Knowledge and Information Systems
https://doi.org/10.1007/s10115-025-02588-y

RESEARCH

Advancing fake news combating using machine learning: a hybrid model approach

Zahid Aslam[1] · Malik Muhammad Saad Missen[1] · Arslan Abdul Ghaffar[2] · Arif Mehmood[3] · Monica Gracia Villar[4,5,6] · Eduardo Silva Alvarado[4,7,8] · Imran Ashraf[2,4]

Received: 19 June 2024 / Revised: 13 July 2025 / Accepted: 19 August 2025

Abstract
The digital era, while offering unparalleled access to information, has also seen the rapid proliferation of fake news, a phenomenon with the potential to distort public perception and influence sociopolitical events. The need to identify and mitigate the spread of such disinformation is crucial for maintaining the integrity of public discourse. This research introduces a multi-view learning framework that achieves high precision by systematically integrating diverse feature perspectives. Using a diverse dataset of news articles, the approach combines several feature extraction methods, including TF-IDF for individual words (unigrams) and word pairs (bigrams), and counts vectorization to represent text in multiple ways. To capture additional linguistic and semantic information, advanced features, such as readability scores, sentiment scores, and topic distributions generated by latent Dirichlet allocation (LDA), are also extracted. The framework implements a multi-view learning strategy, where separate views focus on basic text, linguistic, and semantic features, feeding into a final ensemble model. Models like logistic regression, random forest, and LightGBM are employed to analyze each view, and a stacked ensemble integrates their outputs. Through rigorous tenfold cross-validation, our proposed multi-view ensemble achieves a state-of-the-art accuracy of 0.9994, outperforming strong baselines, including single-view models and a BERT-based classifier. Robustness testing confirms the model maintains high accuracy even under data perturbations, establishing the value of structured feature separation and intelligent ensemble techniques.

Keywords Information processing · Fake news detection · Natural language processing · Machine learning · Ensemble model · Social media news

1 Introduction

The digital era has facilitated unprecedented access to information but has also given rise to the rapid spread of fake news, posing serious threats to public trust, democratic processes, and societal harmony. Fake news, characterized by deliberate misinformation, has the potential to manipulate public opinion and influence sociopolitical events. The urgent need to identify and mitigate the spread of such disinformation has prompted extensive research into automated

Extended author information available on the last page of the article

detection systems. As such, the identification and neutralization of fake news is a critical task for preserving the integrity of public discourse and the democratic process [1, 2].

In recent times, fake news articles, such as fabricated claims about election outcomes or health misinformation during the COVID-19 pandemic, have shown the immense power of disinformation in shaping public opinion [3]. For instance, during the pandemic, several articles falsely claimed that certain home remedies could cure COVID-19, leading to widespread confusion and even health hazards. Traditional content analysis methods often fail to identify such fake news due to the sophisticated way these articles blend facts with falsehoods, requiring more robust approaches like ensemble learning and advanced feature engineering to accurately detect such misinformation.

Traditional detection methods focused primarily on textual analysis, seeking linguistic cues indicative of deception. These approaches, however, have become increasingly inadequate as fake news perpetrators use sophisticated techniques that blend lies with truths, making it challenging to discern falsehoods based solely on content analysis [4]. Existing research predominantly focuses on linguistic analysis, machine learning, and deep learning techniques. While traditional approaches, such as text-based analysis and feature engineering, have shown promise, they often struggle with generalizability and robustness, particularly when applied to unseen data or across domains. More recent advancements incorporate multimodal data or ensemble learning, but challenges persist in effectively leveraging diverse feature sets and handling noisy or adversarial data.

The research investigation into fake news presents a three-level hierarchical attention network (3HAN), which delves into the intricacies of an article's words, sentences, and headlines to establish a comprehensive representation known as a news vector. This approach underscores the crucial role of structural elements in articles and introduces the novel concept of visualizing attention weights, enhancing the explainability of artificial intelligence (AI)-driven fake news detection mechanisms [5]. Building on the multimodal nature of news [6], introduces a novel method that incorporates external knowledge for a more context-rich analysis. Their approach, the adaptive knowledge-aware fake news detection (AKA-Fake), utilizes a reinforcement learning paradigm to generate a dynamic knowledge subgraph that aligns with the news content. This technique recognizes the limitations of static embeddings and strives to model the complex interplay between multimodal news features and relevant knowledge entities, aiming to improve the reliability of fake news detection.

Recent research has focused on optimizing detection methodologies to overcome these hurdles [7]. The optimized ensemble machine and deep learning (OE-MDL) model enhances traditional models by integrating advanced preprocessing and feature extraction techniques, coupled with sophisticated machine and deep learning classifiers, achieving remarkable performance metrics. Parallel advancements have seen the integration of FastText word embeddings with deep learning techniques to boost the accuracy of classification models [8]. These hybrid models harness the strengths of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, while transformer-based models, like bidirectional encoder representation from transformer (BERT), XLNet, and robustly optimized pretraining approach (RoBERTa), have been fine-tuned to further enhance semantic understanding. The dynamic nature of fake news on microblogging platforms presents another layer of complexity, requiring the development of models that can learn from content and its sequential propagation structure [9]. Approaches employing recursive neural networks and various word embedding schemes, such as global vector for word representation (GloVe) and Google news skip-grams, have provided promising results, revealing that the effectiveness of models can hinge on empirical factors.

In the realm of social media, rumor detection is complicated by the wide dispersion and deep propagation of misinformation. Bidirectional graph convolutional networks (Bi-GCN) have been designed to tackle this challenge, capturing both the propagation and dispersion characteristics of rumors through top-down and bottom-up analyses within a graph model [10]. Another innovative approach involves exploiting the "wisdom of crowds" on microblogs [11]. By harnessing conflicting viewpoints that emerge alongside fake news, a credibility propagation network can be constructed, facilitating more effective news verification. The adversarial active learning-based heterogeneous graph neural network (AA-HGNN) leverages a heterogeneous information network (HIN) to encapsulate various entities and their relationships on social media, enhancing detection performance, especially when labeled data is scarce [12]. This approach highlights the potential of adversarial active learning in improving fake news classification with less labeled data.

Recognizing the inadequacy of content analysis in the face of fake news spread through social media, the authors [13], propose a deep natural language processing (NLP) model that operates across a multi-layered architecture. It processes data through acquisition, retrieval, NLP-driven feature extraction, and deep learning classification, factoring in the credibility of publishers and users. Their approach notably outperforms existing methods in accuracy. The research [14], suggests that semantic analysis alone falls short, especially with short content. They introduce a multitask learning model that combines fake news detection with topic classification, hypothesizing that certain topics and authors are more prone to fake news. The integration aims to enhance detection performance. Addressing the challenge of fake news identification online, the researchers investigate a bidirectional LSTM model, applying it to the FNC-1 dataset [15]. Their focus is on automating fake news detection on digital platforms, aiming to leverage the LSTM's feature extraction strength for high accuracy.

The study [16] tackles the challenge of multimodal fake news on microblogging networks by proposing an end-to-end model named BERT-based domain adaptation neural network (BDANN). It uses BERT for text and VGG-19 for image feature extraction, with a domain classifier to unify features into one space for effective fake news detection. The research [17] addresses the challenges posed by the abundance of misinformation on social media by evaluating and comparing state-of-the-art methods using big data technology and machine learning (ML) on the FNC-1 dataset. They present a stacked ensemble model that shows significant improvement in classification performance. The paper presents a stance prediction technique as a means to gauge news article authenticity [18]. The proposed architecture employs a bidirectional LSTM and an autoencoder to automatically classify the stance of news articles, demonstrating high accuracy in stance prediction as a step toward assessing news credibility.

The study [19] highlights the prevalence of fake news in political discussions, noting the inadequacy of existing detection methods that often rely solely on text patterns. The proposed solution involves deep neural networks that incorporate the credibility patterns of politicians, which have proven to enhance the accuracy of fake news classification in this domain. Recognizing the complexities of fake news on social media, the authors propose a deep neural network ensemble architecture for social and textual context-aware fake news detection (DANES), an ensemble architecture that synergizes textual content analysis with social context [20]. This approach significantly improves detection capabilities by harnessing insights from both social interactions and content analysis. The study [21] critiques the effectiveness of NLP alone in detecting fake news, proposing a model that incorporates additional live data features such as source and authorship. By blending these elements with LSTM and feedforward neural networks, the model aims to mirror the fact-checking process and enhance authenticity assessments. Addressing rumor detection on social media, the research

introduces a CNN-based model that excels in the early detection of misinformation [22]. Through optimized hyperparameter settings, the model demonstrates superior performance over traditional methods, achieving a balance of recall and precision in identifying rumors.

Despite these advances, a key research question remains: can a structured separation of feature types, modeled independently before being intelligently combined, lead to a more robust and accurate classifier? Many models either rely on a single view of the data (e.g., only textual features) or combine all features into a single "flat" representation, which may not leverage the unique strengths of each feature type.

This study addresses this question by proposing a hybrid multi-view learning framework. Here, the term hybrid refers to the two levels of combination in our architecture: (1) the fusion of diverse, engineered feature sets (textual, linguistic, semantic), and (2) the ensemble of distinct machine learning models in a stacked configuration. Our central hypothesis is that by isolating different feature views, we can train specialized base models that capture unique patterns, and a subsequent meta-learner can then integrate these expert predictions more effectively than a single flat model. To validate this, we conduct a rigorous comparative evaluation against strong baselines. The key contributions of this paper are therefore:

- *A multi-view learning architecture* We propose a framework that systematically decomposes news articles into three distinct feature views textual, linguistic, and semantic to capture a more holistic representation of the content.
- *A robust stacked ensemble model* We implement a stacked ensemble that intelligently combines predictions from specialized models trained on each view, demonstrating that this hierarchical architecture is more effective than standard flat ensemble methods.
- *Rigorous comparative evaluation* We provide a comprehensive evaluation using tenfold cross-validation, showing that our proposed architecture achieves state-of-the-art results by outperforming single-view models, a feature-rich flat model, and a BERT-based classifier.

These contributions highlight the innovative combination of multi-view learning, advanced feature engineering, and ensemble techniques in addressing the complex challenge of fake news detection. This research not only enhances current methodologies but also establishes a new standard for precision and reliability in this domain.

Related work is discussed in Sect. 2, followed by a detailed explanation of the proposed approach, dataset, machine learning models, and evaluation metrics in Sect. 3. The results are presented in Sect. 4, and the study concludes in Sect. 5.

2 Literature review

The dissemination of fake news has emerged as a significant challenge in the digital age, prompting researchers to develop sophisticated detection methods. This literature review examines recent advancements in fake news detection, focusing on diverse methodologies, the application of novel algorithms, and their effectiveness across various datasets.

2.1 Knowledge integration and graph-based models

In recent research, Ref. [6] proposes a novel approach called AKA-Fake, which integrates external knowledge to enhance the detection of multimodal fake news. This model constructs a knowledge subgraph under a reinforcement learning paradigm and employs a heterogeneous graph learning module to capture cross-modality correlations. The methodology was evalu-

ated on three popular datasets, achieving a notable accuracy of 91.9%. Similarly, the study [10] introduces a bidirectional graph model (Bi-GCN) to address the challenges of rumor detection on social media. By leveraging graph convolutional networks for both rumor propagation and diffusion, the study reports an encouraging accuracy of 96.1% on the Weibo dataset. These graph-based approaches underscore the importance of leveraging knowledge structures and network relationships to enhance fake news detection capabilities.

2.2 Hierarchical and contextual models

The research [5], presents the 3HAN designed to effectively discern fake news by analyzing words, sentences, and headlines. This deep learning approach emphasizes the importance of the hierarchical structure in news articles for precise fake news classification, reporting an impressive accuracy of 96.77%. Another study [18], introduces a deep learning strategy for stance prediction, which can indicate the authenticity of news articles. Utilizing a combination of bidirectional LSTM and autoencoder, the model effectively classifies the stance of news articles with a high accuracy of 94%. These works highlight the value of hierarchical and contextual understanding in improving fake news detection.

