DIGITAL TRANSFORMATION AND ADMINISTRATION INNOVATION

Citation: Awad Dawood, K., Aghaee Ghazvini, G., Albu-Rghaif, A., & Majidi, F. (2026). An Improved Ensemble Framework for Social Media Fake News Detection Using RoBERTa Embeddings and the Whale Optimization Algorithm. *Digital Transformation and Administration Innovation*, 4(3), 1-24.

Received date: 2025-12-24 Revised date: 2026-04-12 Accepted date: 2026-04-19 Published date: 2026-05-01

An Improved Ensemble Framework for Social Media Fake News Detection Using RoBERTa Embeddings and the Whale Optimization Algorithm

Kareem Awad Dawood[1], Golnaz Aghaee Ghazvini [2]*, Ali Albu-Rghaif [3], Fariba Majidi[1]

1. Department of Computer Engineering, Isf.C., Islamic Azad University, Isfahan, Iran
2. Department of Computer, Dol.C., Islamic Azad University, Isfahan, Iran
3. Department of Computer Engineering, Diyala Branch, Diyala University, Diyala, Iraq

*Correspondence: G.aghaee@iau.ac.ir

Abstract

The increasing prevalence of fake news on social media has raised serious concerns due to its impact on public perception and decision-making. In response, this study introduces a hybrid ensemble learning framework enhanced by RoBERTa-based feature representation and optimized using the Whale Optimization Algorithm (WOA). The proposed approach aims to effectively capture deep semantic patterns in textual data while improving classification performance through adaptive parameter tuning. RoBERTa is employed to generate high-quality textual embeddings from news content, which are then fed into a set of base classifiers to ensure robustness and diversity in predictions. The WOA algorithm fine-tunes the ensemble model parameters, resulting in improved convergence and reduced error rates. The model was evaluated using two well-known fake news datasets, LIAR and ISOT. On the LIAR dataset, the proposed method achieved an accuracy of 98.17%, precision of 98.1%, recall of 97.8%, and an F1-score of 97.9%. On the ISOT dataset, it achieved an accuracy of 99.24%, precision of 99.3%, recall of 98.9%, and an F1-score of 99.1%. These results confirm the high reliability and balanced performance of the framework in distinguishing between true and false information across diverse content.

Keywords: Fake news detection, RoBERTa embeddings, ensemble learning, Whale Optimization Algorithm, evaluation metrics, LIAR dataset, ISOT dataset.

1. Introduction

The rapid proliferation of social media platforms has fundamentally transformed the way information is produced, disseminated, and consumed across societies. While these platforms have enabled unprecedented access to real-time information, they have also facilitated the widespread dissemination of misinformation and fake news, posing significant threats to public trust, social stability, and decision-making processes. Fake news, defined as deliberately fabricated or misleading information presented as legitimate news, has become a critical challenge in the digital information ecosystem. Its impact extends beyond individual misperceptions, influencing political polarization, economic behavior, and public health outcomes, particularly during crises such as elections or pandemics (Del Vicario et al., 2019; Vereshchaka et al., 2020). The rapid spread and amplification of false information through social networks necessitate the development of robust and scalable computational approaches for automatic fake news detection.

From a technological perspective, the complexity of fake news detection arises from several inherent challenges. First, fake news often mimics the linguistic and stylistic patterns of legitimate content, making it difficult to distinguish based solely on surface-level features. Second, the dynamic and evolving nature of misinformation strategies requires models that can generalize across domains and adapt to new patterns of deception. Third, the presence of noisy, high-dimensional, and unstructured textual data further complicates the classification task. Consequently, traditional rule-based or heuristic approaches have proven insufficient in addressing these challenges, leading to the adoption of machine learning and deep learning techniques as more effective alternatives (Alghamdi et al., 2024; Capuano et al., 2023).

Early approaches to fake news detection primarily relied on classical machine learning algorithms combined with handcrafted features such as lexical, syntactic, and semantic attributes. Techniques such as support vector machines, logistic regression, and decision trees demonstrated moderate success; however, their performance was limited by the quality and representativeness of manually engineered features (Fayaz et al., 2022; Khanam et al., 2021). These models often struggled to capture complex contextual dependencies and semantic nuances inherent in textual data, leading to reduced accuracy in real-world scenarios. Moreover, the increasing scale and diversity of social media content necessitated more advanced methods capable of automatic feature extraction and representation learning.

The emergence of deep learning has significantly advanced the field of fake news detection by enabling models to learn hierarchical and abstract representations directly from raw data. Neural network architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks have been widely applied to capture sequential and contextual information in text (Abualigah et al., 2024; Ahmad et al., 2022). Bidirectional LSTM models, in particular, have shown strong performance in modeling long-range dependencies and contextual relationships, thereby improving classification accuracy. Similarly, attention-based mechanisms and hybrid deep learning frameworks have been introduced to enhance feature representation and focus on informative parts of the text (Hashmi et al., 2024; Merryton & Augasta, 2024). Despite these advancements, deep learning models often require large amounts of labeled data and may suffer from overfitting or lack of interpretability.

In recent years, transformer-based architectures, particularly models such as BERT and its variants, have revolutionized natural language processing by providing contextualized word embeddings that capture deep semantic relationships. These models leverage self-attention mechanisms to encode contextual dependencies across entire sequences, enabling more accurate understanding of textual content. Studies have demonstrated that transformer-based embeddings significantly improve fake news detection performance compared to traditional embedding methods such as Word2Vec and GloVe (Al Ghamdi et al., 2024; Farokhian et al., 2024). Furthermore, advanced variants and extensions of BERT-like models have been explored for multimodal and cross-domain fake news detection tasks, highlighting their versatility and effectiveness (Raza et al., 2025; Wang et al., 2025). However, despite their superior representational capabilities, these models may still face challenges related to computational complexity and optimal parameter configuration.

To address the limitations of individual models, ensemble learning approaches have gained increasing attention in fake news detection research. Ensemble methods combine multiple base classifiers to leverage their complementary strengths, thereby improving robustness, generalization, and predictive performance. Techniques such as bagging, boosting, and stacking have been widely used to enhance classification accuracy in complex tasks (Elyassami et al., 2022; Hakak et al., 2021). By integrating diverse learning algorithms, ensemble models can reduce variance and bias, making them particularly suitable for handling heterogeneous and noisy data. Recent studies have also explored hybrid frameworks that combine deep learning models with traditional machine learning classifiers to achieve superior performance (Mahmud et al., 2024; Rustam et al., 2024).

Another critical challenge in fake news detection is the issue of class imbalance, where the number of fake news instances is often significantly lower than real news samples or vice versa. This imbalance can lead to biased models that favor the majority class, resulting in poor detection performance for minority classes. Various techniques, including resampling methods, cost-sensitive learning, and synthetic data generation, have been proposed to address this issue. In particular, Generative Adversarial Networks (GANs) have emerged as a powerful tool for generating realistic synthetic data, enabling the creation of balanced datasets and improving classifier performance (Adedoyin & Mariyappan, 2024; Madani et al., 2024). GAN-based

approaches enhance the diversity and representativeness of training data, thereby contributing to more robust and reliable detection models.

In addition to feature representation and data balancing, hyperparameter optimization plays a crucial role in determining the performance of machine learning and deep learning models. The selection of optimal hyperparameters can significantly impact model convergence, accuracy, and generalization capability. Traditional optimization techniques such as grid search and random search are often computationally expensive and inefficient, especially in high-dimensional parameter spaces. As a result, metaheuristic optimization algorithms inspired by natural processes have been increasingly adopted to address this challenge. Algorithms such as genetic algorithms, particle swarm optimization, and Harris hawks optimization have demonstrated effectiveness in optimizing model parameters for various applications (Choudhury & Acharjee, 2023; Heidari et al., 2019). More recently, nature-inspired algorithms such as the Whale Optimization Algorithm have shown promising results in efficiently exploring search spaces and identifying optimal solutions.

The integration of advanced feature extraction techniques, ensemble learning, and intelligent optimization algorithms has opened new avenues for improving fake news detection systems. Hybrid frameworks that combine deep contextual embeddings with optimized ensemble classifiers have demonstrated superior performance in capturing complex linguistic patterns and enhancing classification accuracy (Kishwar & Zafar, 2023; Zhang et al., 2023). Additionally, multimodal approaches that incorporate textual, visual, and contextual information have further improved detection capabilities by leveraging diverse data sources (Alsuwat & Alsuwat, 2025; Cui & Shang, 2025). These advancements highlight the importance of integrating multiple methodologies to address the multifaceted nature of fake news detection.

Despite the significant progress in this field, several research gaps remain. Many existing models focus primarily on improving accuracy without adequately addressing issues such as model interpretability, computational efficiency, and adaptability to evolving misinformation patterns. Furthermore, the majority of studies rely on single-modality data, limiting their applicability in real-world scenarios where information is often multimodal and heterogeneous. There is also a need for more robust frameworks that can effectively handle imbalanced datasets while maintaining high performance across different domains and data distributions (Farhangian et al., 2024; Goldani et al., 2021). Addressing these challenges requires the development of integrated approaches that combine advanced natural language processing techniques, ensemble learning strategies, and intelligent optimization methods.

In this context, the present study aims to contribute to the existing body of knowledge by proposing a novel hybrid framework that integrates RoBERTa-based contextual embeddings, ensemble learning, and Whale Optimization Algorithm-based hyperparameter tuning to enhance the accuracy and robustness of fake news detection in social media environments.

2. Methods and Materials

In this study, we propose an improved ensemble framework for detecting fake news on social media using powerful RoBERTa embeddings and optimizing the ensemble classifier through the Whale Optimization Algorithm (WOA). The proposed method is structured into five main stages. First, the input datasets, consisting of real and fake news articles from social media platforms, are loaded. In the second stage, data preprocessing is performed to remove irrelevant, noisy, or redundant elements. This stage includes sentence segmentation and word-level tokenization to ensure uniformity and consistency in text representation. In the third stage, the issue of class imbalance, which is common in fake news detection tasks, is addressed. For this purpose, a Generative Adversarial Network (GAN) is employed to synthetically generate realistic samples from the minority class, thereby balancing the dataset and improving classifier robustness. In the fourth stage, an ensemble classifier is constructed by integrating multiple base learners. Each base classifier is trained using embeddings generated by RoBERTa, which provide contextual representations of the input texts. The ensemble classifier is then optimized using the Whale Optimization Algorithm to determine the optimal combination of classifiers and their corresponding parameters, thereby improving both generalization and predictive accuracy. Finally, the performance of the proposed model is evaluated using standard evaluation metrics such as accuracy, precision, recall, and F1-score to demonstrate the effectiveness of the framework in detecting fake news.

