*, vol. 6, no. 2, pp. 3795–3803, Feb. 2025, doi: 10.55248/gengpi.6.0225.0971.

[11] M. Al-alshaqi, D. B. Rawat, and C. Liu, "Ensemble Techniques for Robust Fake News Detection: Integrating Transformers, Natural Language Processing, and Machine Learning", *Sensors*, vol. 24, no. 18, p. 6062, Sept. 2024, doi: 10.3390/s24186062.

[12] Y. Kumar, "Fake News Detection Model Ndetect Using Ensemble Machine Learning Techniques", *International Journal of Environmental Sciences*, vol. 11, no. 12, 2025.

[13] M. Q. Alnabhan and P. Branco, "Fake News Detection Using Deep Learning: A Systematic Literature Review", *IEEE Access*, vol. 12, pp. 114435–114459, 2024, doi: 10.1109/ACCESS.2024.3435497.

[14] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects", *IEEE Access*, vol. 10, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[15] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", Aug. 27, 2019, *arXiv:1908.10084*, doi: 10.48550/arXiv.1908.10084.

[16] A. Cutler, D. Cutler, and J. Stevens, "Random Forests", in *Machine Learning - ML*, vol. 45, 2011, pp. 157–176. doi: 10.1007/978-1-4419-9326-7_5.

[17] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview", *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, June 2024, doi: 10.58496/BJML/2024/007.

[18] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[19] C.-C. Chang, Y.-Z. Li, H.-C. Wu, and M.-H. Tseng, "Melanoma Detection Using XGB Classifier Combined with Feature Extraction and K-Means SMOTE Techniques", *Diagnostics*, vol. 12, no. 7, p. 1747, July 2022, doi: 10.3390/diagnostics12071747.

[20] L. Dube and T. Verster, "Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models", *DSFE*, vol. 3, no. 4, pp. 354–379, 2023, doi: 10.3934/DSFE.2023021.

[21] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis, "Multilayer Perceptron and Neural Networks", vol. 8, no. 7, 2009.

[22] M. Maalouf, "Logistic regression in data analysis: an overview", *IJDATS*, vol. 3, no. 3, p. 281, 2011, doi: 10.1504/IJDATS.2011.041335.

[23] D. Berrar, "Cross-Validation", in *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, 2019, pp. 542–545. doi: 10.1016/B978-0-12-809633-8.20349-X.

[24] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: Word Embedding Over Linguistic Features for Fake News Detection", *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 4, pp. 881–893, Aug. 2021, doi: 10.1109/TCSS.2021.3068519.

[25] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics", *AJBR*, pp. 4023–4031, Nov. 2024, doi: 10.53555/AJBR.v27i4S.4345.

[26] Y. Wu, Y. Xiao, M. Hu, M. Liu, P. Wang, and M. Liu, "Towards Robust Evidence-Aware Fake News Detection via Improving Semantic Perception".

[27] S. Zhu, G. Mu, J. Ma, and X. Li, "An Enhanced MIBKA-CNN-BiLSTM Model for Fake Information Detection", *Biomimetics*, vol. 10, no. 9, p. 562, Aug. 2025, doi: 10.3390/biomimetics10090562.