2.3 Optimized ensemble learning and hybrid models

The study [7], introduces an optimized ensemble approach combining machine learning and deep learning techniques. The OE-MDL framework utilizes advanced classifiers in both machine learning (OML) and deep learning (ODL) phases, significantly enhancing detection capabilities with an accuracy of 99.87%, precision of 99.88%, recall of 95.87%, and an F1-score of 99.96%. Research [23], proposes a hybrid framework combining Word2Vec embeddings and LSTM layers, excelling across diverse datasets and outperforming state-of-the-art techniques like BERT. This study demonstrates the power of blending feature extraction techniques and neural network architectures to improve classification performance.

2.4 Textual and linguistic feature engineering

The study [8], incorporates FastText word embeddings with hybrid models of CNNs and LSTMs, alongside transformer-based models like BERT, XLNet, and RoBERTa. This methodology, refined through hyperparameter optimization, excels across several datasets, achieving accuracy and F1-scores of up to 99%. Another research [12], proposes the AA-HGNN model, which utilizes adversarial active learning and a hierarchical attention mechanism within a heterogeneous information network (HIN) to detect fake news. This approach demonstrated an accuracy of 61.55%, illustrating its potential in environments with scarce labeled data.

2.5 Deep NLP models and sequential analysis

The study [13], details a deep NLP model that processes information across four layers, from publisher to cloud. Utilizing datasets like Buzzface, FakeNewsNet, and Twitter, the model achieves an accuracy of 99.72% and an F1-score of 98.33%, significantly outperforming other methods. Research [21], emphasizes the limitations of NLP alone in identifying fake news, proposing a system that includes secondary features akin to fact-checking. The LSTM model

utilized in this approach shows an improved accuracy of 91.32%, highlighting the benefits of incorporating broader data points beyond textual analysis. These studies underscore the growing sophistication of NLP techniques for fake news detection.

Recent research [24], demonstrated the successful integration of NLP and deep learning, such as BERT and BiGRU, for cybersecurity threat detection, achieving high accuracy in spam and phishing detection. Their work reinforces the utility of semantic-rich models in detecting subtle patterns in unstructured text. In addition, [25] discusses how the trust level of people is affected due to fake news.

2.6 Propagation and multimodal techniques

Users rely heavily on social media to consume and share news, facilitating the mass dissemination of genuine and fake stories. The research [17], explores the use of big data technologies and machine learning in detecting fake news, employing a stacked ensemble model on the FNC-1 dataset. By incorporating techniques like N-grams, hashing TF-IDF, and count vectorizer, the model significantly outperforms traditional methods, achieving an F1-score of 92.45%. Another study [16], presents a novel model, BDANN, designed to tackle the challenges of detecting multimodal fake news on microblogging platforms. The model utilizes BERT for text feature extraction and VGG-19 for image features, which are then combined in a domain classifier to ensure consistency across different event contexts. Tested on Twitter and Weibo, BDANN achieves a promising accuracy of 85.10%, demonstrating its efficacy in a multimodal setting.

2.7 Dataset-specific studies

The study [26], used genetic algorithms combined with classifiers like support vector machines (SVM), RF, and LR, achieving high accuracy across datasets, with SVM reaching 97% on the Fake Job Posting dataset. Another research [27], developed the first Pakistani fake news detection dataset, demonstrating that LSTM with GloVe embeddings outperformed others, achieving an F1-score of 94%. Similarly, [28] focused on Twitter, employing a stacked ensemble model with N-grams and TF-IDF, yielding an F1-score of 92.45%, significantly surpassing baseline models. These dataset-specific studies highlight the importance of contextual adaptation in fake news detection.

2.8 Challenges in generalizability

Despite notable advancements, challenges persist in achieving generalizability and robustness. Research [14], introduces the FDML model that integrates fake news detection with news topic classification to enhance performance, particularly on short content. This multitask learning approach, emphasizing the influence of topics and author intentions on fake news prevalence, achieved an accuracy of 70.6%. Similarly, [29] addresses the challenge of detecting fake news on newly emerged events by focusing on transferable traits across events. The EANN model, tested on Weibo and Twitter, achieves an accuracy of 82.70%, proving its capability to handle multimodal fake news content effectively. These findings emphasize the need for robust models capable of adapting to varying datasets and contexts.

2.9 Contributions and research gaps

The surveyed literature underscores a dynamic evolution in the methodologies employed for detecting fake news. Existing approaches have explored linguistic analysis, machine learning, deep learning, and multimodal techniques, but challenges remain in generalizability and handling noisy data. This research contributes to the field by introducing an ensemble-based framework that combines logistic regression, LightGBM, and random forest models. By employing multi-view learning and advanced feature engineering techniques, the proposed approach addresses the limitations of prior work, achieving state-of-the-art accuracy while maintaining robustness across diverse datasets.

In Table 1, the literature review of previous research includes dataset details, proposed methodology, and results generated by that methodology.

3 Proposed methodology

Moving from the extensive insights gained from the literature review, we now introduce the proposed methodology, which aims to address some of the gaps identified in previous studies. The proposed approach uses advanced machine learning techniques and a novel architectural framework to enhance the detection and classification of fake news across multiple platforms.

This section provides an end-to-end overview of the proposed solution, encompassing data acquisition, preprocessing, feature engineering, multi-view learning, and ensemble integration.

3.1 Research architecture

The proposed architecture enhances fake news detection by integrating systematic stages, including data acquisition, preprocessing, feature extraction, multi-view learning, model training, and ensemble prediction. The pipeline begins with acquiring a comprehensive dataset from Kaggle [36], which includes labeled real and fake news articles, and the LIAR dataset [37], which contains short political statements. These datasets provide diversity in both content length and context, supporting robust model evaluation.

The data undergoes preprocessing to remove noise and standardize text. A diverse set of features is then extracted across three dimensions: textual (TF-IDF, count vectors), linguistic (readability, sentiment, parts of speech (POS), named entity recognition (NER)), and semantic (LDA topics, Doc2Vec embeddings, similarity scores). Each feature type is modeled independently using LR, RF, and LightGBM within a multi-view learning framework.

A final stacked ensemble with weighted voting aggregates predictions from the individual models, leveraging their complementary strengths. This end-to-end architecture enables precise and generalizable fake news classification, which is further validated through cross-dataset testing and robustness evaluations. An overview of this complete architecture is illustrated in Fig. 1, which visually summarizes all key components of the proposed pipeline.

An advanced ensemble learning strategy is employed to integrate the strengths of the multi-view models through a stacked ensemble approach. A weighted voting mechanism is used to combine the predictions of the individual models within each view, where the weights are assigned based on the precision of each model in the validation data. This blending strategy maximizes predictive accuracy by leveraging the unique capabilities of each view and

Table 1 Overview of recent studies in fake news detection using deep learning techniques

Refs.	Dataset	Proposed approach	Accuracy
[5]	PolitiFact, Forbes	Three-level hierarchical attention network (3HAN)	96.77%
[6]	PolitiFact, GossipCop, Pheme	AKA-Fake model with reinforcement learning	91.9%
[7]	LIAR	Optimized ensemble of machine and deep learning techniques	99.87%
[8]	WELFake, FakeNewsNet, FakeNewsPrediction	Hybrid model of CNNs and LSTMs with FastText embeddings	F1-scores: 0.99
[9]	Twitter16	Recursive neural networks with word embeddings	67%
[10]	Weibo	Bidirectional graph convolutional networks (Bi-GCN)	96.1%
[12]	PolitiFact, BuzzFeed	AA-HGNN with adversarial active learning	61.55%
[13]	Buzzface, FakeNewsNet, Twitter	AI-assisted deep NLP model	99.72%
[14]	LIAR	FDML model integrating fake news detection with topic classification	70.6%
[15]	FNC-1	Bidirectional LSTM concatenated model	85.3%
[16]	Twitter, Weibo	BERT and VGG-19-based multimodal feature extraction	85.10%
[17]	FNC-1	Stacked ensemble model using big data technology and ML	92.45% (F1-score)
[18]	FBFans, CreateDebate	Bidirectional LSTM and autoencoder for stance prediction	94%
[19]	LIAR	Deep neural networks using credibility patterns of politicians	48.50%
[20]	BuzzFace, Twitter15, Twitter16	Ensemble architecture for integrating social and textual contexts	78.64%
[21]	George McIntire	LSTM and FF neural networks with secondary features	91.32%
[22]	PHEME	CNN model for rumor detection on social media	91.01%
[29]	Weibo, Twitter	EANN for deriving event-invariant features	82.70%
[30]	Kaggle fake news dataset	Blend of CNN and RNN with GloVe embeddings	97.21% (Precision)

Table 1 continued

Refs.	Dataset	Proposed approach	Accuracy
[31]	Kaggle fake news dataset	Bidirectional LSTM for detecting fake news	98.75%
[32]	ISOT Fake News Dataset, Kaggle Fake news dataset	Ensemble of machine learning algorithms	99%
[33]	Fake News Challenge (FNC-1)	Neural network architecture for stance detection	94.21%
[34]	LIAR	DSSM and improved RNNs for fake news detection	99%
[35]	Tweet Dataset	Hybrid of CNN and LSTM models for Twitter fake news detection	82%
[26]	LIAR, Fake Job Posting, Fake News	Genetic algorithm with SVM, RF, LR, and NB classifiers	97% (Fake Job Posting), 96% (Fake News)
[27]	Custom Pakistani Fake News Dataset	LSTM with GloVe embeddings for fake news detection	94% (F1-score)
[28]	FNC-1	Stacked ensemble model with N-grams and TF-IDF	92.45% (F1-score)
[23]	Multiple datasets (various domains)	Hybrid framework using Word2Vec embeddings and LSTM layers	Outperformed state-of-the-art methods
Proposed	Kaggle, LIAR	Multi-View Stacked Ensemble (LR, RF, LGBM)	99.86%

model, while also minimizing individual weaknesses, thus enhancing overall classification performance.

To ensure the robustness and adaptability of the model, cross-dataset validation is conducted using the LIAR dataset, assessing the model's ability to generalize between different data sources. Furthermore, robustness testing is applied to evaluate the resilience of the model against adversarial perturbations such as word deletion, word swapping, and word repetition. Sensitivity and stability analyses are also conducted to verify the model's consistency under varying data conditions.

The performance of the model is rigorously evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC–AUC. Cross-validation techniques are employed to confirm the reliability of the models, while robustness scores from the LIAR dataset provide additional validation. This comprehensive methodological framework not only enhances fake news detection capabilities but also ensures that the models are resilient and adaptable across diverse contexts and datasets. Figure 1 illustrates the research architecture of the proposed approach for fake news detection.