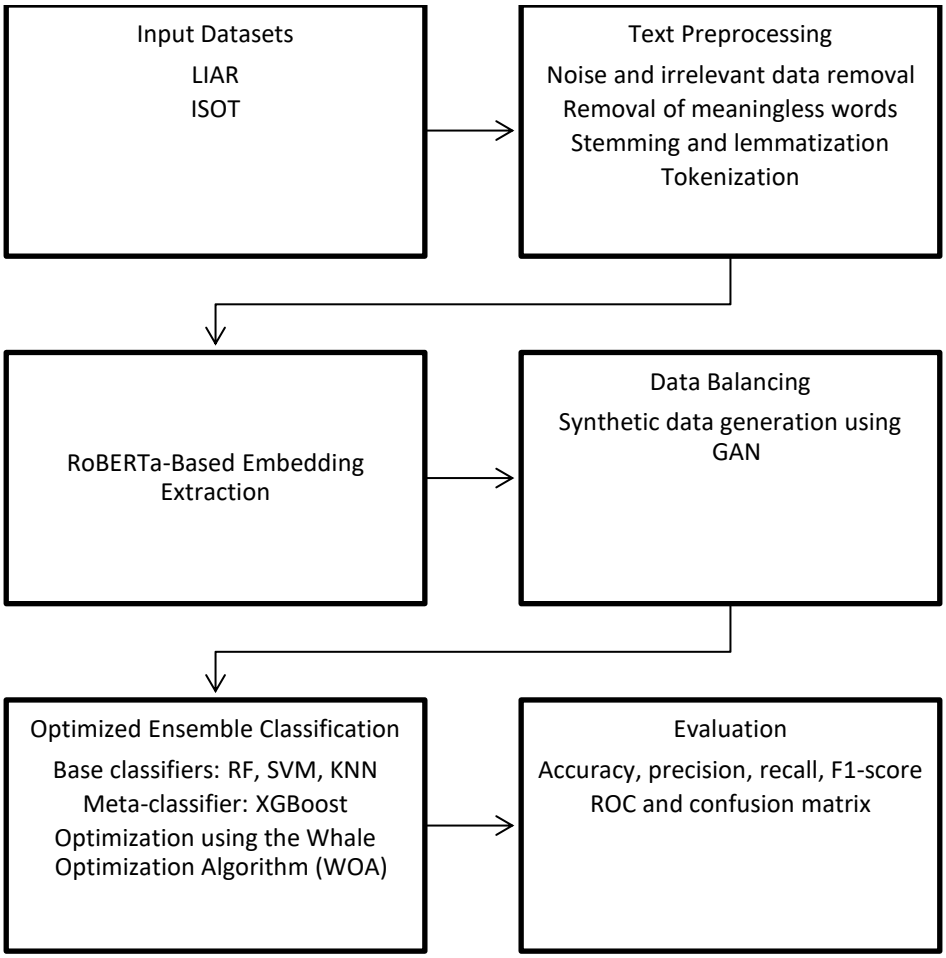

Figure 1. Block Diagram of the Proposed Method

Data Preprocessing

Data preprocessing is a critical step in preparing unstructured text from social media platforms for effective fake news detection. Since the proposed method leverages textual embeddings from the RoBERTa language model, the preprocessing pipeline is designed to preserve meaningful textual information while removing irrelevant noise. The steps are as follows:

Dataset Selection: A labeled social media dataset containing examples of fake and real news is selected. The dataset includes textual posts along with metadata and ground truth labels.

Tokenization: Each sentence is split into individual tokens using a tokenizer compatible with RoBERTa, preserving linguistic structure and subword units.

Stopword Removal: Common words that do not significantly contribute to semantic meaning, such as “the,” “is,” and “and,” are removed to reduce dimensionality.

Noise Removal: Punctuation marks, hyperlinks, user mentions, hashtags, emojis, and other non-alphabetic characters are removed to enhance textual clarity.

Lemmatization: Instead of stemming, lemmatization is applied to preserve the correct word forms, which better aligns with the pre-trained vocabulary of RoBERTa.

Vectorization (Embedding): Unlike traditional bag-of-words or TF-IDF approaches, the preprocessed text is directly transformed into dense textual embeddings using the pre-trained RoBERTa model, capturing deep semantic relationships between words.

Final Preprocessed Dataset: The resulting dataset consists of RoBERTa embeddings for each text sample, ready for subsequent training and optimization tasks.

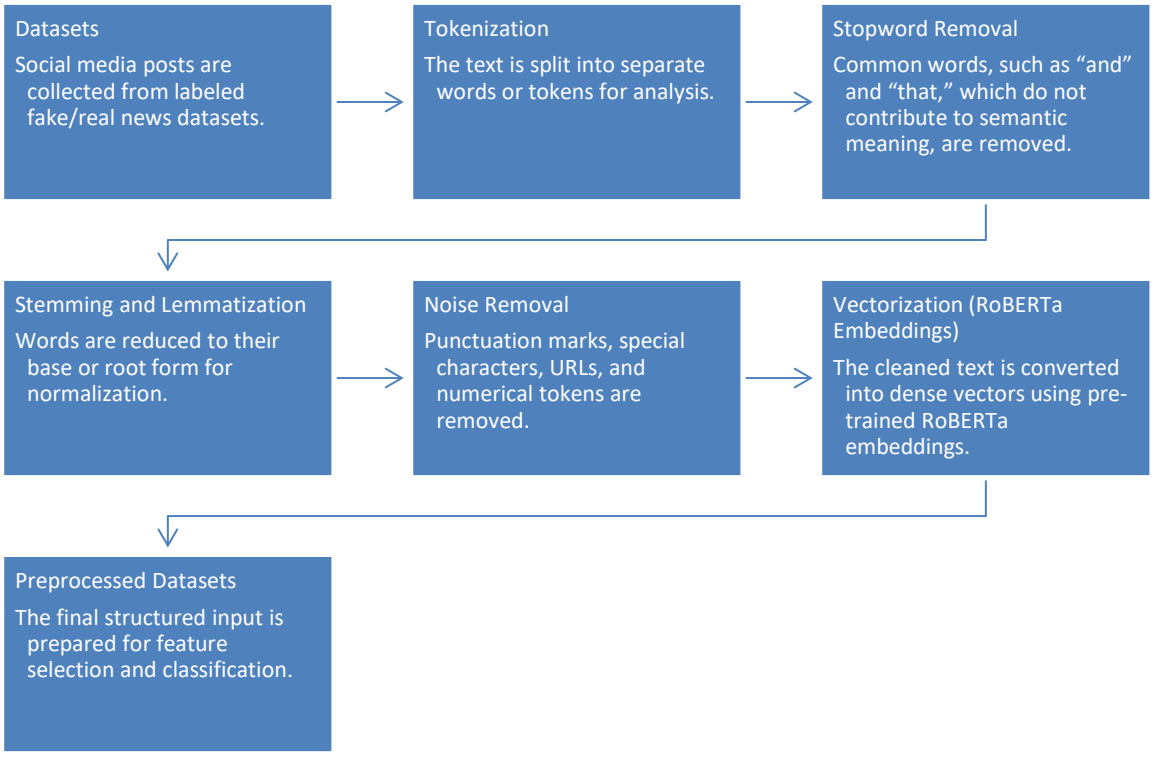

Figure 2. Overview of the Text Preprocessing Procedure for RoBERTa-Based Fake News Detection

Dataset Description

This study utilized two publicly available fake news datasets, namely LIAR and ISOT, each offering diverse characteristics for evaluating the robustness of the proposed detection model.

LIAR Dataset: Introduced by Wang (2017), LIAR is a large-scale benchmark dataset consisting of 12,836 short political statements labeled with six fine-grained categories (pants-fire, false, barely-true, half-true, mostly-true, and true). Each sample is accompanied by metadata such as speaker, context, and subject. Its brevity and rich label diversity make it particularly suitable for sentence-level fake news detection.

ISOT Dataset: Developed by Ahmed et al., the ISOT dataset comprises 44,898 news articles, divided into 21,417 labeled as "fake" and 23,481 labeled as "real." These articles are longer and more narrative, enabling the detection of deceptive writing patterns at the paragraph level. Unlike LIAR, ISOT lacks metadata but provides richer contextual information within textual content.

Tokenization

Both datasets were tokenized using the WordPiece tokenizer integrated into the BERT architecture. This tokenizer is capable of handling out-of-vocabulary words by decomposing them into subword units, thereby improving generalization. For the LIAR dataset, tokenization was performed at the sentence level, preserving the short-text nature of the statements. In contrast, the ISOT dataset, due to its longer articles, required document-level tokenization. Special tokens such as [CLS] and [SEP] were added to the tokenized sequences for compatibility with BERT.

Stopword Removal

Given the context-sensitive nature of fake news, a conservative approach to stopword removal was adopted. For both datasets, the NLTK stopword list was used; however, contextually significant function words were retained when they contributed to meaning (e.g., negations such as "not" and "never"). For instance, in the LIAR dataset, preserving negations was crucial due to the short sentence length. In the ISOT dataset, stopword removal was slightly more aggressive, balancing textual clarity and semantic preservation.

Lemmatization and Stemming

To normalize textual content, lemmatization was preferred over stemming to maintain grammatical correctness, which is essential for BERT-based models. The spaCy library was used for lemmatization in both datasets. For LIAR, the focus was on reducing redundancy in frequently occurring political terms. For ISOT, lemmatization helped reduce vocabulary size while preserving subtle linguistic structures used in fabricated narratives. Since BERT utilizes contextual embeddings, lemmatization also ensured cleaner input sequences without compromising syntactic integrity.

Text Denoising and Normalization

Text denoising involves removing irrelevant or redundant characters such as digits, punctuation, and special symbols that do not enhance semantic interpretation. Simultaneously, normalization techniques such as case folding (e.g., converting all text to lowercase) are applied to ensure consistency across datasets. This unified textual representation reduces inconsistencies in feature encoding and supports more robust and reliable model learning.

Feature Extraction and Data Balancing

In this stage, the objective is to represent the preprocessed textual data in a meaningful numerical format suitable for classification, while also addressing potential class imbalance issues in the training data. For this purpose, a hybrid feature extraction approach is employed that combines semantic embeddings from RoBERTa with statistical term weighting using TF-IDF. This integration enables the model to retain both contextual meaning and the relative importance of terms across the corpus. After extracting feature vectors, the next step involves applying Generative Adversarial Networks (GANs) to synthetically generate samples for the minority class (e.g., fake news samples), ensuring class balance and preventing classifier bias toward the majority class. To obtain rich, context-aware document representations, a hybrid embedding scheme is utilized that integrates RoBERTa embeddings with TF-IDF weighting. While RoBERTa captures contextual meaning within sentences, TF-IDF ensures that more informative words contribute more significantly to the final representation. The integration of these two components results in a more expressive and discriminative feature space.

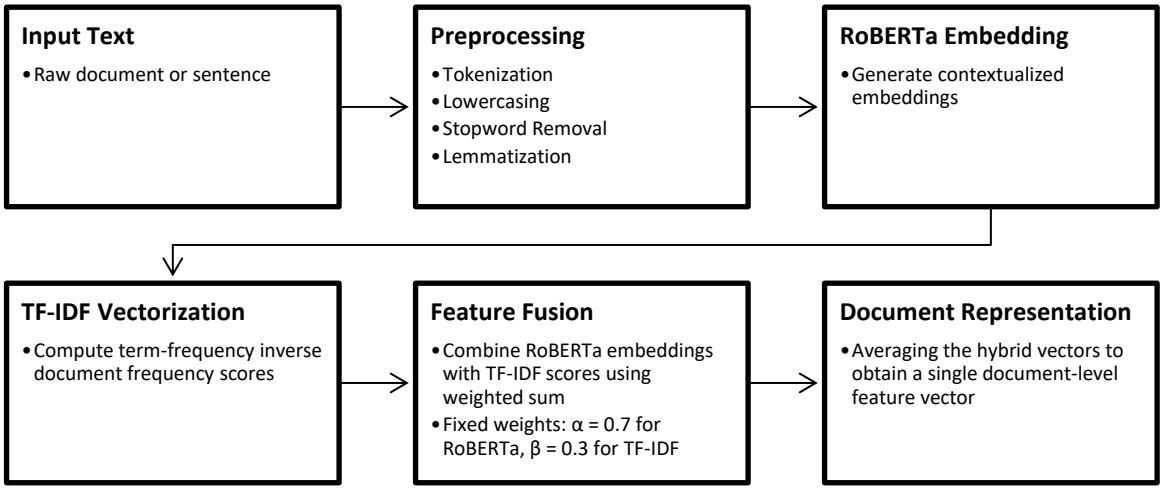

Figure 3. Block Diagram: Hybrid Embedding with RoBERTa and TF-IDF

Assume that $d \in D$is a document composed of a sequence of tokens $t_1, t_2, \ldots, t_n$. For each token t_i, the following is computed:

(2) $\text{TF-IDF}(t_i, d) = \text{TF}(t_i, d) \times \log \frac{N}{DF(t_i)}$

where $\text{TF}(t_i, d)$ is the term frequency of token $t_i$in document d, $\text{DF}(t_i)$ is the number of documents containing t_i, and $N$is the total number of documents in the dataset.

Textual embedding using RoBERTa:

(2) $V_{RoB}(t_i) = RoBERTa(t_i)$

Weighted hybrid vector for document d:

(3) $V_{hybrid}(d) = \sum_{i=1}^{n} TF\text{-}IDF(t_i, d) \times V_{RoB}(t_i)$

This operation results in a dense vector representation $V_{hybrid} \in \mathbb{R}^d$, where $d$corresponds to the embedding dimension (typically 768 for RoBERTa-base).

After computing the weighted sum of token embeddings, an averaging method was used to generate a single document-level vector. This approach was selected due to its simplicity, effectiveness in preserving global contextual information, and proven performance in prior studies involving sentence and document representation. Additionally, a simple grid search mechanism was employed to fine-tune the scalar coefficients α and β, which control the contributions of RoBERTa embeddings and TF-IDF scores, respectively.

(3) $V_{final} = \alpha \cdot V_{RoBERTa} + \beta \cdot V_{TF-IDF}$

This allows the model to emphasize either semantic or statistical information depending on downstream task performance.

Table 1. Steps of the Proposed Hybrid Feature Extraction Algorithm Based on BERT and TF-IDF

Step	Output Example
Input Sentence	"The news article was completely false and misleading."
Text Preprocessing	"news article completely false misleading"
Tokenization	['news', 'article', 'completely', 'false', 'misleading']
TF-IDF Computation	{'news': 0.35, 'article': 0.30, 'completely': 0.20, 'false': 0.40, 'misleading': 0.45}
BERT Vectorization for Each Token	{'news': [..., 0.56, 0.12], 'article': [..., 0.44, 0.22], ...}
Weighted Vector Computation	{'news': 0.35 × [0.12, 0.56, ...], ..., 'misleading': 0.45 × [0.28, 0.73, ...]}
Aggregation of Weighted Vectors	Final document vector (e.g., [0.42, 1.37, ..., 0.83])
Hybrid Embedding Output	$V_{hybrid} = [0.42, 1.37, \dots, 0.83]$

Generation of Synthetic Samples Using GAN

To improve the performance of the ensemble classifier and reduce the problem of data imbalance, Generative Adversarial Networks (GANs) are used to generate synthetic samples. GAN is a class of deep learning models capable of generating high-quality synthetic data similar to the distribution of real data. This process balances the dataset, particularly by increasing the number of minority-class samples, which is essential for improving the learning capacity of the classifier.

A standard GAN consists of two neural networks:

Generator (G): It receives a random noise vector $z \sim p_z(z)$and generates synthetic samples $G(z)$that are intended to resemble real data.

Discriminator (D): It determines whether a sample is real (from the actual dataset) or fake (generated by the generator). Its task is to maximize the probability of correctly distinguishing real and synthetic data.

The objective function for GAN training is a minimax optimization problem and is defined as follows:

$$\min_G \max_D V(D,G) = E_{x \sim p_{data}(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log (1 - D(G(z)))]$$

where $x$denotes real samples, $z$denotes input noise, $p_{data}(x)$is the real data distribution, and $p_z(z)$is the noise distribution. By simultaneously training the generator and discriminator, the generator gradually learns to produce samples that are highly similar to real data. In our proposed framework, the generated synthetic samples are selectively added to the training data to strengthen underrepresented classes and reduce model bias, thereby increasing predictive accuracy.

Whale Optimization Algorithm (WOA)-Optimized Two-Layer Ensemble Learning Framework

This section introduces a two-layer ensemble learning framework enhanced by the Whale Optimization Algorithm (WOA). This framework is specifically designed to improve the accuracy of fake news detection on social media. The overall structure of the model is shown in Figure 7. In the first layer, the input is processed after preprocessing and feature extraction using RoBERTa-based contextual embeddings to extract the semantic and contextual nuances of the text. A diverse set of base classifiers, such as SVM, Random Forest, and other classifiers, is then applied to these features. Their outputs are not hard labels, but rather a probability distribution over the classes. These soft outputs are transferred to the next stage as meta-features. To maximize the performance of each base classifier, the Whale Optimization Algorithm (WOA) is used to tune the hyperparameters. This algorithm simulates the collective hunting behavior of humpback whales and efficiently searches the hyperparameter space to find the best values. In the second layer, the optimized probability vectors are fed into a meta-classifier. This meta-classifier aggregates the outputs of the first layer and provides the final prediction. This hierarchical structure

combines the power of multiple classifiers with the optimization capability of WOA, resulting in a more robust and accurate model for fake news detection.

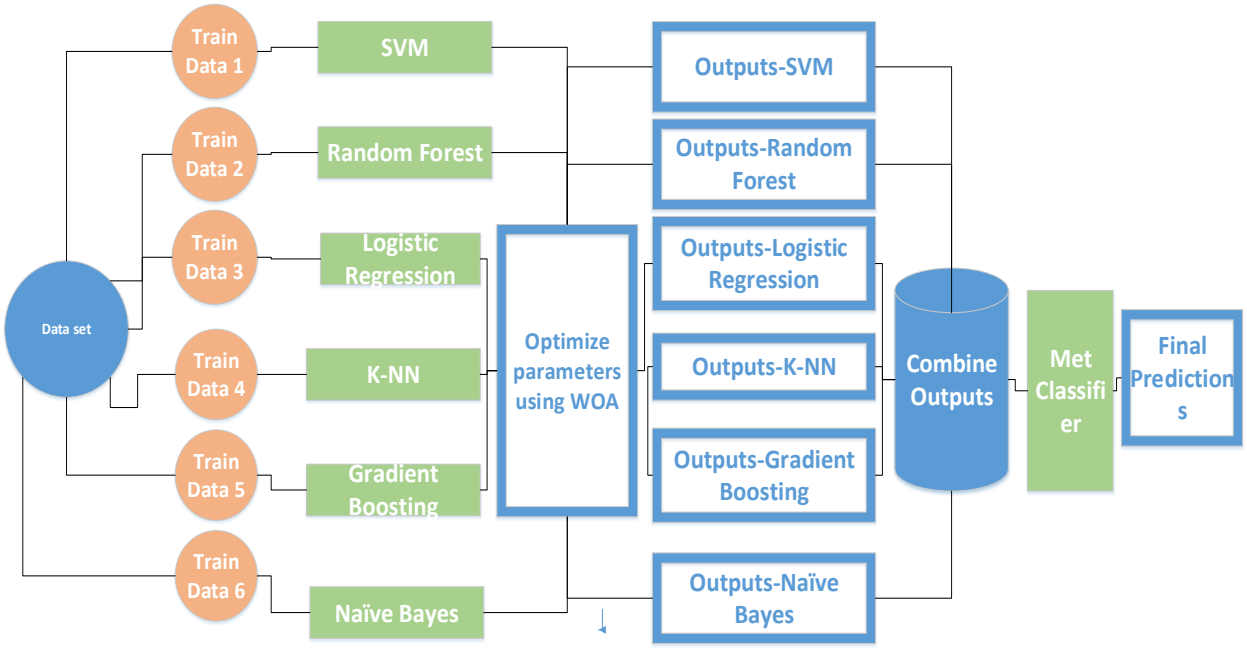

Figure 4. Overview of the Proposed Ensemble Architecture

As shown in Figure 4, the proposed architecture adopts a structured two-stage ensemble framework designed for robust and accurate fake news detection on social media platforms.

This framework consists of two main layers:

First Stage – Construction and Optimization of Base Models: In this stage, the preprocessed and feature-extracted data, using hybrid RoBERTa-TF-IDF embedding, are first fed into a set of base classifiers. At this step, the Whale Optimization Algorithm (WOA) is applied to tune the hyperparameters of each classifier. The objective of this stage is to identify the best hyperparameter configuration that maximizes classifier performance in the fake news detection task. Instead of producing hard labels, each base model outputs a class-probability vector, which provides a richer probabilistic interpretation.

Second Stage – Meta-Learning and Final Classification: The probability vectors generated by the base models are concatenated to form a higher-level feature representation. This meta-feature vector is fed into a high-level meta-classifier, which produces the final prediction by combining the outputs of the base models. This two-layer structure enables the model to simultaneously benefit from the individual strengths of each classifier and their collective decision-making capability.

Construction of the Base Model and Its Features

In the first stage of the ensemble framework, six diverse machine learning algorithms are used as base models:

Support Vector Machine: Known for its high efficiency in high-dimensional spaces and highly effective for large-scale feature sets.

Random Forest: An ensemble of decision trees that is resistant to overfitting and performs well on imbalanced datasets.

Logistic Regression: A lightweight, interpretable model suitable for linearly separable data.

K-Nearest Neighbors: A non-parametric method suitable for local patterns and small datasets.

Gradient Boosting: A powerful technique for modeling complex nonlinear relationships through the additive combination of decision trees.

Naïve Bayes: Highly efficient when the feature-independence assumption holds and provides fast and scalable classification.

This set of classifiers was selected to provide broad algorithmic diversity so that the ensemble model can generalize appropriately across different types of inputs.