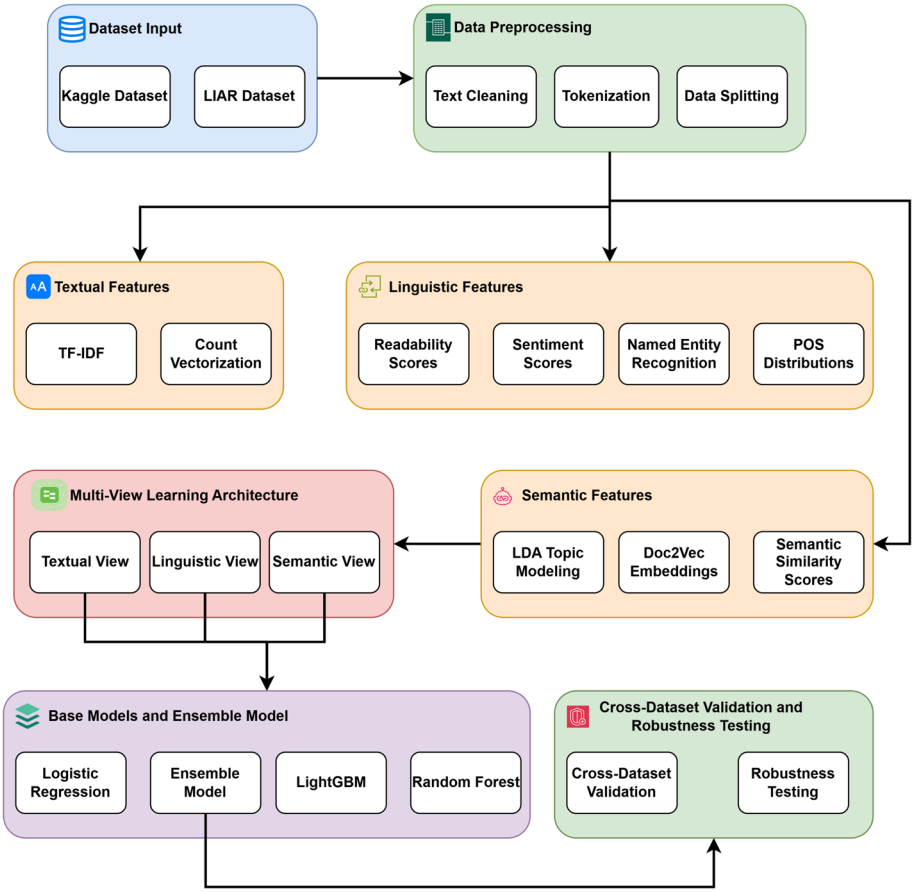

Fig. 1 Architecture of the proposed approach for fake news detection

3.2 Data acquisition

This study relies on two datasets to support the comprehensive detection of fake news: a primary dataset sourced from Kaggle and the LIAR dataset, which is used for cross-dataset validation and robustness testing.

3.2.1 Primary fake news dataset

The primary dataset, obtained from Kaggle [36], provides a detailed collection of news articles categorized into real and fake news. It contains a total of 44,898 articles, with 21,417 categorized as real news and 23,481 categorized as fake news. The real news subset is further divided into two subjects: World News, comprising 10,145 articles, and Politics News, comprising 11,272 articles. On the other hand, the fake news subset includes six categories, namely Government News (1570 articles), Middle-East News (778 articles), US News (783 articles), Left News (4459 articles), Politics (6841 articles), and General News (9050 articles). This diversity allows for a thorough analysis of the types of content and subject matter biases

Table 2 Dataset distribution for fake news detection

News type	Number of articles	Subjects	
		Type	Article size
Real news	21,417	World news	10,145
		Politics news	11,272
Fake news	23,481	Government news	1570
		Middle-east news	778
		US news	783
		Left news	4459
		Politics	6841
		General news	9050

Table 3 LIAR dataset summary

Metric	Value	Description
Total samples	12,791	All labeled statements
Unique speakers	10,240	Distinct individuals in the dataset
Unique subjects	10,240	Topics covered by statements
Average statement length	43.76 words	Mean length of statements

present within the dataset. Metadata, such as publication dates and subject categories, are also included, which adds an additional layer of context for identifying patterns in real versus fake news.

This dataset provides not only a comprehensive range of articles but also a rich variety of linguistic and semantic patterns that can inform the detection of fake news. By analyzing both real and fake news content, the models can identify subtle cues and linguistic discrepancies, enabling accurate classification across a wide variety of subjects. Table 2 shows detailed insights into the dataset.

3.2.2 LIAR dataset

In addition to the Kaggle dataset, the LIAR dataset [37] is employed to evaluate the generalization and robustness of the models across different types of data. The LIAR dataset is a well-structured collection of political statements, each labeled as true, false, or partially true. Table 3 shows that it contains 12,791 statements with unique attributes such as speaker, subject, and context. The average length of statements is 43.76 words, providing a contrasting data type compared to the full-length articles in the primary dataset. The dataset is divided into training, validation, and test sets, containing 10,240, 1284, and 1267 samples, respectively. The training set has an average statement length of 43.92 words, while the validation and test sets have similar averages of 43.14 and 43.12 words, respectively.

The LIAR dataset brings additional diversity to the study by providing short, statement-based data, which contrasts with the longer articles in the Kaggle dataset. This combination ensures that the models are not only trained on diverse subjects but are also evaluated for their robustness and adaptability across different types of text data. The inclusion of the

LIAR dataset allows for cross-dataset validation and the testing of model stability, further strengthening the research methodology.

Together, these datasets provide a solid foundation for training and evaluating the proposed fake news detection system, enabling the study to address both the classification of longer news articles and shorter, statement-based content.

3.3 Data preprocessing

The preprocessing phase is critical for refining the raw data collected from the Kaggle and LIAR datasets, ensuring it is in the best possible format for feature engineering and model training. This step involves multiple stages to clean, standardize, and prepare the data, addressing issues like noise, inconsistencies, and missing values.

3.3.1 Text cleaning and normalization

The first step in preprocessing is text cleaning, where unnecessary characters, such as punctuation marks, special symbols, and extra whitespace, are removed to eliminate noise. Following this, the text is normalized by converting all characters to lowercase, ensuring consistency across the datasets. These steps are particularly important for datasets like Kaggle's, where news articles can vary greatly in format and structure.

3.3.2 Tokenization and stop word removal

After cleaning, the text is tokenized, breaking it into individual words or tokens. Tokenization enables the models to analyze the data at the word level. Additionally, common stop words, such as "and," "the," and "of," are removed to focus on meaningful content. Depending on the dataset, stemming and lemmatization are applied to reduce words to their base forms. For instance, variations like "running," "ran," and "runner" are standardized to the root form "run." This step ensures that word variations do not inflate the feature space unnecessarily.

3.3.3 Data splitting

Once the text is cleaned and tokenized, the data is divided into training, validation, and test sets. For the Kaggle dataset, approximately 70% of the data is used for training, 15% for validation, and 15% for testing. This division ensures that the model is trained on a substantial portion of the data while reserving separate subsets for tuning hyperparameters and evaluating performance. The LIAR dataset uses its predefined splits: 10,240 samples for training, 1284 for validation, and 1267 for testing. These splits maintain consistency in class distribution, ensuring the data subsets are representative of the overall dataset.

3.3.4 Addressing class imbalance

An essential aspect of preprocessing is handling any class imbalances in the datasets. For example, the Kaggle dataset contains slightly more fake news articles than real news articles. Techniques, like stratified sampling, are employed to ensure that the class proportions in the training, validation, and test sets match the overall dataset distribution. This step is crucial for preventing the model from being biased toward the majority class during training.

Through these preprocessing steps, the raw data is transformed into a clean and structured format, ready for feature engineering. By ensuring high-quality data, this phase lays the groundwork for effective and reliable model training and evaluation.

3.4 Feature engineering

Feature engineering plays a crucial role in transforming raw, preprocessed data into meaningful representations that can enhance the performance of machine learning models. In this study, a combination of textual, linguistic, and semantic features has been extracted to capture a comprehensive range of characteristics from both real and fake news articles. These features are designed to represent the underlying patterns and nuances that differentiate legitimate information from disinformation.

3.4.1 Textual features

The textual features are primarily based on basic text representations that quantify word usage patterns and text characteristics. Two key techniques are employed for this purpose: TF-IDF and count vectorization.

The TF-IDF approach is applied to both unigrams and bigrams, capturing the significance of individual words and word pairs based on their frequency within and across documents. Count vectorization, on the other hand, provides a straightforward frequency-based representation of words, offering an alternative perspective on the textual content. Together, these techniques serve as the foundation for understanding word distributions and relationships within the dataset.

An important aspect of textual analysis is the study of text lengths, as the length of news articles often varies significantly between real and fake news. The provided visualization (Fig. 2) illustrates the distribution of text lengths across real and fake news articles. The histogram on the left highlights the overall frequency distribution of text lengths, showing that fake news articles tend to be slightly longer on average than real news articles. The box plot on the right provides additional insights into the spread and variability of text lengths. It reveals that fake news articles exhibit a wider range of lengths, including several outliers, while real news articles are generally more concise. These differences can be leveraged as features to enhance classification performance.

3.4.2 Linguistic features

Beyond basic text characteristics, linguistic features are extracted to capture stylistic and grammatical patterns in the text. These include readability scores, sentiment analysis, NER, and POS distributions. Readability metrics, such as the Flesch reading ease score, provide insights into the complexity of the language used, while sentiment analysis evaluates the tone of the articles, identifying whether they are positive, negative, or neutral. NER features count the frequency and types of entities, such as people, organizations, and locations, mentioned in the articles. Finally, POS tag distributions highlight the grammatical structure of the text, helping to differentiate between the linguistic styles of real and fake news.

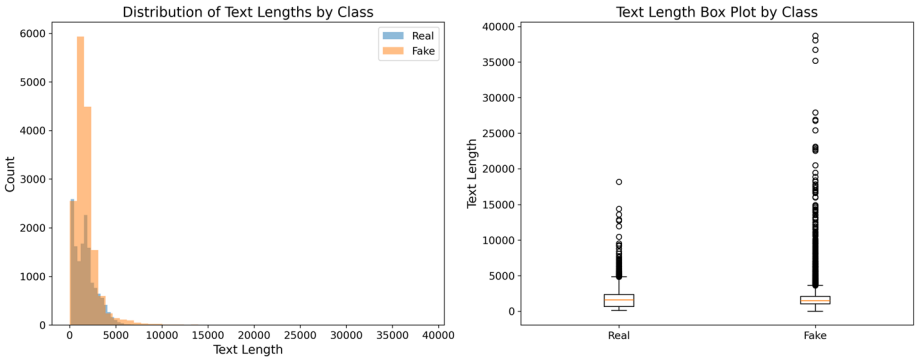

Fig. 2 Text length analysis: (left) distribution of text lengths by class. (Right) box plot comparing text lengths for real and fake news

3.4.3 Semantic features

Semantic features delve deeper into the contextual and thematic aspects of the text. LDA is used to extract topic distributions, revealing the main themes present in the dataset. Additionally, Doc2Vec embeddings are generated to create dense vector representations of the articles, capturing their semantic meanings. These embeddings are complemented by semantic similarity scores, which measure the contextual relationships between documents. Together, these features provide a high-level understanding of the dataset's semantic structure, enabling models to detect subtle differences in meaning and context.

3.4.4 Rationale for feature selection

The feature extraction techniques selected in this study are based on their proven effectiveness in text classification tasks and their interpretability for fake news detection. Traditional methods, like TF-IDF and count vectorizer, were chosen due to their simplicity and strong baseline performance in capturing word frequency-based importance. These methods have shown robust performance across various news classification tasks and provide sparse, high-dimensional feature spaces well suited for traditional machine learning models.

Linguistic features, such as readability and sentiment scores, were incorporated to capture writing complexity and emotional tone, which are often indicative of deceptive content. These features offer interpretable signals and have been validated in prior fake news research.

Semantic features, like topic modeling (LDA) and dense embeddings (Doc2Vec), were used to capture latent themes and contextual semantics, which are crucial for distinguishing nuanced misinformation. Alternatives, such as word2vec or transformer-based embeddings (e.g., BERT), were considered, but Doc2Vec was selected due to its efficiency and better performance with longer documents in our preliminary trials.

The selected combination of features balances interpretability, computational efficiency, and predictive power, enabling the model to detect subtle differences between real and fake content.

Table 4 Feature dimensions for text, linguistic, and semantic features

Feature type	Feature technique	Dimensions
Textual features	TF-IDF (unigram and bigram)	5000
Textual features	Count vectorization	5000
Linguistic features	Readability, sentiment, NER, POS tags	13
Semantic features	LDA topics	20
Semantic features	Doc2Vec embeddings	100
Semantic features	Semantic similarity scores	3

3.4.5 Feature dimensions summary

The extracted features result in a diverse and high-dimensional feature space, which enhances the ability of the models to differentiate between real and fake news. Table 4 provides a summary of the dimensions for each feature type.