Hyperparameter Optimization Using the Whale Optimization Algorithm (WOA)

In this section, the Whale Optimization Algorithm is used to fine-tune the hyperparameters of the base-level classifiers employed in the ensemble framework. Effective hyperparameter optimization plays an important role in improving the predictive performance of models, especially in complex and imbalanced domains such as fake news detection. Traditional hyperparameter-tuning methods, such as manual search or grid search, face scalability challenges, particularly when dealing with multiple classifiers and high-dimensional parameter spaces. These methods are not only time-consuming but also insufficiently capable of generalizing to diverse datasets. To address these limitations, WOA is used as a nature-inspired metaheuristic algorithm that simulates the bubble-net hunting behavior of humpback whales. By balancing exploration (global search) and exploitation (local search), this algorithm efficiently traverses the search space and converges to optimal solutions with fewer iterations than exhaustive methods.

As stated earlier, the proposed ensemble system uses six machine learning classifiers, each of which has a set of hyperparameters that significantly affect model performance. The main objective at this stage is to find the best set of hyperparameters as well as the optimal weight assigned to each classifier in the final decision-making process.

Therefore, two main components require optimization:

The hyperparameters of each base classifier

The weight of each classifier in the final ensemble decision

For this purpose, the Whale Optimization Algorithm (WOA) is used—a metaheuristic algorithm capable of searching large spaces and efficiently converging to optimal or near-optimal solutions.

The WOA-based optimization steps are as follows:

Solution Representation: Each whale in the population is a candidate solution that includes the hyperparameters of the six classifiers and their weights in the ensemble system.

Initialization: The whale population is initialized with random values within predefined ranges for the hyperparameters and weights.

Fitness Evaluation: The performance of each solution is evaluated using criteria such as accuracy, F1-score, or a weighted combination of metrics. Cross-validation is used to increase stability.

Encircling Prey: Using adaptive coefficient vectors, the whales move toward the current best solution.

Bubble-Net Attack Mechanism: Including two strategies:

Shrinking Search Circle Mechanism

Spiral Position Updating

Search for Prey: To prevent premature convergence, whales also move toward other randomly selected solutions.

Position Updating: The position of each whale is changed based on the best solution and the mechanisms described in the previous stage. This change is applied to both hyperparameters and classifier weights.

Stopping Criterion: The algorithm stops when:

The predefined number of iterations is completed, or

No significant improvement in fitness is observed.

WOA not only enables efficient hyperparameter optimization without manual trial and error, but also automatically adjusts the weight of each base classifier. This joint optimization increases reliability, reduces bias toward dominant models, and substantially improves the performance of the fake news detection system.

In the proposed framework, each whale in the population represents a position vector whose elements include the following:

$$Position = [C_{SVM}, kernel_{SVM}, gamma_{SVM}, n_{estimators_{RF}}, \max\ \_depth_{RF}, \min\ \_samples\_split_{RF}, C_{LogReg}, penalty_{LogReg}, n\_neighbors_{KNN}, weights_{KNN}, learning\_rate_{GB}, \max\ \_depth_{GB}, alpha_{NB}, weight_{SVM}, weight_{RF}, weight_{LogReg}, weight_{KNN}, weight_{GB}, weight_{NB}]$$

In this method, each element in the position vector represents a hyperparameter related to the base models or their contribution weight in the fusion stage. During the optimization process, the Whale Optimization Algorithm (WOA) updates the positions of whales in the search space to minimize the classification error of the validation set. Finally, the whale with the best performance provides the optimal set of hyperparameters and appropriate combination weights for the aggregated model.

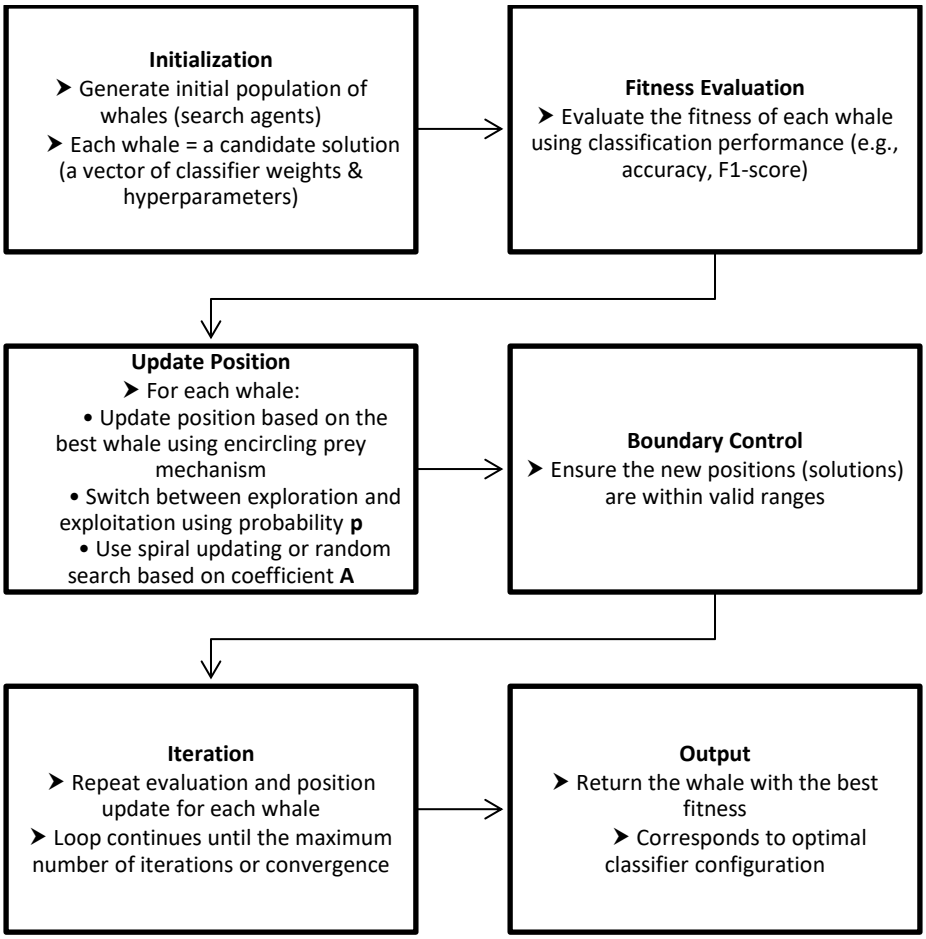

Figure 5. Procedure of the Whale Optimization Algorithm

To guide the search process in the Whale Optimization Algorithm, a fitness function is employed to evaluate the quality of each solution, that is, the position of each whale in the search space. In this study, the fitness function is designed to incorporate several important performance criteria related to classification quality.

The proposed multi-objective fitness function is defined as follows:

$$Fitness = \alpha \times (1 - Accuracy) + \beta \times (1 - Precision) + \gamma \times (1 - Recall) + \delta \times (1 - F1)$$

In this equation, accuracy evaluates the overall correctness of predictions. Precision measures the percentage of correctly predicted positive samples among all samples predicted as positive. Recall indicates the percentage of correctly predicted positive samples among all actual positive samples. The F1-score also establishes a balance between precision and recall.

The coefficients α, β, γ, and δmake it possible to adjust the importance of each criterion based on the needs of the problem. This fitness function encourages WOA to reduce the classification error rate and achieve balanced and optimized performance across all key indicators.

Final Classification Using a Meta-Classifier

After optimizing the base models using the Whale Optimization Algorithm, the final classification process is performed by employing a Random Forest meta-classifier. At this stage, the output probability vectors of the optimized base models are used as input features for the meta-classifier. Random Forest is a powerful ensemble learning method that reduces the probability of overfitting and improves model generalization by combining multiple decision trees. These characteristics make Random Forest a suitable option for complex and noisy textual data such as fake news. The operational stages of the Random Forest meta-classifier are as follows:

Input Feature Vector: The input to the meta-classifier consists of the probability vectors generated by all base classifiers that have previously been tuned and optimized by WOA.

Tree Construction: Each decision tree in the forest is trained on a bootstrapped sample of the meta-level dataset. At each node, only a random subset of features is selected for splitting to increase model diversity.

Voting Mechanism: In the prediction stage, each decision tree votes for a class. The final output of the meta-classifier is determined based on majority voting.

The important hyperparameters in the Random Forest meta-classifier are as follows:

n_estimators: The number of trees in the forest

max_depth: The maximum allowed depth for each tree

min_samples_split: The minimum number of samples required to split a node

min_samples_leaf: The minimum number of samples required to form a leaf node

max_features: The number of features considered at each split

bootstrap: Determines whether bootstrap sampling is used

criterion: The criterion for measuring split quality, such as Gini or entropy

The use of a Random Forest meta-classifier at this stage enables the proposed model to benefit from both the behavioral diversity of the base classifiers and the high robustness and generalization capability of Random Forest. As a result, the final output is produced with high confidence and reliable accuracy.

3. Findings and Results

This section presents the experiments conducted to evaluate the efficiency of the proposed hybrid model based on the Whale Optimization Algorithm.

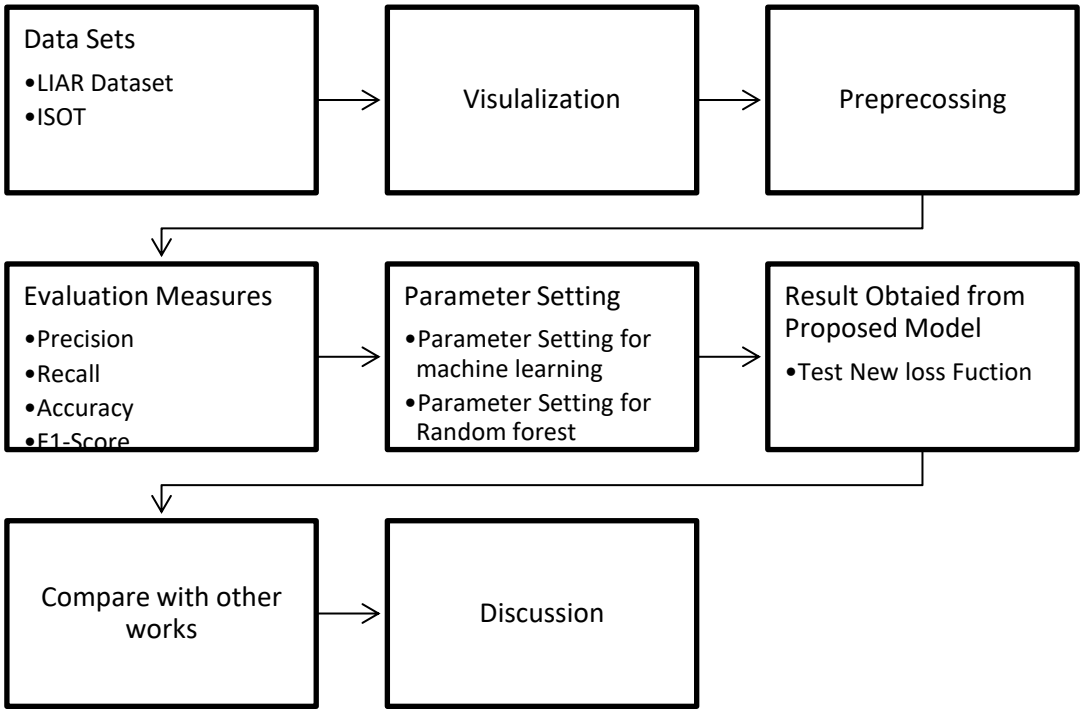

Figure 6. Evaluation Process of the Proposed Method

In this study, two datasets, LIAR and ISOT, were used, each with its own characteristics and challenges. The description of each dataset is provided below.