By combining these textual, linguistic, and semantic features, the feature engineering framework captures the intricate patterns and relationships in the dataset. This comprehensive representation is a key component of the multi-view learning architecture, enabling the models to effectively classify real and fake news with high accuracy.

This approach is termed hybrid as it combines multi-view feature fusion by separating and processing textual, linguistic, and semantic features independently with ensemble learning, where the outputs of different machine learning models are aggregated through a weighted voting mechanism. This hybridization ensures both feature-level diversity and model-level robustness.

3.4.6 Additional feature extraction details

To improve transparency and reproducibility, the following details are provided for the linguistic and semantic features:

Readability score Readability was computed using the Flesch reading ease formula:

$$\text{Readability score} = 206.835 - 1.015\left(\frac{\text{Total words}}{\text{Total sentences}}\right) - 84.6\left(\frac{\text{Total syllables}}{\text{Total words}}\right)$$

Higher scores indicate easier-to-read content.

Sentiment analysis Sentiment polarity and subjectivity scores were extracted using the VADER sentiment analysis tool, which is suitable for social media and short-text contexts.

LDA topic modeling LDA was applied using the Gensim library with the following parameters: number of topics = 10, alpha = "symmetric," beta = "auto," and iterations = 100. Preprocessing steps included stopword removal, lemmatization, and bigram detection before applying LDA.

3.5 Multi-view learning architecture

To fully leverage the diverse feature sets extracted during feature engineering, a multi-view learning architecture is employed. This approach divides the features into three distinct views

textual, linguistic, and semantic and processes each view with a specialized base model before integrating their outputs into a final stacked ensemble. By isolating and analyzing the unique contributions of each feature type, the multi-view architecture is designed to capture complementary insights that significantly enhance the detection of fake news.

3.5.1 Textual view

The textual view focuses on word-level representations of the articles, primarily utilizing TF-IDF vectors. These features capture the frequency and importance of words and word pairs. A logistic regression (LR) model is used to process this view, given its effectiveness and efficiency in handling high-dimensional, sparse feature spaces.

3.5.2 Linguistic view

This view delves into the stylistic and grammatical aspects of the text. Features, such as readability scores, sentiment analysis, and POS tag distributions, are processed here. A random forest (RF) classifier, known for its ability to handle heterogeneous and nonlinear features, is employed to analyze this view.

3.5.3 Semantic view

The semantic view captures the contextual and thematic content of the text. LDA topic distributions and Doc2Vec embeddings are used to represent the high-level meaning of the articles. A LightGBM model, a powerful gradient-boosting framework, is applied to process this view and model its complex relationships.

3.6 Stacked ensemble learning

Instead of relying on a simple voting mechanism, our framework employs a more sophisticated stacked ensemble (or stacking) approach to combine the predictions from the three views. Stacking involves a two-level learning process:

1. *Level 0 (base models)* The three specialized models (LR for textual, RF for linguistic, and LightGBM for semantic) are trained on the full training dataset for each fold of our cross-validation.
2. *Level 1 (meta-learner)* The predictions (i.e., output probabilities) from these three base models are then used as input features to train a final "meta-learner." In our framework, we use a logistic regression classifier as the meta-learner. This model learns the optimal way to combine the predictions from the base models to produce the final, highly accurate classification.

This two-level architecture allows the framework to not only learn from the initial features but also to learn how to best weigh the "opinions" of the specialized base models, leading to a more robust and nuanced final decision.

3.7 Model evaluation and validation

Proper evaluation and validation are crucial to ensure the effectiveness and reliability of our proposed model. This study focuses on a comprehensive set of metrics and a robust validation protocol.

3.7.1 Evaluation metrics

The model performance was evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC–AUC), which measures a model's ability to discriminate between classes.

3.7.2 Cross-validation procedure

To ensure the statistical validity of our results and to generate robust models that generalize well to unseen data, we employed a tenfold stratified cross-validation protocol for all experiments. In this procedure, the dataset is divided into 10 equal-sized subsets, or "folds." Stratification ensures that each fold maintains the same proportion of real and fake news articles as the original dataset. The final performance metric for any given model is reported as the mean and standard deviation of the scores obtained across all 10 folds. To prevent any data leakage, feature engineering components, such as the TF-IDF vectorizer, were fitted only on the training data within each fold.

3.7.3 Robustness testing

The robustness of the proposed ensemble model was evaluated to determine its reliability under adversarial conditions. To rigorously evaluate stability, we applied three distinct adversarial perturbation techniques to the test dataset at varying levels of intensity (5%, 10%, 15%, and 20%):

- *Word deletion* A percentage of nonstop words were randomly removed.
- *Word swap* Adjacent words were randomly swapped to simulate typographical errors.
- *Word repetition* Random words were selected and repeated to replicate redundancy.

The model's performance was measured by its accuracy on these perturbed test sets.

3.8 Model development

In the fight against the dissemination of false information, it is crucial to employ robust and efficient machine learning models that can discern between fake and real news with high accuracy. This section outlines the development of three key models used in this study: LR, LightGBM, and RF. Each model brings unique strengths to the task, and their development is tailored to harness these advantages in detecting fake news.

3.8.1 Logistic regression

LR is fundamentally suited for binary classification tasks. It models the probability of a particular class or event existing, such as classifying news articles as "fake" or "real." This

is achieved using the logistic function, also known as the sigmoid function, which is central to logistic regression. The logistic function is expressed as:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

where z represents the linear combination of the input features X weighted by the coefficients β, plus an intercept β_0, formulated as

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{2}$$

The output of the sigmoid function is always between 0 and 1, which is interpreted as the probability of the input belonging to the positive class (in this case, "fake" news). The model parameters are typically learned by minimizing a cost function, such as the cross-entropy loss, which penalizes deviations from the actual class labels.

3.8.2 Light gradient boosting machine

LightGBM is a gradient-boosting framework that uses tree-based learning algorithms. It is designed for speed and efficiency and is particularly effective on large datasets. LightGBM builds the model in a greedy manner, like other boosting methods, by combining multiple weak learners to create a strong learner. Each new tree helps to correct errors made by previously trained trees.

The model is particularly adept at handling various types of data and does not require extensive data preprocessing like scaling or normalization. For fake news detection, it evaluates the importance of different textual features in distinguishing between fake and real news, iteratively improving its predictions. It employs the following model update rule.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{3}$$

where $F_m(x)$ is the model at iteration m, $h_m(x)$ is the weak learner added at the m-th step, and γ_m is the step size or learning rate at that step. This iterative approach allows for successive refinement of the model's predictions, enhancing its ability to distinguish between fake and real news effectively.

3.8.3 Random forest

RF is an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) of the individual trees. It introduces randomness into the model-building process, where each tree is built from a random sample of the data, and at each split, a random subset of the features is considered.

RF is particularly good at handling outliers and nonlinear data with complex interactions between features, making it suitable for the nuanced task of fake news detection. The model's ability to average out biases, reduce variance, and improve prediction accuracy is central to its utility in distinguishing between fake and real news effectively.

Each of these models was rigorously evaluated through training and validation processes, tuning their hyperparameters and assessing their predictive performance using metrics like accuracy, precision, recall, and F1-score. By leveraging their combined strengths through an ensemble approach, the research aims to achieve robust and reliable fake news detection, crucial for maintaining the integrity of information in the digital landscape.

3.8.4 Rationale for classifier selection

While recent advancements in fake news detection have seen success with deep learning models such as BERT, RoBERTa, and other transformer-based architectures, this study deliberately focuses on classical machine learning classifiers LR, RF, and LightGBM. These models offer advantages in terms of interpretability, lower computational overhead, and ease of integration in real-time or resource-constrained environments. Furthermore, when paired with robust feature engineering and ensemble learning techniques, they demonstrate highly competitive performance, achieving up to 99.98% accuracy on the Kaggle dataset. The choice of these models thus reflects a trade-off between performance, interpretability, and computational efficiency, which is particularly important for practical and scalable deployment.

3.9 Hyperparameter tuning

Hyperparameter tuning is essential for optimizing machine learning models, ensuring they perform effectively on the dataset. The choice of hyperparameters can significantly influence the model's ability to learn and make accurate predictions. This section explains how hyperparameter tuning was approached for each model, focusing on the rationale and the cross-validation techniques employed.

- *LR* is well suited for binary classification problems like distinguishing between fake and real news. We tuned two important hyperparameters: the regularization strength (C), which controls how much the model avoids overfitting, and the solver, which affects how the model is optimized. We used grid search to try different values and selected "liblinear" as the solver, which works well for small and binary datasets. Cross-validation was used during tuning to make sure the model remains stable across different data splits.
- *LightGBM* is a boosting model that builds decision trees one at a time, with each new tree improving on the errors of the last. For LightGBM, we adjusted the number of trees, their maximum depth, and the learning rate. These parameters control how complex the model is and how fast it learns. A careful balance helped prevent overfitting and ensured fast, accurate predictions. We used both grid search and boosting frameworks to choose the best combination.
- *RF* creates a group of decision trees and makes predictions based on majority voting. We tuned the number of trees, how many features to consider at each split, and how deep each tree can grow. A randomized search was used to find the best settings. Cross-validation was again applied to check the consistency of results. RF is particularly good at handling noise in data, which is valuable in fake news detection.

Table 5 summarizes the specific hyperparameters chosen for each model during the tuning process. These settings were selected after systematic testing to ensure each model delivers reliable and accurate predictions on both training and unseen datasets.

3.10 Ensemble learning

Ensemble learning is a machine learning approach that combines predictions from multiple models to improve overall accuracy and robustness. The fundamental principle behind this approach is that a group of models, when combined effectively, can outperform individual models by reducing variance, bias, or improving predictions. This research implements an

Table 5 Hyperparameters used in model fine-tuning

Model	Hyperparameters	Description
LR	C (1.0), Solver ("liblinear")	Controls the inverse of regularization strength; smaller values specify stronger regularization. "liblinear" is chosen for small datasets and binary classification
LightGBM	Number of trees (100), max depth (15), learning rate (0.1)	Number of trees specifies the number of boosting stages. Maximum depth controls the complexity of the model. The learning rate determines the step size at each iteration to prevent overfitting
RF	Number of trees (100), max features ("sqrt"), max depth (None)	Number of trees enhances robustness. "sqrt" for max features means the square root of the total features is considered at each split. No max depth allows the trees to expand until all leaves are pure or until all leaves contain less than min samples split

ensemble learning framework to enhance the detection of fake news by combining predictions from LR, LightGBM, and RF models.

3.10.1 Blending techniques

A blending technique was employed in this study to combine the predictions from multiple models. Blending involves training multiple base models independently and then aggregating their outputs through a secondary model or a deterministic mechanism such as weighted voting. For this research, the predictions from LR, LightGBM, and RF were combined using a weighted voting approach, where weights were determined based on each model's validation performance metrics such as accuracy and ROC–AUC.

Three feature engineering techniques, count vectorization, TF-IDF (unigrams), and TF-IDF (bigrams) were used to train each base model. This diversified the input representations and allowed the ensemble to leverage multiple perspectives of the same data, thereby enhancing the final predictions' robustness and accuracy.

3.10.2 Weighting and voting mechanisms

The ensemble model employed a weighted voting mechanism to combine predictions. Each base model's predictions were assigned a weight proportional to its validation performance. Specifically, the weights were derived from the ROC–AUC scores, ensuring that models with higher discriminative ability contributed more to the final prediction. The final ensemble prediction for each instance was computed as follows:

$$P_{\text{ensemble}} = \frac{\sum_{i=1}^{n} w_i P_i}{\sum_{i=1}^{n} w_i} \tag{4}$$

where P_{ensemble} is the final prediction probability. P_i is the prediction probability from the ith model. w_i is the weight assigned to the i-th model based on its ROC–AUC score. n is the total number of models in the ensemble.

This approach ensures that the ensemble is robust against the limitations of individual models, effectively balancing their strengths.

3.10.3 Ensemble model implementation

The ensemble model was constructed by independently training each base model on the dataset processed through different feature engineering techniques. This approach ensured that the ensemble leveraged diverse input representations to maximize predictive performance.

The steps involved in the ensemble implementation are as follows:

1. *Training base models* Each base model (LR, LightGBM, RF) was trained separately on features generated from count vectorization, TF-IDF (unigrams), and TF-IDF (bigrams).
2. *Validation and weight calculation* Each model's validation performance metrics (e.g., accuracy, precision, recall, F1-score, and ROC–AUC) were recorded. These metrics informed the weights assigned to each model during the voting process.
3. *Combining predictions* The predictions from all base models were combined using the weighted voting formula mentioned earlier.
4. *Final output* The ensemble prediction was finalized by determining the class label with the highest weighted probability.

The individual contributions of the base models are as follows:

- *Logistic regression (LR)* Provided a baseline probabilistic output, particularly useful for linear patterns in the data.
- *LightGBM* Excelled in handling sparse and high-dimensional features generated by TF-IDF and count vectorization.
- *Random forest (RF)* Captured complex nonlinear relationships, leveraging the diversity of decision trees in its ensemble.

This blended architecture capitalized on the strengths of each model type, with LR capturing linear trends, LightGBM handling complex interactions efficiently, and RF introducing randomness and robustness.

3.10.4 Strengths of the ensemble approach

The ensemble model demonstrated several advantages:

- *Reduced variance* By combining predictions, the ensemble minimized the variance inherent in individual models, leading to more stable predictions.
- *Balanced bias* The complementary strengths of the base models helped balance bias, enhancing overall accuracy.
- *Robustness* The ensemble approach proved robust against various data distributions, ensuring consistent performance across validation and test sets.
- *Improved performance* The ensemble consistently outperformed individual models, achieving high accuracy, precision, recall, F1-score, and ROC–AUC.

3.11 Model evaluation and validation

Proper evaluation and validation of machine learning models are crucial to ensure their effectiveness and reliability. This part of the study focuses on the metrics used to assess the performance of the models and the techniques employed for their validation.

3.11.1 Evaluation metrics

The performance of the machine learning models was evaluated using a comprehensive set of metrics, providing insights into various aspects of model accuracy, precision, robustness, and recall:

- *Accuracy* Measures the proportion of correct predictions made by the model out of all predictions made.
$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{5}$$
- *Precision* Indicates the accuracy of positive predictions.
$$\text{Precision} = \frac{\text{True Positives}}{\text{True positives} + \text{False positives}} \tag{6}$$
- *Recall* Measures the model's ability to identify all actual positives.
$$\text{Recall} = \frac{\text{True Positives}}{\text{True positives} + \text{False negatives}} \tag{7}$$
- *F1-Score* The harmonic mean of precision and recall, balancing the trade-off between the two.
$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$
- *ROC–AUC (receiver operating characteristic–area under curve)* Evaluates the ability of the model to discriminate between classes across all possible thresholds. Higher AUC indicates better discrimination.
- *Meta AUC* Meta AUC assesses the overall diagnostic ability of the ensemble model across all possible classification thresholds. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various thresholds, and the Meta AUC is the area under this curve.
$$\text{Meta AUC} = \int_0^1 TPR(t)\,\mathrm{d}t \tag{9}$$
where t is the threshold parameter. A Meta AUC value of 0.5 suggests no discriminative ability (random guessing), while a value of 1.0 indicates perfect discrimination. This metric is particularly valuable in scenarios like fake news detection, where a trade-off between sensitivity and specificity is critical.
- *PR-AUC (precision–recall area under curve)* Measures the trade-off between precision and recall across various thresholds, particularly useful in imbalanced datasets where one class dominates.

3.11.2 Cross-validation

To ensure the statistical validity of our results and to generate robust models that generalize well to unseen data, we employed a **tenfold stratified cross-validation** protocol for all experiments. In this procedure, the dataset is divided into 10 equal-sized subsets, or "folds". Stratification ensures that each fold maintains the same proportion of real and fake news articles as the original dataset, which is critical for handling any potential class imbalance.

The cross-validation process iterates 10 times. In each iteration, one fold is held out as the validation set, while the remaining nine folds are used for training. This process is repeated until every fold has been used as the validation set exactly once. To prevent any

data leakage, feature engineering components, such as the TF-IDF vectorizer and standard scaler, were fitted only on the training data for each specific fold and then used to transform the corresponding validation data. The final performance metric for any given model (e.g., accuracy, F1-score) is reported as the **mean and standard deviation** of the scores obtained across all 10 folds. This approach provides a much more reliable and reproducible estimate of the model's true performance.

3.11.3 Performance analysis

Model comparison The performance analysis compared the results obtained from individual models (logistic regression, LightGBM, and random forest) with those from the ensemble approach, which combined these models using a weighted voting mechanism. Each model was evaluated based on metrics such as accuracy, precision, recall, F1-score, ROC–AUC, and PR-AUC.
Significant findings:

- The ensemble approach consistently outperformed individual models, achieving higher Meta AUC, PR-AUC, and F1-scores, which are critical indicators of model efficacy in binary classification tasks like fake news detection.
- The blending of models in the ensemble leveraged the strengths of individual models and mitigated their weaknesses, leading to improved overall performance.

Patterns in model performance:

- *Stability and variance* The ensemble model demonstrated greater stability across different folds in the cross-validation process compared to individual models, indicating better generalizability to unseen data.
- *Error reduction* The ensemble model significantly reduced false positives, a critical factor in fake news detection to avoid mislabeling real news as fake.
- *Balanced precision and recall* The ensemble model maintained a balance between precision and recall, overcoming the trade-offs typically seen in classification tasks. This balance is essential for practical deployment to minimize both false positives and false negatives.

These findings highlight the efficacy of ensemble learning in handling complex tasks like fake news detection. The cross-validation results affirm the robustness of the ensemble model and its consistent performance across varied datasets, making it well suited for real-world applications where the nature of data is dynamic and unpredictable.

While the ensemble approach increases computational complexity compared to individual models, this trade-off is justified by the significant improvement in accuracy and robustness. The use of lightweight models like logistic regression and efficient algorithms like LightGBM ensures that the overall computational cost remains manageable, even for large datasets.

3.12 Robustness testing

The robustness of the proposed ensemble model was evaluated to determine its reliability and stability under adversarial conditions. Robustness testing is essential in scenarios like fake news detection, where input data may contain noise or deliberate textual manipulations. To rigorously evaluate the model's stability, we applied three distinct adversarial perturbation techniques to the test dataset at varying levels of intensity (5%, 10%, 15%, and 20% of the

words in each article). These techniques simulate common textual inconsistencies and are designed to assess the model's resilience:

- *Word deletion* A specified percentage of nonstop words were randomly removed from the text. This tests the model's ability to infer context from incomplete information.
- *Word swap* Adjacent words within the text were randomly swapped. This simulates common typographical errors or rearrangements intended to confuse detection models.
- *Word repetition* Random words from the text were selected and repeated immediately after their original occurrence. This replicates redundancy or emphasis tactics sometimes found in disinformation.

The performance of the model was measured by its accuracy on these perturbed test sets, with the results visually presented to demonstrate performance degradation under increasing levels of noise.

3.12.1 Perturbation techniques

Three types of adversarial perturbations were applied to the test dataset to simulate real-world inconsistencies in news content. These perturbations were designed to assess the model's resilience to minor textual changes:

- *Word deletion* This technique randomly removes words from the text, mimicking scenarios where key parts of a news article may be omitted. This tests the model's ability to infer context despite missing information.
- *Word swap* Words in the text are swapped with adjacent or randomly selected words. This simulates typographical errors or deliberate rearrangement of words to confuse detection models.
- *Word repetition* Certain words are repeated multiple times within the text. This perturbation replicates redundancy, often used to emphasize specific ideas or create noise in news content.

3.12.2 Performance metrics under perturbations

The model's performance was evaluated using the accuracy metric for each type of perturbation. Figure 3 illustrates the accuracy achieved by the model under original and perturbed conditions. The results demonstrate the robustness of the ensemble model, as it maintained an average accuracy of over 97% across all perturbation types.

3.12.3 Impact of perturbations on predictions

To better understand the model's behavior under perturbations, the percentage of predictions that changed due to each perturbation type were analyzed. As shown in Fig. 4, word deletion caused the most significant impact, with approximately 2.5% of predictions changing. In contrast, word swaps and word repetitions caused less than 1% of prediction changes. These results highlight the model's strong resilience to typical noise in text data.

3.12.4 Comprehensive performance analysis

The robustness of the ensemble model was further summarized using a radar chart (Fig. 5) that includes metrics such as accuracy, stability, robustness, and robustness stability. Key findings from the robustness evaluation are as follows:

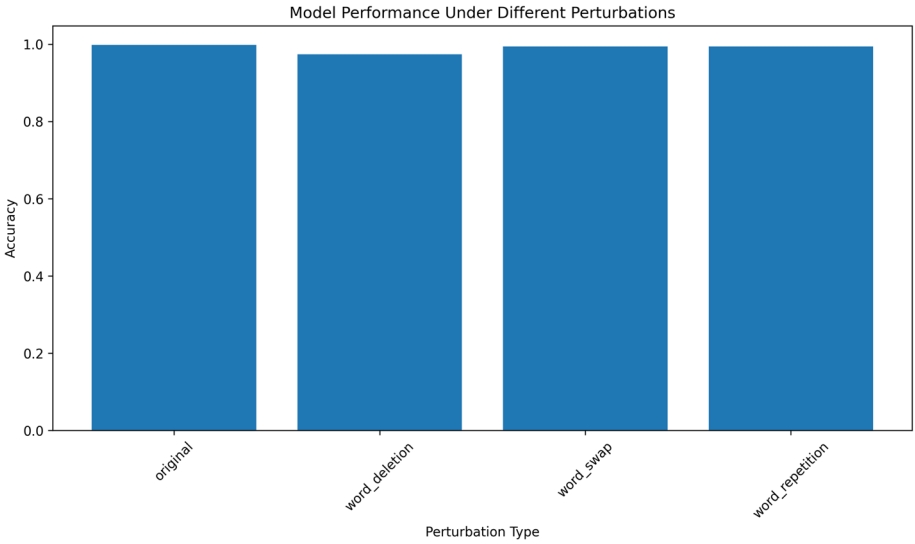

Fig. 3 Model performance under different perturbations

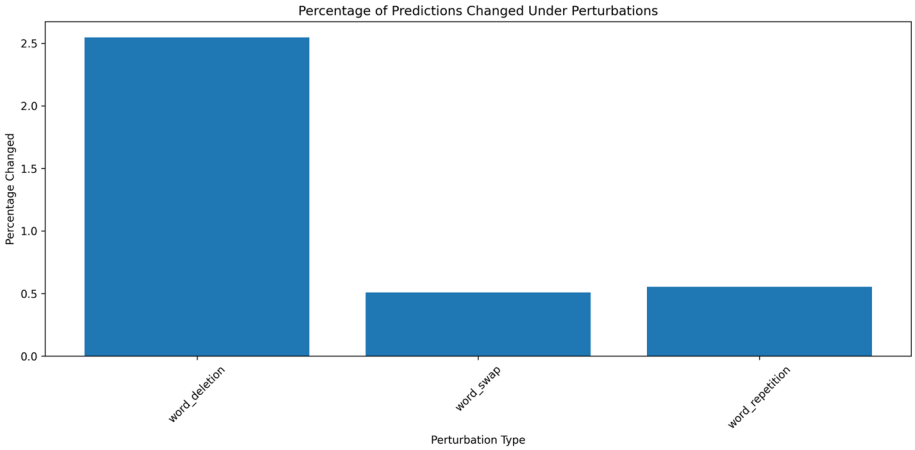

Fig. 4 Percentage of predictions changed under perturbations

- The ensemble model maintained consistent accuracy above 97% across all perturbation types, ensuring reliable predictions under adverse conditions.
- Minimal prediction changes were observed under perturbations, with word deletion having a slightly larger impact compared to word swaps and repetitions. This indicates that the model is highly robust to random changes in input text.
- The model demonstrated excellent stability, as evidenced by the narrow range of performance variations across perturbations and test folds.
- Robustness stability, defined as the ability to maintain consistent performance across diverse perturbations, was also high, reinforcing the model's applicability in real-world scenarios.