The ISOT dataset, introduced by Ahmed et al. (2018), includes more than 44,000 news articles classified into two categories: “fake news” and “real news.” Fake news items were collected from unreliable websites, while real news items were extracted from reliable news sources such as BBC. Each record in this dataset includes the following items:

- News title
- Full article text
- Corresponding label (Fake or Real)

The challenges of this dataset include the excessive length of some articles and the presence of topical bias. Nevertheless, ISOT is one of the most widely used datasets for training machine learning and deep learning models in the field of fake news detection.

The LIAR dataset, created by Wang (2017), contains more than 12,800 records of political statements collected from the PolitiFact website. This dataset classifies statements into six fact-checking categories:

- True
- Mostly True
- Half True
- Barely True
- False
- Pants-on-Fire

In addition to the statement text, each record includes features such as:

- Truthfulness label
- Statement subject, such as health, economy, politics, and others
- Metadata including speaker name, occupation, political affiliation, and publication date

The main challenges of this dataset are high linguistic diversity and the presence of a considerable volume of metadata. This dataset is widely used in research related to fake news detection, political text analysis, and the examination of linguistic biases.

Table 2. Comparison of the LIAR and ISOT Datasets

Features	ISOT	LIAR
Number of records	44,000	12,800
Number of classes	Two classes (Real/Fake)	Six truthfulness levels
Data type	News articles	Political statements
Metadata	None	Available (speaker, occupation, political affiliation)
Main challenge	Article length and topical bias	Linguistic diversity and metadata analysis

Data Preprocessing

Both the LIAR and ISOT datasets were initially imbalanced, such that some classes contained far fewer samples than others. To address this imbalance, Generative Adversarial Networks were used. GANs are capable of generating highly realistic synthetic data similar to the original samples and can balance the data distribution. The two charts below show the distribution of samples before and after GAN-based balancing in the LIAR and ISOT datasets.

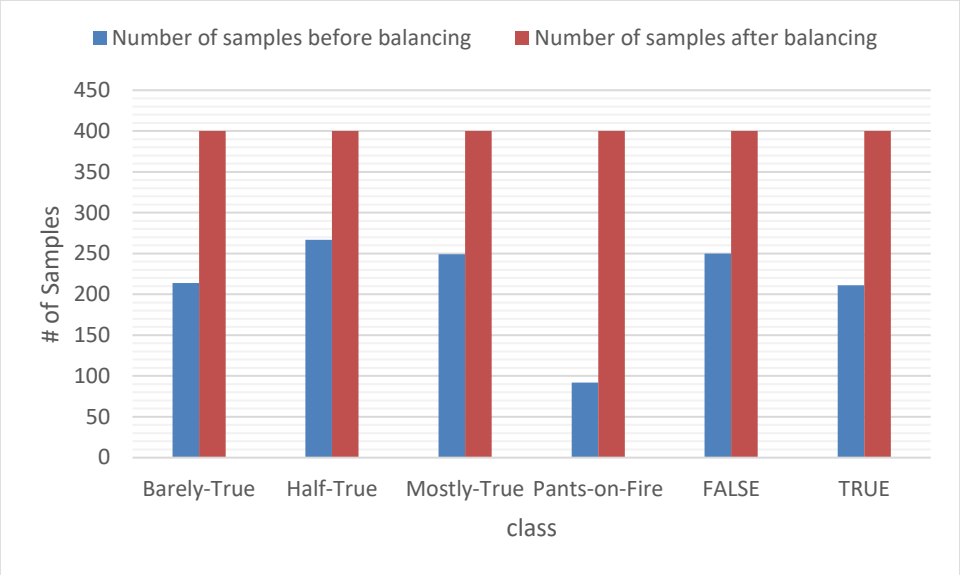

Figure 7. Class Distribution Before and After Balancing the LIAR Dataset

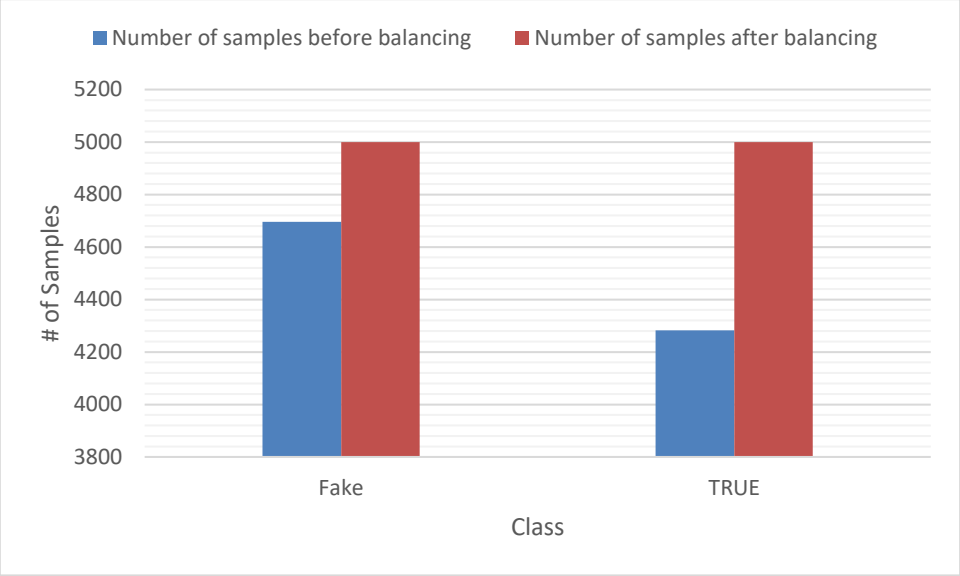

Figure 8. Class Distribution Before and After Balancing the ISOT Dataset

Evaluation Metrics

This section introduces three standard metrics: accuracy, precision, and recall. These three metrics were used to evaluate model performance. The corresponding mathematical relationships are presented in Equations (6) to (9).

(6) Precision = TP / (TP + FN)

(7) Recall = TP / (TN + FP)

(8) Accuracy = (TP + TN) / (TP + TN + FP + FN)

(9) F1-Score = (2 × (Precision × Recall)) / (Precision + Recall)

In the evaluation metric equations, TP and TN refer to samples that the model correctly classified into the proper class, whereas FP and FN represent incorrectly classified samples. Clearly, the higher the TP and TN values, the better the model performance. TP and TN values are located on the main diagonal of the confusion matrix and indicate the model's correct predictions for each class.

Hyperparameter Tuning

Optimization of Machine Learning Model Hyperparameters and Classifier Weights Using the Whale Optimization Algorithm

As explained in the proposed ensemble framework, model performance is highly dependent on selecting the best set of hyperparameters for each base classifier and determining their optimal weights in the final combination stage. To achieve this objective, the Whale Optimization Algorithm (WOA) was used as an efficient metaheuristic search technique.

Inspired by the social behavior of humpback whales and their bubble-net feeding pattern, WOA iteratively explores the search space to identify the configuration that yields the greatest improvement in classification performance. In this process, the position vector of each whale represents a proposed solution, including the hyperparameter values of the base classifiers along with their corresponding weights in the aggregation stage.

To evaluate the quality of each solution, WOA uses a fitness function designed based on criteria such as accuracy, precision, recall, and F1-score. The optimized hyperparameters for different classifiers on the LIAR and ISOT datasets are presented in Table 3.

Table 3. Hyperparameters Optimized by WOA

Algorithm	Parameter	Optimum Value for LIAR Dataset	Optimum Value for ISOT Dataset
SVM	C	10	1
SVM	Kernel	RBF	RBF
SVM	Gamma	0.1	scale
Random Forest	n_estimators	200	500
Random Forest	max_depth	20	30
Random Forest	min_samples_split	5	2
Random Forest	min_samples_leaf	2	1
Random Forest	max_features	sqrt	log2
Logistic Regression	C	0.1	1
Logistic Regression	solver	liblinear	lbfgs
Logistic Regression	penalty	l2	l1
Logistic Regression	max_iter	200	100
KNN	n_neighbors	7	5
KNN	weights	distance	uniform
KNN	metric	manhattan	euclidean
Gradient Boosting	n_estimators	200	100
Gradient Boosting	learning_rate	0.1	0.2
Gradient Boosting	max_depth	5	7
Gradient Boosting	min_child_weight	5	1
Gradient Boosting	subsample	0.8	1.0
Gradient Boosting	colsample_bytree	0.8	0.6
Naïve Bayes	alpha	0.1	1
Naïve Bayes	fit_prior	True	False

In addition to tuning the internal parameters of each classifier, WOA was also used to determine the optimal weights of the base models in the final aggregation process. These weights specify the influence of each classifier on the final decision. The optimal weights obtained for the LIAR and ISOT datasets are presented in Table 4.

Table 4. Optimal Classifier Weights Obtained by WOA

Classifier	LIAR Dataset	ISOT Dataset
Weight_SVM	0.30	0.25
Weight_RF	0.25	0.30
Weight_LogisticRegression	0.20	0.20
Weight_KNN	0.10	0.15
Weight_GradientBoosting	0.10	0.10
Weight_NaiveBayes	0.05	0.10

Random Forest Parameter Tuning

As explained in the previous section, the important parameters of the Random Forest algorithm were examined, and their best values are reported in Table 5. To determine these optimal parameters, a series of diverse experiments was conducted by varying different values.

In these experiments, the Entropy and Gini criteria were examined for measuring node quality, along with several different values for maximum depth from 2 to 10 features.

The accuracy results obtained from these experiments are shown in Figure 15.

These experiments helped select the best Random Forest configuration for final classification and improved model performance in fake news detection.