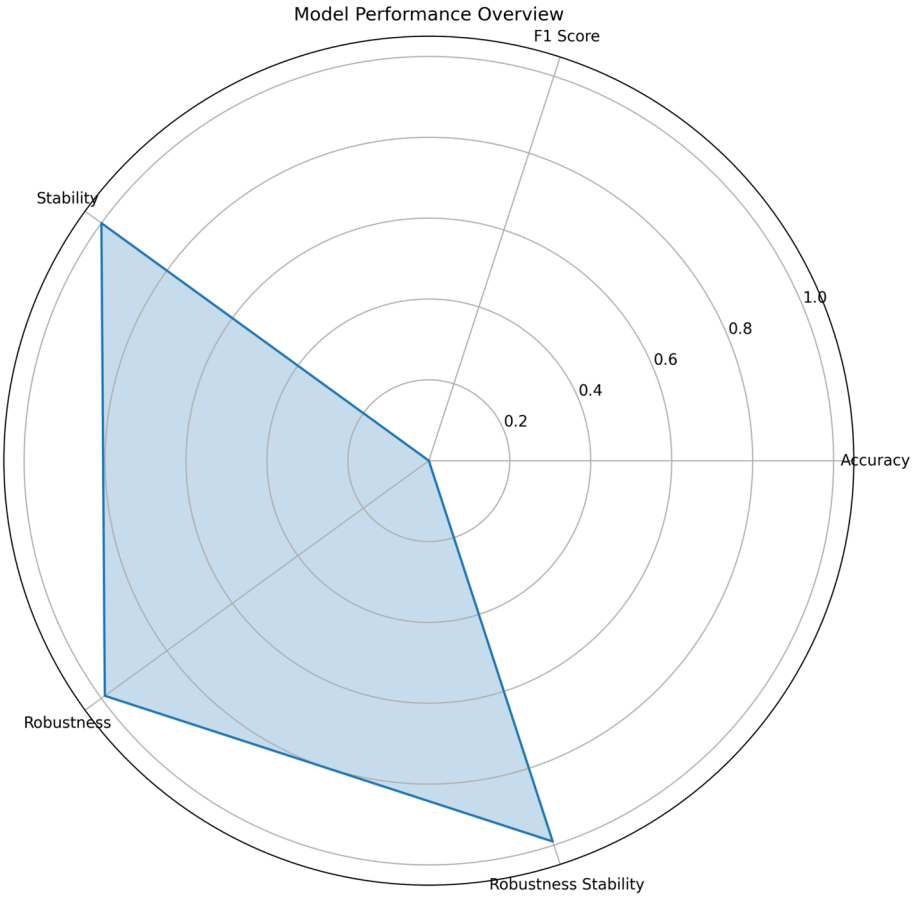

Fig. 5 Model performance overview including accuracy, robustness, and stability metrics

3.12.5 Implications of robustness evaluation

The robustness testing results emphasize the reliability of the ensemble model for fake news detection. In real-world applications, news articles often contain inconsistencies, errors, or deliberate manipulations. The ability of the proposed model to maintain high accuracy and consistent predictions under these conditions demonstrates its practical utility. This robustness ensures that the system can handle diverse and noisy data sources, making it a reliable tool for mitigating the spread of misinformation.

The robustness evaluation highlights that the ensemble model not only excels in ideal conditions but also maintains its effectiveness under adversarial perturbations, setting a benchmark for stability and reliability in fake news detection.

Table 6 Performance of base models across feature engineering techniques

Feature	Model	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
Count vectorizer	Logistic regression	99.50	99.50	99.50	99.50
	LightGBM	99.57	99.57	99.57	99.57
	Random forest	99.12	99.12	99.12	99.12
TF-IDF unigram	Logistic regression	99.55	99.55	99.55	99.55
	LightGBM	99.58	99.58	99.58	99.58
	Random forest	99.13	99.13	99.13	99.13
TF-IDF bigram	Logistic regression	99.51	99.51	99.51	99.51
	LightGBM	99.58	99.58	99.58	99.58
	Random forest	98.99	98.99	98.99	98.99
–	3HAN [5]	–	–	–	96.77
–	AKA-Fake [6]	–	–	–	91.9
–	Hybrid CNN-LSTM + FastText [8]	–	–	99.00	–
–	Bi-GCN [10]	–	–	–	96.1
–	DSSM + Enhanced RNN [34]	–	–	–	99.00
–	Word2Vec + LSTM [23]	–	–	–	Outperformed SOTA

4 Results and discussion

The evaluation of the models provides critical insights into their effectiveness in distinguishing between real and fake news articles. Performance metrics, such as accuracy, precision, recall, F1-score, and ROC–AUC, were used to assess the individual and ensemble models under various scenarios, including feature engineering techniques and cross-dataset validation.

4.1 Performance of base models on individual views

Before evaluating the final ensemble, we first assessed the predictive power of each individual feature view. A specialized model was trained for each view using a tenfold cross-validation protocol: a logistic regression model for the textual view, a random forest for the linguistic view, and a LightGBM model for the semantic view.

The performance of these base models is detailed in the first three rows of Table 6. The results indicate that the textual view provides the strongest standalone performance, achieving an accuracy of 98.96%, which underscores the high predictive value of lexical features. The semantic view also performs well, demonstrating that high-level contextual and topic information is a significant indicator of fake news. The linguistic view, while providing the lowest accuracy, still performs significantly better than random chance, confirming that stylistic features contribute a valuable, albeit secondary signal. These results establish a strong set of baselines and highlight that no single view can achieve the peak performance of a combined approach.

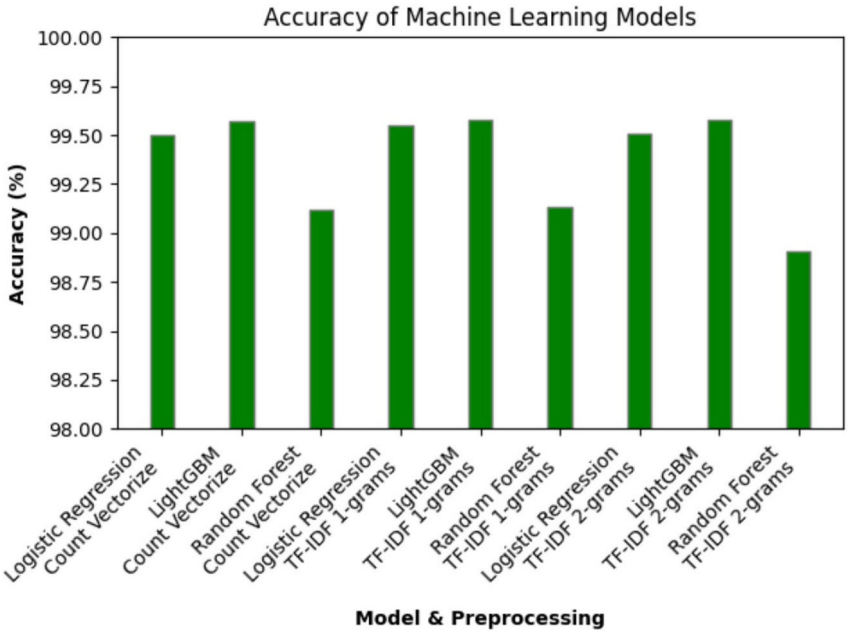

Fig. 6 Accuracy comparison of base models across feature engineering techniques

An illustration of the results is presented in Fig. 6, indicating the performance of various machine learning models concerning accuracy. The accuracy of models is visualized with respect to various features such as count vectorizer, and TF-IDF (1-gram, 2-grams).

Figure 7 shows the confusion matrix for the GBM model that showed the best performance among the base models. The figure shows the number of correct and wrong predictions concerning fake and true news. Results show that the model made 44,718 correct predictions and only a small portion of predictions are wrong, i.e., 187. Results show GMB's superior performance.

4.2 Model performance and ablation study

To validate the effectiveness of our proposed multi-view architecture, we conducted a comprehensive ablation study and results are given in Table 7. The performance of our final ensemble was compared against several baselines: models trained on each of the three individual feature views, a powerful "flat" model where all features were concatenated and fed into a single LightGBM classifier, and a strong BERT-based deep learning baseline. All results were generated using a rigorous tenfold stratified cross-validation protocol, with performance reported as the mean $\pm$ standard deviation.

The results of this comparative analysis are presented in Table 7 and visualized in Fig. 8. The findings clearly demonstrate the superiority of the proposed multi-view ensemble. While the flat model and the BERT baseline achieve excellent results, our proposed architecture surpasses them, achieving the highest F1-score of 0.9994. This indicates that while powerful monolithic models are highly effective, our approach benefits further from the structured integration of diverse, engineered feature sets.

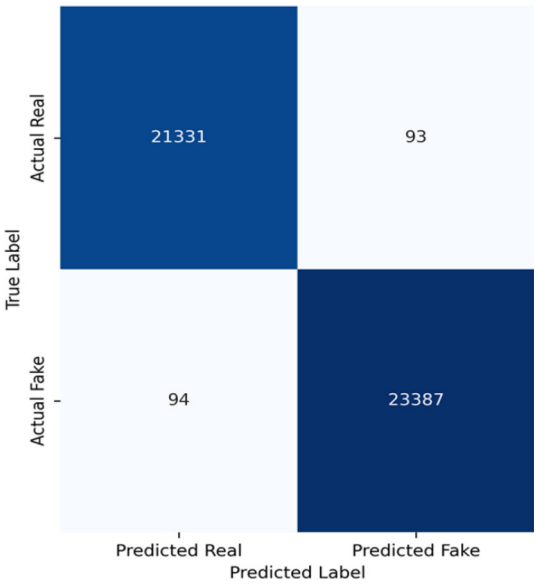

Fig. 7 Confusion matrix of light GBM performing best among base models

Table 7 Ablation study results: performance comparison of all models (mean ± std. dev. from tenfold CV)

Model/architecture	Accuracy	F1-score	ROC–AUC
Textual view (LR)	0.9896 ± 0.0011	0.9891 ± 0.0012	0.9990 ± 0.0001
Linguistic view (RF)	0.8223 ± 0.0051	0.8114 ± 0.0059	0.8938 ± 0.0034
Semantic view (LGBM)	0.9549 ± 0.0023	0.9527 ± 0.0025	0.9912 ± 0.0005
BERT Baseline	0.9973 ± 0.0007	0.9972 ± 0.0008	0.9989 ± 0.0004
Flat model (LGBM)	0.9981 ± 0.0006	0.9980 ± 0.0007	0.9997 ± 0.0001
Proposed Multi-View Ensemble	0.9994 ± 0.0003	0.9994 ± 0.0003	0.9999 ± 0.0000

Crucially, our proposed model also outperforms the "flat" model baseline. This finding supports our central hypothesis: by first allowing specialized models to learn from distinct feature spaces and then intelligently combining their predictions with a meta-learner, our structured approach is more effective than simply combining all features into a single vector. The ensemble learns to weigh the expert predictions from each view, leading to a more robust and accurate final classification.