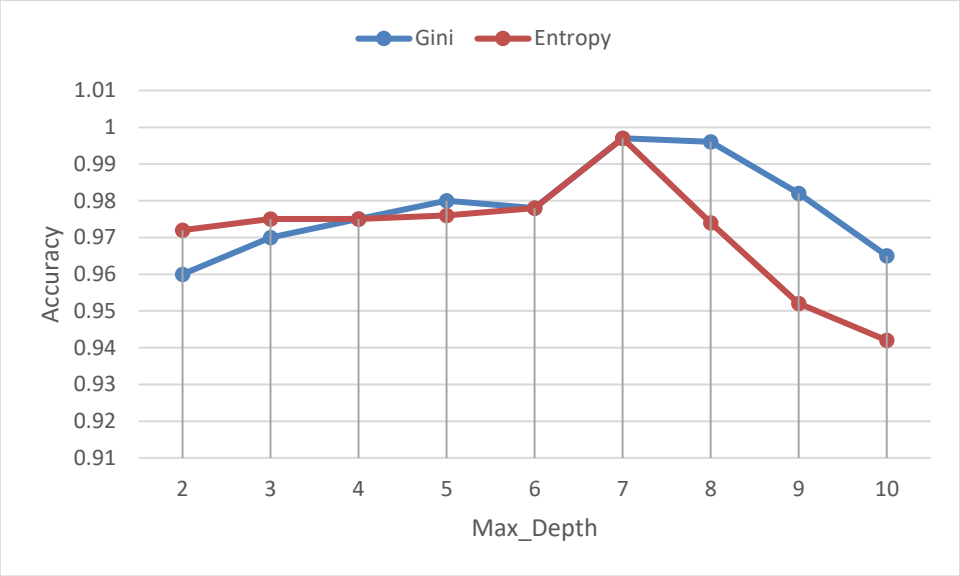

Figure 9. Effect of Gini and Entropy Criteria with Different Max_Depth Values on the Dataset

As shown in Figure 15, accuracy increases as Max_Depth increases. The highest accuracy in this experiment corresponds to the Gini criterion and Max_Depth equal to 7. After Max_Depth reaches 7, the increasing trend in accuracy stops and even declines. Therefore, the best value for this parameter was determined to be 7. Based on other experiments, the optimal values for the other Random Forest parameters are presented in Table 5.

Table 5. Optimal Hyperparameter Values of the Random Forest Algorithm

Hyperparameter	Value
n_estimators	100
max_features	Log2
max_depth	7
min_samples_split	2
min_samples_leaf	1
bootstrap	True
criterion	Gini

Experimental Results on the LIAR Dataset

This section presents the experimental results. A 5-fold cross-validation method was used to evaluate the model. Figure 16 shows the confusion matrix related to the LIAR dataset and provides an overall view of model performance in identifying each class.

Table 6 shows the results obtained by the model on the training and test sets. In addition, precision, recall, and F1-score are plotted in figures below, respectively, to visually compare model performance for each class.

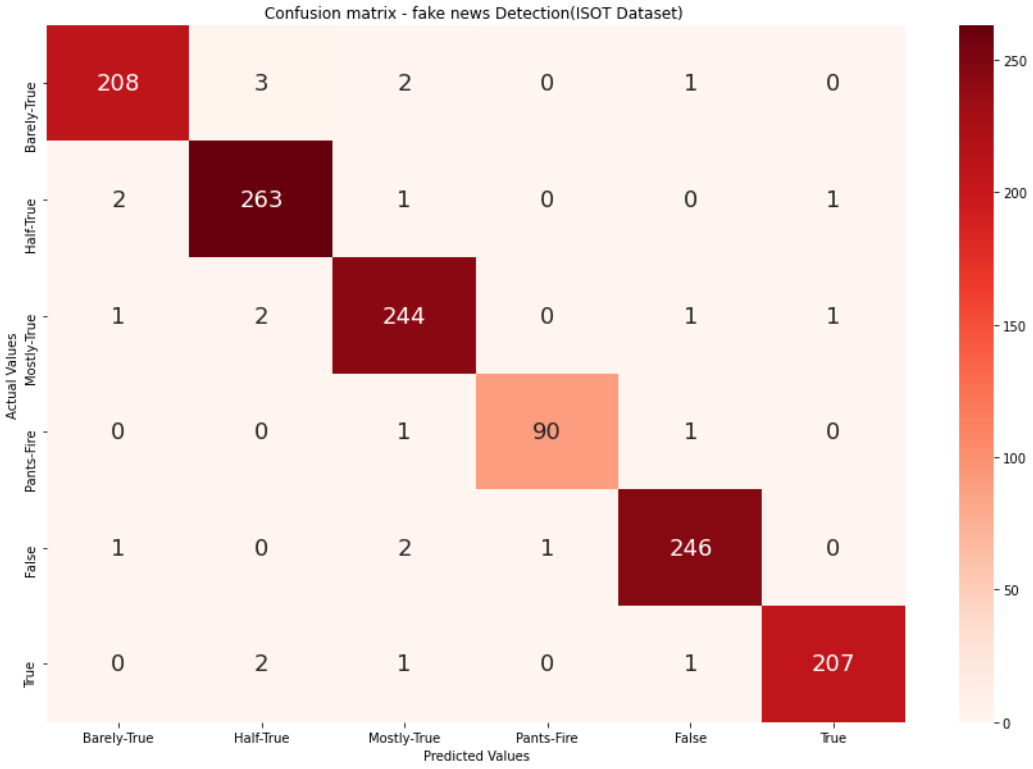

Figure 10. Confusion Matrix of the Final Model for the LIAR Dataset

Table 6. Final Accuracy Obtained in the Training and Testing Stages for the LIAR Dataset

Proposed Model	Test Set (%)	Training Set (%)
Proposed Model	98.17	100

Table 7. Final Model Performance in the Training and Testing Stages for the LIAR Dataset

Class	Precision (Test)	Recall (Test)	F1-Score (Test)	Precision (Training)	Recall (Training)	F1-Score (Training)
Barely True	98.1	97.8	97.9	99.1	99.5	99.3
Half True	98.1	97.8	97.9	98.8	99.3	99.0
Mostly True	98.1	97.8	97.9	98.3	99.0	98.6
Pants-on-Fire	98.1	97.8	97.9	99.5	100	99.7
False	98.1	97.8	97.9	99.2	99.6	99.4
True	98.1	97.8	97.9	99.3	99.8	99.5

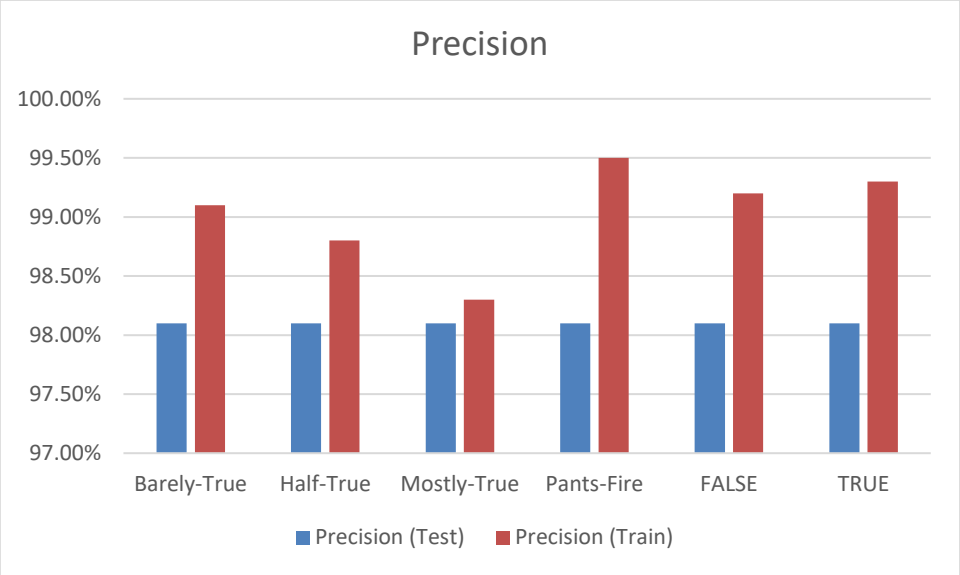

Figure 11. Precision Values Obtained for the Training and Test Data in the LIAR Dataset

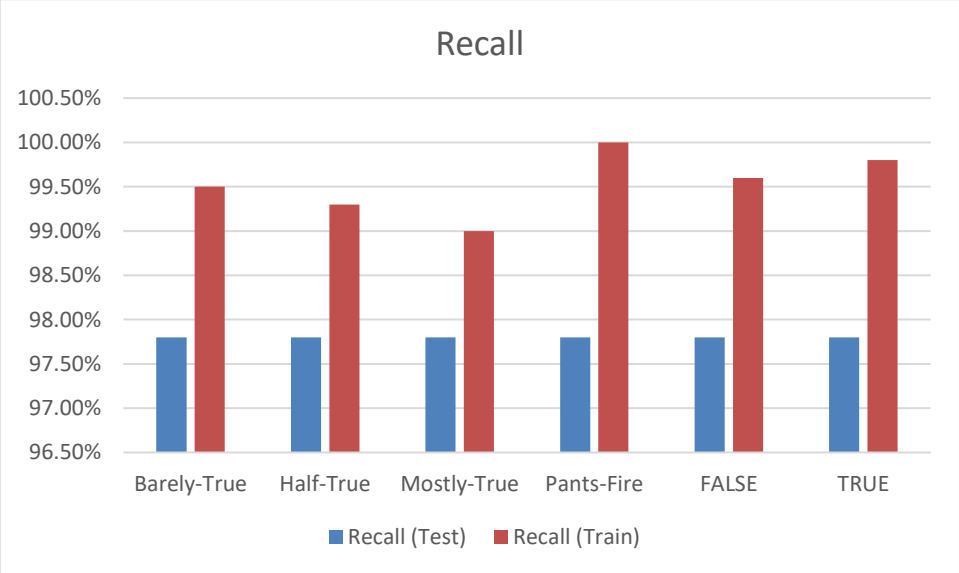

Figure 12. Recall Values Obtained for the Training and Test Data in the LIAR Dataset

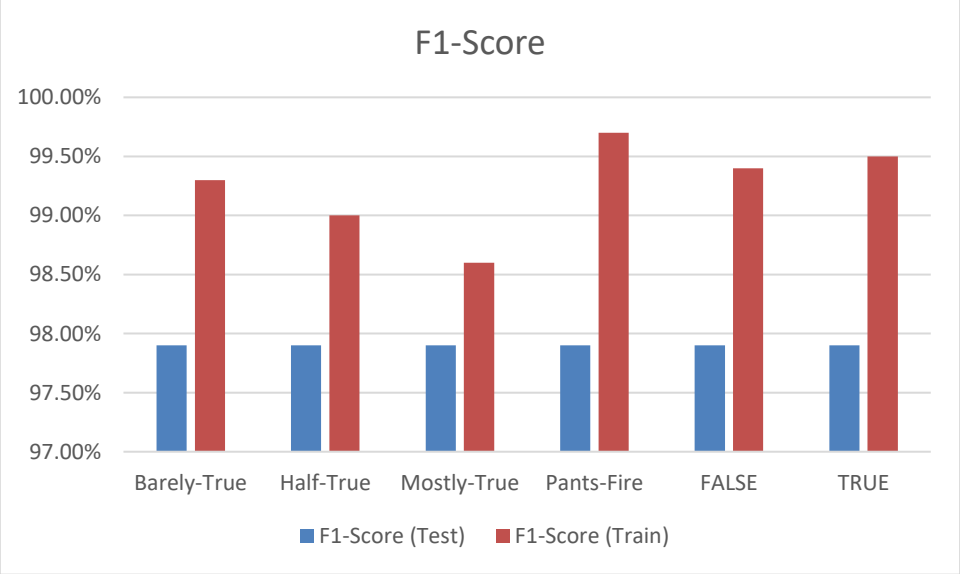

Figure 13. F1-Score Values Obtained for the Training and Test Data in the LIAR Dataset

Experimental Results on the ISOT Dataset

The proposed ensemble framework for fake news detection on social networks, which relies on deep textual embeddings and intelligent parameter optimization, demonstrated very strong performance on the LIAR dataset. This dataset, with its complex textual structure and six different semantic classes, provides a challenging setting for model evaluation. The proposed model achieved high accuracy in the testing stage, indicating its ability to generalize and learn diverse semantic patterns in real data. The results obtained across different evaluation metrics indicate the model's balanced performance and high predictive power across all classes. The model's consistent performance across semantically close classes shows that the proposed framework can effectively distinguish subtle differences among true, half-true, and fake news. Furthermore, the model's high accuracy in low-frequency classes indicates that the designed strategy effectively managed the data imbalance problem and prevented performance degradation in minority classes. This desirable performance is the result of integrating deep textual feature extraction with intelligent model-structure optimization. The ensemble approach reduced the weaknesses of the base classifiers, while the parameter optimization process prevented overfitting. Overall, the obtained results confirm the effectiveness of the proposed framework in automatic fake news detection in complex multiclass textual environments and demonstrate its reliability for practical applications in social networks.

Figure 14 shows the confusion matrix for the ISOT dataset. Table 8 presents the results obtained in the training and test sets. Precision, recall, and F1-score are also presented in Figures 21, 22, and 23 to enable a detailed analysis of model performance.

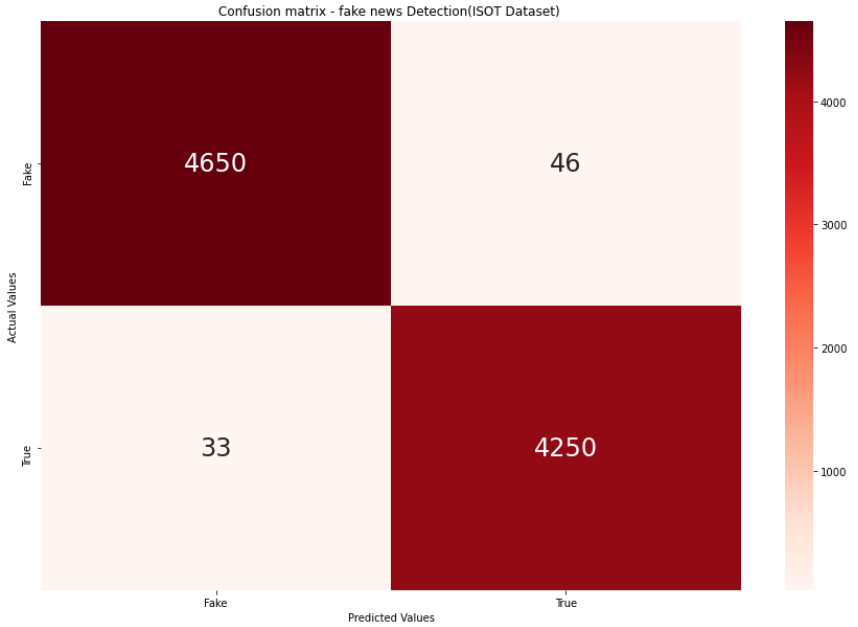

Figure 14. Confusion Matrix of the Final Model for the ISOT Dataset

Table 8. Final Accuracy Obtained in the Training and Testing Stages for the ISOT Dataset

Proposed Model	Test Set (%)	Training Set (%)
Proposed Model	99.24	100

The model achieved an accuracy of 99.24% in predicting the test data, indicating very strong performance in detecting real and fake news.

Table 9. Final Model Performance in the Training and Testing Stages for the ISOT Dataset

Metric	Test Precision	Test Recall	Test F1-Score	Training Precision	Training Recall	Training F1-Score
Fake	99.3	98.9	99.1	100	100	100
Real	99.3	98.9	99.1	100	100	100

Figure 15. Precision Values Obtained for the Training and Test Data in the ISOT Dataset

Figure 16. Recall Values Obtained for the Training and Test Data in the ISOT Dataset

Figure 17. F1-Score Values Obtained for the Training and Test Data in the ISOT Dataset

Experimental Results on the ISOT Dataset

The results obtained from the ISOT dataset show that all evaluation metrics for both fake and real classes are above 99%. This confirms the ability of the proposed model to effectively detect real and fake news while maintaining balance between the two classes. The small number of errors in predicting fake and real samples indicates that the model can establish a clear boundary between true and false information.

The model's precision, recall, and F1-score for both classes are very high and close to each other, indicating stable performance and a balance between false positives and false negatives. Out of 8,979 test samples, only a small number of fake and real news items were misclassified, demonstrating the model's generalization power and stability. This balance across the metrics is especially important for practical applications on social networks.

The success of the model results from combining RoBERTa textual embeddings with an ensemble classifier optimized by the Whale Optimization Algorithm. RoBERTa provides a deep understanding of textual meaning, while WOA enables effective optimization of hyperparameters and classifier weights. The results on the ISOT dataset show that the proposed framework is flexible, accurate, and reliable, with strong generalization capability on real and complex data.

Comparison with Other Methods

This section presents a final comparison between the proposed method and other methods. In recent years, various methods have been proposed for fake news detection using machine learning and deep learning algorithms; however, most of these methods have limitations in terms of accuracy or performance on real-world data.

The results of this study are compared with several previous methods in Table 10.

Table 10. Comparison of the Proposed Method with Existing Methods

Source	Method	LIAR Dataset	ISOT Dataset
(Adedoyin & Mariyappan, 2024)	Capsule Neural Networks	78.2%	98.7%
(Hakak et al., 2021)	Ensemble ML with feature extraction	84%	99%
(Rustam et al., 2024)	Enhanced features via text-to-image with customized models	92%	98.7%
(Alsuwat & Alsuwat, 2025)	Improved multi-modal framework using NLP and Bi-LSTM	95.6%	96.3%
(Merryton & Augasta, 2024)	Attribute-wise Attention model with BiLSTM	60.31%	99%
Proposed Method	Enhanced Ensemble Classifier optimized by WOA with RoBERTa embeddings	98.17%	99.24%

By combining RoBERTa-based deep embeddings with intelligent hyperparameter optimization through the Whale Optimization Algorithm, the proposed method demonstrates superior performance compared with previous methods. The use of RoBERTa enables the model to effectively understand semantic relationships and complex linguistic structures, while the Whale Optimization Algorithm contributes to adaptive optimization of classifier hyperparameters, resulting in faster convergence and reduced classification error. This framework provides high accuracy, stability, and generalizability, and performs particularly well in diverse and noisy social media data environments. The results show that combining powerful feature extraction with intelligent optimization offers an effective pathway for developing high-performance fake news detection systems.

4. Discussion and Conclusion

The findings of the present study demonstrate that the proposed hybrid ensemble framework, integrating RoBERTa-based contextual embeddings with Whale Optimization Algorithm (WOA)-based hyperparameter tuning, achieved exceptionally high performance across both datasets, LIAR and ISOT. The model attained an accuracy of 98.17% on the LIAR dataset and 99.24% on the ISOT dataset, alongside consistently high values of precision, recall, and F1-score. These results indicate that the proposed framework is highly effective in distinguishing between true and false information, even in complex and diverse textual environments. The robustness of the model is particularly evident in its balanced performance across all evaluation metrics, suggesting that it successfully mitigates the common trade-offs between precision and recall. This balanced performance is critical in fake news detection tasks, where both false positives and false negatives can have significant real-world implications.

One of the key factors contributing to the superior performance of the proposed model is the use of RoBERTa embeddings for feature extraction. Unlike traditional feature engineering approaches or static word embeddings, RoBERTa provides deep

contextual representations that capture semantic nuances and long-range dependencies within text. This capability allows the model to effectively differentiate subtle linguistic variations between fake and real news, which are often difficult to detect using surface-level features. Previous studies have similarly reported the effectiveness of transformer-based models in enhancing fake news detection performance, particularly due to their ability to model contextual relationships and complex language structures (Farokhian et al., 2024; Raza et al., 2025). The results of this study further reinforce these findings by demonstrating that contextual embeddings play a crucial role in improving classification accuracy and generalization capability.

In addition to feature representation, the ensemble learning structure significantly contributed to the improved performance of the proposed model. By combining multiple base classifiers, the ensemble approach leveraged the strengths of different algorithms while compensating for their individual weaknesses. This diversity among classifiers enhances the model's ability to generalize across various data distributions and reduces the risk of overfitting. The effectiveness of ensemble learning in fake news detection has been widely documented, with studies indicating that ensemble-based models often outperform individual classifiers in terms of accuracy and robustness (Elyassami et al., 2022; Hakak et al., 2021). The findings of this research align with these studies, demonstrating that the integration of heterogeneous classifiers leads to more reliable and stable predictions.

The incorporation of the Whale Optimization Algorithm for hyperparameter tuning further enhanced the performance of the ensemble model. Traditional hyperparameter tuning methods, such as grid search or manual tuning, are often computationally expensive and may fail to identify optimal configurations in high-dimensional search spaces. In contrast, WOA efficiently explores the search space by balancing exploration and exploitation, enabling the model to converge toward optimal solutions with fewer iterations. The results indicate that WOA not only improved classification accuracy but also contributed to the stability and consistency of the model across different datasets. Similar benefits of metaheuristic optimization techniques have been reported in previous studies, where algorithms inspired by natural processes have been shown to effectively optimize model parameters and improve predictive performance (Choudhury & Acharjee, 2023; Heidari et al., 2019). The successful application of WOA in this study highlights its potential as a powerful tool for optimizing complex machine learning systems.

Another important aspect of the proposed framework is the use of GAN-based data balancing to address class imbalance issues. The results demonstrate that the model maintained high performance even in datasets with initially imbalanced class distributions, indicating that the synthetic data generated by GAN effectively enhanced the representation of minority classes. This improvement is particularly evident in the LIAR dataset, where the model achieved consistent performance across all six classes, including those with fewer samples. Previous research has emphasized the importance of addressing class imbalance in fake news detection, as imbalanced datasets can lead to biased models that favor majority classes (Adedoyin & Mariyappan, 2024; Madani et al., 2024). The findings of this study confirm that GAN-based augmentation is an effective strategy for improving model robustness and ensuring fair representation of all classes.

The performance of the proposed model on the LIAR dataset is particularly noteworthy due to the dataset's complexity and fine-grained classification structure. Unlike binary classification tasks, the LIAR dataset requires the model to distinguish among six levels of truthfulness, which introduces additional challenges in capturing subtle semantic differences. The high accuracy achieved in this study indicates that the proposed framework is capable of handling such complexity, effectively identifying nuanced distinctions between categories such as "half true," "mostly true," and "false." This capability is consistent with findings from prior studies that highlight the importance of deep contextual modeling in multi-class fake news detection tasks (Kishwar & Zafar, 2023; Zhang et al., 2023). The results suggest that the integration of advanced embeddings and ensemble learning is particularly beneficial in addressing fine-grained classification problems.