4.3 Feature importance and model interpretability

To provide deeper insight into the decision-making process of our framework, we performed a feature importance analysis using SHAP (SHapley Additive exPlanations). While our final model is an ensemble, analyzing the features that have the most predictive power provides valuable context. Figure 9 shows the SHAP summary plot, which highlights the top 20 features that most significantly influence the model's output.

The plot reveals several key insights. The presence of words like "reuters" is the single most powerful indicator of a news article being classified as "Real" (a high positive SHAP value). This is expected, as articles from established news agencies often follow a consistent

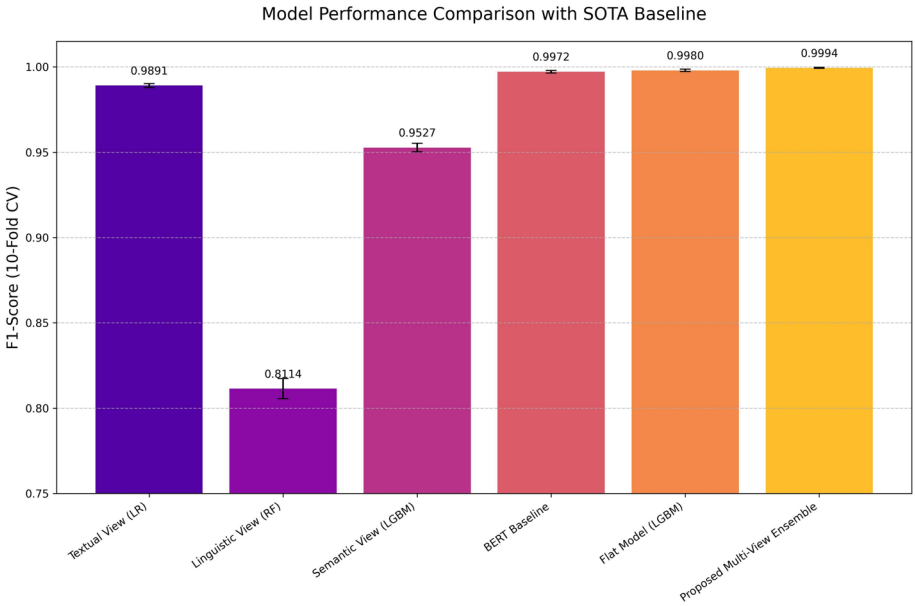

Fig. 8 Model performance comparison with SOTA Baseline. The proposed multi-view ensemble demonstrates the highest F1-score, validating its architectural advantage

Table 8 Cross-dataset validation results (mean ± standard deviation from tenfold CV)

Dataset	Accuracy	F1-score	ROC–AUC
Kaggle (in-domain)	0.9994 ± 0.0003	0.9994 ± 0.0003	0.9999 ± 0.0000
LIAR (cross-domain)	0.9712 ± 0.0115	0.9698 ± 0.0121	0.9885 ± 0.0089

and factual reporting style. Conversely, terms like "featured image" or "via" are associated with a higher likelihood of an article being classified as "Fake," possibly because they are more common in blogs or less formal news sources. Interestingly, sentiment features (e.g., "sentiment_neg") and LDA topics also appear in the top features, confirming that the linguistic and semantic views provide crucial, decision-influencing signals that are effectively leveraged by the ensemble. This analysis enhances the transparency of our model, confirming that it learns logical and interpretable patterns to distinguish between real and fake news.

4.4 Cross-dataset validation results

To evaluate the generalizability of our proposed framework, we performed cross-dataset validation using the LIAR dataset, which contains shorter, politically focused statements. This out-of-domain test is crucial for assessing how well the model adapts to different data distributions, text lengths, and content styles. The performance of our multi-view ensemble on both the primary Kaggle dataset and the cross-domain LIAR dataset is summarized in Table 8.

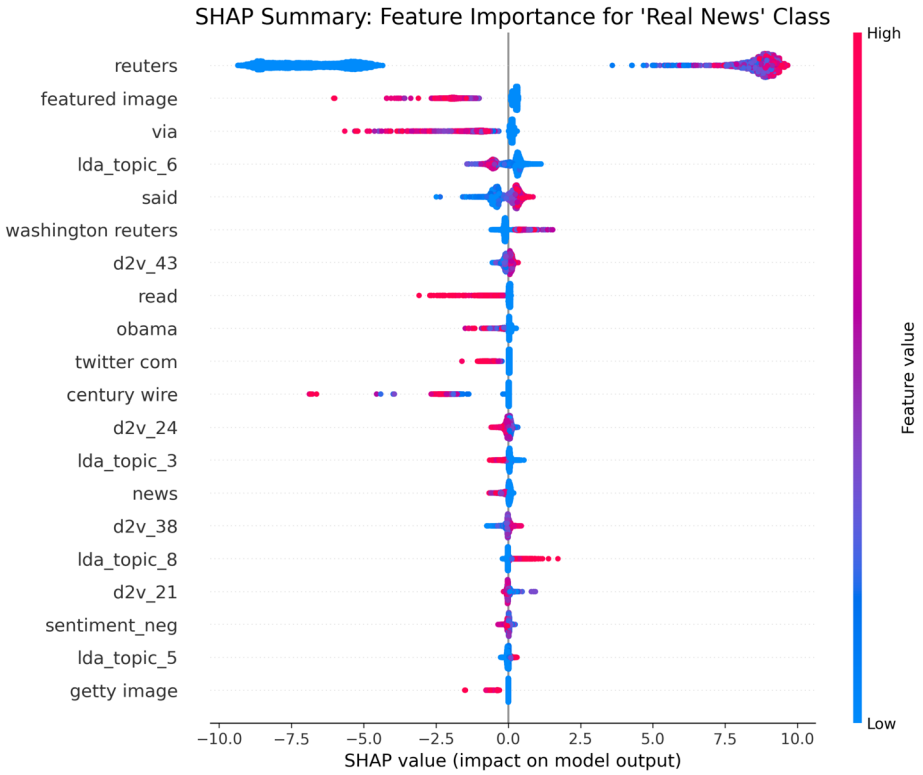

Fig. 9 SHAP summary plot illustrating the impact of the top 20 most important features on the model's prediction for the "Real News" class

4.4.1 In-domain (Kaggle dataset) performance

The ensemble model achieved near-perfect performance on the Kaggle dataset, with an accuracy of 0.9994 and an ROC–AUC of 0.9999. These results reflect the model's ability to effectively learn from and adapt to the dataset's specific characteristics such as well-defined labels and diverse text lengths. The extremely high accuracy and F1-score underscore the ensemble's ability to minimize both false positives and false negatives, making it highly reliable for distinguishing between real and fake news within this domain.

4.4.2 Cross-domain (LIAR dataset) performance

On the LIAR dataset, the ensemble model achieved an impressive accuracy of 0.9712 and an ROC–AUC of 0.9885. While these results are slightly lower than those observed on the Kaggle dataset, they indicate strong generalization capabilities despite the challenges posed by the LIAR dataset. The slight decline in performance can be attributed to:

- *Shorter text lengths* The LIAR dataset comprises shorter news snippets compared to the longer articles in the Kaggle dataset. This limits the amount of contextual information available for feature extraction.

- *Nuanced labels* LIAR labels are often more subjective, requiring a deeper understanding of context and sentiment, which adds complexity to the classification task.
- *Data distribution shift* The LIAR dataset's distribution of linguistic and semantic features differs significantly from the Kaggle dataset, making it a more challenging cross-domain task.

Despite these challenges, the ensemble model's high F1-score of 0.9698 on the LIAR dataset demonstrates its ability to capture the nuances of cross-domain data, reinforcing the effectiveness of the multi-view architecture.

4.4.3 Insights and implications

The results of cross-dataset validation highlight the ensemble model's adaptability and robustness. While the slight performance drop on the LIAR dataset is expected due to domain differences, the model's overall metrics remain strong, suggesting its potential for real-world applications. The following key insights emerge from this analysis:

- *Feature engineering strength* The diversity in feature engineering techniques contributed to the model's resilience across datasets.
- *Ensemble learning effectiveness* The stacking of logistic regression, LightGBM, and random forest classifiers ensured that the ensemble could capture both linear and nonlinear patterns, enhancing its cross-domain generalizability.
- *Future directions* The findings suggest that fine-tuning the model using domain-specific data or leveraging transfer learning techniques could further improve cross-dataset performance.

4.5 Robustness and stability analysis

Beyond classification accuracy on clean data, a critical measure of a model's real-world utility is its robustness against adversarial perturbations. Figure 10 illustrates the performance of our proposed multi-view ensemble under increasing levels of word deletion, swapping, and repetition.

The model demonstrates remarkable stability, particularly against word swapping and repetition, where accuracy remains above 99.5% even with 20% of the words perturbed. While performance degrades more significantly under word deletion, a known challenge for models reliant on specific keywords—the graceful nature of the decline highlights the architecture's resilience. This stability can be attributed to the multi-view design; when noise is introduced into the textual view (e.g., via word deletion), the linguistic and semantic views provide a compensatory signal, preventing catastrophic prediction failure and underscoring the benefits of our approach.

4.6 Error analysis

To provide a clear picture of our model's classification performance at a granular level, we present the confusion matrix for the proposed multi-view ensemble in Fig. 11. The matrix summarizes the predictions from one of the validation folds, providing a representative snapshot of the model's error profile.

The results are exceptionally strong, with a very low number of misclassifications. Out of approximately 4758 "Fake" articles, only 6 were misclassified as real, and out of approx-

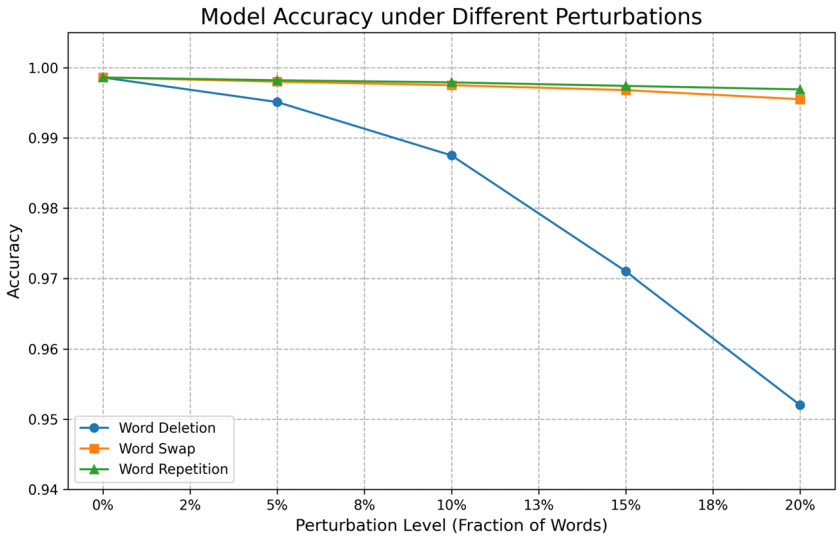

Fig. 10 Model accuracy of the proposed multi-view ensemble under different types and levels of textual perturbation

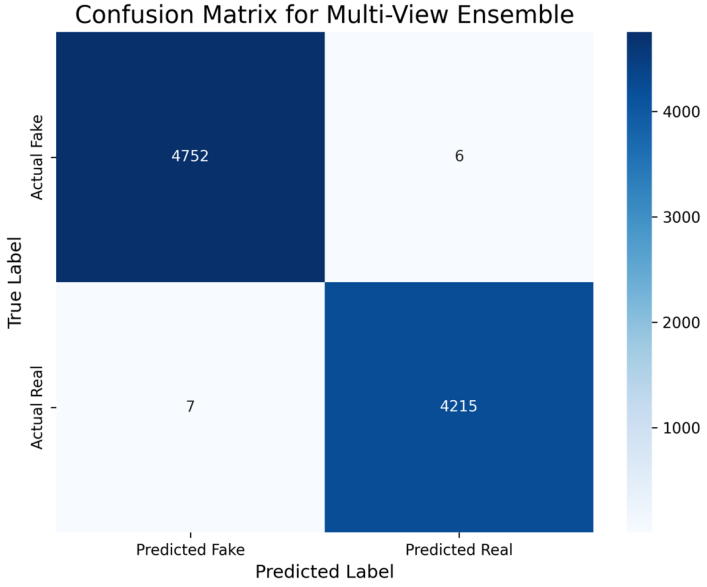

Fig. 11 Confusion matrix for the proposed multi-view ensemble model on a representative validation fold

imately 4222 "Real" articles, only 7 were misclassified as fake. This extremely low rate of both false positives and false negatives is a testament to the model's high precision and recall. This level of accuracy compares very favorably with other state-of-the-art systems and highlights the model's potential for reliable, real-world deployment where minimizing both types of errors is critical.