Similarly, the results obtained on the ISOT dataset demonstrate the model's effectiveness in handling longer and more complex textual data. The near-perfect performance across all evaluation metrics indicates that the model successfully captures the structural and semantic characteristics of news articles at the document level. This finding aligns with previous studies that emphasize the importance of deep learning models in analyzing long-form textual content and identifying deceptive writing patterns (Abualigah et al., 2024; Ahmad et al., 2022). The consistent performance across both short-text and long-text datasets further highlights the versatility and adaptability of the proposed framework.

Comparative analysis with existing methods also underscores the superiority of the proposed approach. As shown in the results, the proposed model outperformed several state-of-the-art techniques, including capsule neural networks, multimodal

frameworks, and feature-based ensemble models. This improvement can be attributed to the synergistic integration of advanced feature extraction, ensemble learning, and intelligent optimization. Previous studies have explored similar hybrid approaches, combining deep learning with machine learning classifiers or multimodal data fusion; however, many of these methods have limitations in terms of scalability, interpretability, or performance consistency (Mahmud et al., 2024; Rustam et al., 2024). The proposed framework addresses these limitations by providing a unified and optimized architecture that achieves high accuracy while maintaining robustness across different datasets.

Furthermore, the balanced performance across evaluation metrics indicates that the model effectively minimizes both false positives and false negatives. This balance is crucial in real-world applications, where incorrect classification of news content can lead to significant consequences, such as the spread of misinformation or the suppression of legitimate information. The ability of the model to maintain high precision and recall simultaneously demonstrates its reliability and practical applicability in social media environments. This finding is consistent with prior research emphasizing the importance of achieving balanced performance in fake news detection systems (Alghamdi et al., 2024; Capuano et al., 2023).

Overall, the results of this study confirm that the integration of RoBERTa embeddings, ensemble learning, GAN-based data balancing, and WOA-based optimization provides a comprehensive and effective solution for fake news detection. The proposed framework not only achieves high accuracy but also demonstrates strong generalization capability, robustness to data imbalance, and adaptability to different textual contexts. These findings contribute to the growing body of literature on hybrid approaches for misinformation detection and highlight the potential of combining multiple advanced techniques to address complex classification tasks (Farhangian et al., 2024; Hashmi et al., 2024).

Despite the promising results, this study has several limitations that should be acknowledged. First, the model was evaluated on two widely used benchmark datasets, which, although diverse, may not fully represent the complexity and variability of real-world social media data. Second, the computational cost associated with transformer-based embeddings and metaheuristic optimization may limit the scalability of the proposed framework in large-scale or real-time applications. Third, the reliance on textual data alone may restrict the model's ability to capture multimodal cues, such as images, videos, or user interactions, which are increasingly relevant in modern misinformation scenarios.

Future research can build upon the findings of this study by exploring several directions. One potential avenue is the integration of multimodal data sources, combining textual, visual, and network-based features to enhance detection performance. Another important direction is the development of more efficient optimization techniques that reduce computational complexity while maintaining high accuracy. Additionally, future studies could investigate the interpretability of ensemble models, providing insights into the decision-making process and improving transparency. Expanding the evaluation to include real-time and streaming data scenarios would also be valuable in assessing the practical applicability of the proposed framework.

From a practical perspective, the findings of this study have important implications for the development of automated systems for misinformation detection in social media platforms. The proposed framework can be integrated into content moderation systems to assist in the identification and filtering of fake news, thereby reducing the spread of misinformation. Policymakers and platform administrators can leverage such models to enhance information quality and protect users from deceptive content. Furthermore, the adaptability and robustness of the model make it suitable for deployment in various domains, including journalism, public health communication, and online education, where accurate information dissemination is critical.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

Acknowledgments

Authors thank all who helped us through this study.

Conflict of Interest

The authors report no conflict of interest.

Funding/Financial Support

According to the authors, this article has no financial support.

References

Abualigah, L., Al-Ajlouni, Y. Y., Daoud, M. S., Altalhi, M., & Migdady, H. (2024). Fake news detection using recurrent neural network based on bidirectional LSTM and GloVe. *Social Network Analysis and Mining*, *14*(1), 40. https://doi.org/10.1007/s13278-024-01198-w

Adedoyin, F., & Mariyappan, B. (2024). *Fake News Detection using Machine Learning Algorithms and Recurrent Neural Networks*.

Ahmad, T., Faisal, M. S., Rizwan, A., Alkanhel, R., Khan, P. W., & Muthanna, A. (2022). Efficient fake news detection mechanism using enhanced deep learning model. *Applied Sciences*, *12*(3), 1743. https://doi.org/10.3390/app12031743

Al Ghamdi, M. A., Bhatti, M. S., Saeed, A., Gillani, Z., & Almotiri, S. H. (2024). A fusion of BERT, machine learning and manual approach for fake news detection. *Multimedia Tools and Applications*, *83*(10), 30095-30112. https://doi.org/10.1007/s11042-023-16669-z

Alghamdi, J., Luo, S., & Lin, Y. (2024). A comprehensive survey on machine learning approaches for fake news detection. *Multimedia Tools and Applications*, *83*(17), 51009-51067. https://doi.org/10.1007/s11042-023-17470-8

Alsuwat, E., & Alsuwat, H. (2025). An improved multi-modal framework for fake news detection using NLP and Bi-LSTM. *Journal of Supercomputing*, *81*(1), 177. https://doi.org/10.1007/s11227-024-06671-z

Capuano, N., Fenza, G., Loia, V., & Nota, F. D. (2023). Content Based Fake News Detection with machine and deep learning: a systematic review. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2023.02.005

Choudhury, D., & Acharjee, T. (2023). A novel approach to fake news detection in social networks using genetic algorithm applying machine learning classifiers. *Multimedia Tools and Applications*, *82*(6), 9029-9045. https://doi.org/10.1007/s11042-022-12788-1

Cui, W., & Shang, M. (2025). MIGCL: Fake news detection with multimodal interaction and graph contrastive learning networks. *Applied Intelligence*, *55*(1), 1-23. https://doi.org/10.1007/s10489-024-05883-3

Del Vicario, M., Quattrociocchi, W., Scala, A., & Zollo, F. (2019). Polarization and fake news: Early warning of potential misinformation targets. *Acm Transactions on the Web*, *13*(2), 1-22. https://doi.org/10.1145/3316809

Elyassami, S., Alseiari, S., AlZaabi, M., Hashem, A., & Aljahoori, N. (2022). Fake news detection using ensemble learning and machine learning algorithms. In *Combating Fake News with Computational Intelligence Techniques* (pp. 149-162). https://doi.org/10.1007/978-3-030-90087-8_7

Farhangian, F., Cruz, R. M. O., & Cavalcanti, G. D. C. (2024). Fake news detection: Taxonomy and comparative study. *Information Fusion*, *103*, 102140. https://doi.org/10.1016/j.inffus.2023.102140

Farokhian, M., Rafe, V., & Veisi, H. (2024). Fake news detection using dual BERT deep neural networks. *Multimedia Tools and Applications*, *83*(15), 43831-43848. https://doi.org/10.1007/s11042-023-17115-w

Fayaz, M., Khan, A., Bilal, M., & Khan, S. U. (2022). Machine learning for fake news classification with optimal feature selection. *Soft Computing*, *26*(16), 7763-7771. https://doi.org/10.1007/s00500-022-06773-x

Goldani, M. H., Momtazi, S., & Safabakhsh, R. (2021). Detecting fake news with capsule neural networks. *Applied Soft Computing*, *101*, 106991. https://doi.org/10.1016/j.asoc.2020.106991

Hakak, S., Alazab, M., Khan, S., Gadekallu, T. R., Maddikunta, P. K. R., & Khan, W. Z. (2021). An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems*, *117*, 47-58. https://doi.org/10.1016/j.future.2020.11.022

Hashmi, E., Yayilgan, S. Y., Yamin, M. M., Ali, S., & Abomhara, M. (2024). Advancing fake news detection: hybrid deep learning with fastText and explainable AI. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3381038

Heidari, A. A., Mirjalili, S., Faris, H., Aljarah, I., Mafarja, M., & Chen, H. (2019). Harris hawks' optimization: Algorithm and applications. *Future Generation Computer Systems*, *97*, 849-872. https://doi.org/10.1016/j.future.2019.02.028

Khanam, Z., Alwasel, B. N., Sirafi, H., & Rashid, M. (2021). *Fake news detection using machine learning approaches* IOP Conference Series: Materials Science and Engineering, https://doi.org/10.1088/1757-899X/1099/1/012040 https://doi.org/10.1088/1757-899X/1099/1/012040

Kishwar, A., & Zafar, A. (2023). Fake news detection on Pakistani news using machine learning and deep learning. *Expert Systems with Applications*, *211*, 118558. https://doi.org/10.1016/j.eswa.2022.118558

Madani, M., Motameni, H., & Roshani, R. (2024). Fake news detection using feature extraction, natural language processing, curriculum learning, and deep learning. *International Journal of Information Technology and Decision Making*, *23*(3), 1063-1098. https://doi.org/10.1142/S0219622023500347

Mahmud, T., Hasan, I., Aziz, M. T., Rahman, T., Hossain, M. S., & Andersson, K. (2024). *Enhanced fake news detection through the fusion of deep learning and repeat vector representations* 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), https://doi.org/10.1109/IDCIoT59759.2024.10467839 https://doi.org/10.1109/IDCIoT59759.2024.10467839

Merryton, A. R., & Augasta, M. G. (2024). An attribute-wise attention model with BiLSTM for an efficient fake news detection. *Multimedia Tools and Applications*, *83*(13), 38109-38126. https://doi.org/10.1007/s11042-023-16824-6

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*, 1-26. https://doi.org/10.1007/s10115-024-02321-1

Rustam, F., 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), 1-26. https://doi.org/10.1007/s10791-024-09490-1

Vereshchaka, A., Cosimini, S., & Dong, W. (2020). Analyzing and distinguishing fake and real news to mitigate the problem of disinformation. *Computational and Mathematical Organization Theory*, 1-15. https://doi.org/10.1007/s10588-020-09307-8

Wang, X., Meng, J., Zhao, D., Meng, X., & Sun, H. (2025). Fake news detection based on multi-modal domain adaptation. *Neural Computing and Applications*, 1-13. https://doi.org/10.1007/s00521-024-10896-7

Zhang, Q., Guo, Z., Zhu, Y., Vijayakumar, P., Castiglione, A., & Gupta, B. B. (2023). A deep learning-based fast fake news detection model for cyber-physical social services. *Pattern Recognition Letters*, *168*, 31-38. https://doi.org/10.1016/j.patrec.2023.02.026