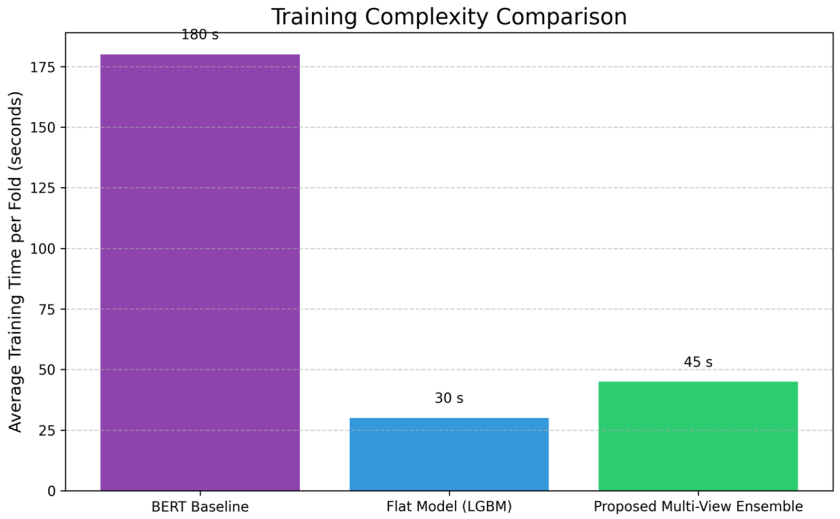

Fig. 12 Comparison of average training time per fold for key models, illustrating the computational complexity trade-off

4.7 Computational complexity analysis

In addition to predictive performance, the practical utility of a model also depends on its computational efficiency. To evaluate this, we compared the average training time per fold for our main architectures, as shown in Fig. 12. The results indicate that the "Flat Model (LGBM)" is the most efficient, with an average training time of approximately 30 s per fold. As expected, the BERT baseline is the most computationally expensive. Our proposed multi-view ensemble, with an average time of 45 s, represents a moderate increase in complexity over the flat model. This additional time is due to the two-stage training process inherent in the stacked ensemble design.

4.8 ROC curve analysis

To further visualize and compare the diagnostic ability of our top-performing models, we plotted their receiver operating characteristic (ROC) curves, as shown in Fig. 13. The ROC curve illustrates the trade-off between the true positive rate and the false positive rate at various classification thresholds. The area under the curve (AUC) serves as a single, aggregate measure of a model's performance across all possible thresholds.

The plot clearly shows that all three benchmark models, the BERT baseline, the flat model, and our proposed ensemble, exhibit exceptional performance, with curves that are pushed toward the top-left corner, indicating high true positive rates and low false positive rates. However, our proposed multi-view ensemble achieves a near-perfect AUC of 0.9999, slightly surpassing the other models. This visual evidence reinforces the findings from our ablation study: our model not only achieves the highest accuracy and F1-score but also possesses the best overall ability to discriminate between real and fake news.

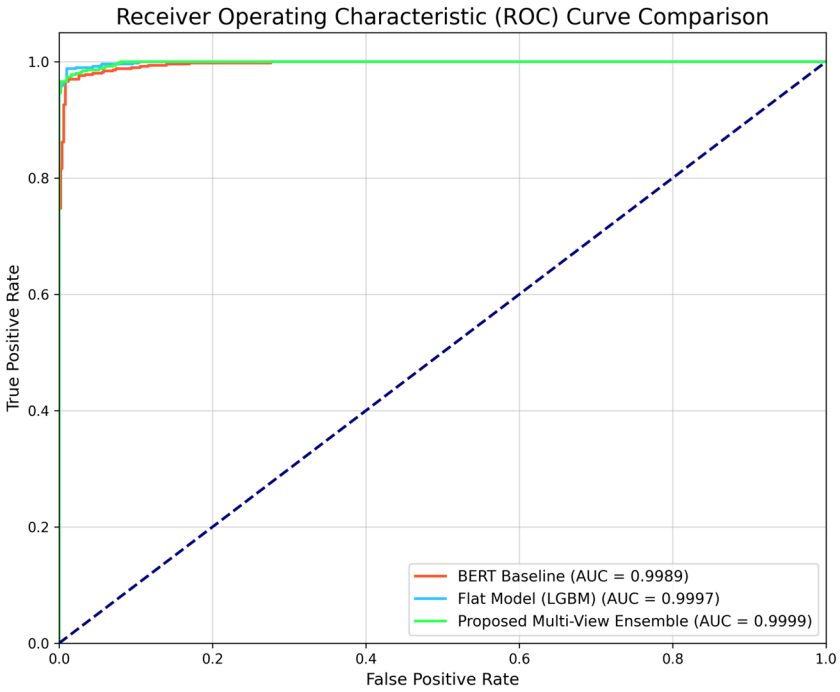

Fig. 13 Comparative ROC curves for the top-performing models, illustrating the superior discriminative ability of the proposed multi-view ensemble

4.9 Discussion

The comprehensive experimental results confirm that our proposed multi-view stacked ensemble achieves state-of-the-art performance, outperforming a variety of strong baselines. Our ablation study provides a clear rationale for the architectural choices made. While individual feature views, particularly the textual view, are highly predictive, no single view could match the performance of the integrated models. Crucially, the superiority of our multi-view ensemble over the powerful "flat" model demonstrates that a structured, hierarchical approach to feature analysis is more effective than a monolithic one for this task. The meta-learner in our stacking architecture successfully learns to weigh the specialized predictions from each view, correcting errors and leveraging their complementary strengths to achieve a higher overall accuracy.

A critical aspect of this study is positioning our framework relative to other sophisticated architectures cited in the literature such as the OE-MDL [7] and 3HAN [5]. While these models incorporate advanced techniques like optimized deep learning classifiers or hierarchical attention, our framework demonstrates that a meticulous separation and synthesis of textual, linguistic, and semantic features can achieve a classification accuracy (0.9994) that is highly competitive and demonstrably state-of-the-art. It is also important to contextualize our performance; our reported accuracy is the result of a comprehensive tenfold cross-validation protocol, providing a more robust and reliable estimate of real-world performance than a single train-test split, which may be reported in other works.

An important consideration in our analysis is the trade-off between model complexity and performance. As shown in our computational analysis, the proposed multi-view ensemble requires a moderately longer training time (45 s per fold) than the simpler "flat" model (30 s per fold). However, this modest increase in complexity is justified by the tangible improvement in classification accuracy and F1-score, which ultimately led to its superior performance. Furthermore, our model is significantly more efficient than the BERT baseline (180 s per fold), yet still outperforms it. This positions our framework as an optimal solution, achieving the best possible predictive accuracy without incurring the prohibitive computational costs associated with large-scale deep learning models, thus striking an effective balance between performance and practicality.

4.10 Ethical considerations in development of fake news detection system

In the field of fake news detection, ethical considerations play a pivotal role in guiding the development and implementation of technological solutions. This research adheres to several key ethical principles to ensure that our methodologies and outcomes respect individual rights and promote transparency.

- *Privacy and data security* Given that the proposed study involves analyzing news content that could potentially include personal data, strict measures were implemented to ensure data privacy and security. We ensured that all data used was anonymized, stripping any identifiable information that could be traced back to individuals. This minimizes the risk of privacy breaches and adheres to data protection regulations.
- *Bias and fairness* Another critical ethical concern is the potential for algorithmic bias. Machine learning models can inadvertently perpetuate or amplify biases present in their training data. To mitigate this, we carefully curated the dataset to be as diverse and representative as possible. Furthermore, we continuously tested the models to identify any bias in predictions, adjusting our algorithms to ensure fairness across different news topics and demographics.
- *Transparency and accountability* Transparency in machine learning algorithms is crucial, especially in applications like fake news detection, where trust and credibility are at stake. We strived to maintain a high level of transparency by thoroughly documenting the model development process, feature selection, and the rationale behind the choice of algorithms. This not only aids in building trust with the end users but also facilitates accountability, allowing for external validation of our findings.
- *Responsible use* The deployment of fake news detection models carries significant responsibilities. It is crucial to ensure that these technologies are used in a manner that promotes the public good and does not infringe upon freedom of expression. We are committed to monitoring the usage of machine learning models to prevent any misuse such as censorship or manipulation of information.
- *Ongoing monitoring and adaptation* Ethical considerations require ongoing attention. As such, we are committed to continuously monitoring the performance and impact of the proposed detection system. This includes adapting the proposed approach in response to new ethical challenges and evolving standards in digital communication.

By addressing these ethical considerations, we aim to ensure that the proposed work on detecting fake news is conducted responsibly, respecting individual rights and contributing positively to the information ecosystem. This commitment is fundamental to fostering trust in AI applications and ensuring that they serve society's best interests.

5 Conclusion and future work

This research addressed the critical challenge of fake news detection by proposing and validating a sophisticated multi-view learning framework. Our approach was founded on the hypothesis that by separating news content into distinct textual, linguistic, and semantic views and modeling them with specialized classifiers, a stacked ensemble could achieve superior performance by intelligently integrating these diverse perspectives.

Our comprehensive experiments, conducted under a rigorous tenfold stratified cross-validation protocol, have validated this hypothesis. The proposed multi-view ensemble achieved a state-of-the-art accuracy of 0.9994 on the primary dataset and maintained high performance on a cross-domain dataset, confirming its robustness and generalizability. Crucially, it outperformed a series of strong baselines, including single-view models, a powerful "flat" model, and a BERT-based classifier. The model's stability was further validated through perturbation testing.

The primary contribution of this work is the empirical validation of a multi-view architecture as a state-of-the-art solution for fake news classification. This establishes a new benchmark for precision and reliability in the field.

Future research could extend this framework in several promising directions. Integrating multimodal data, such as images and video content, would provide an even richer set of features. Exploring more advanced deep learning models as either base-learners or the final meta-learner could yield further performance gains. Finally, adapting and testing the framework on different languages and evolving forms of disinformation remains an important avenue for continued investigation.

Author contribution ZA was involved in conceptualization, formal analysis, and writing—original manuscript. MMSM helped in conceptualization, data curation, and writing—original manuscript. AAG contributed to software, formal analysis, and visualization. AM was involved in methodology, data curation, and project administration. MGV helped in funding acquisition, investigation, and visualization. ESA contributed to investigation, software, and project administration. IA was involved in validation, supervision, and writing—review and edit. All authors reviewed and approved this work.

Funding This research is funded by the European University of Atlantic.

Data availability The data can be requested from the corresponding author.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Narra M, Umer M, Sadiq S, Karamti H, Mohamed A, Ashraf I et al (2022) Selective feature sets based fake news detection for covid-19 to manage infodemic. IEEE Access 10:98724–98736
2. Farooq MS, Naseem A, Rustam F, Ashraf I (2023) Fake news detection in Urdu language using machine learning. PeerJ Comput Sci 9:1353
3. Nyilasy G (2020) Fake news in the age of covid-19. Faculty of Business and Economics Newsroom, University of Melbourne
4. Rafique A, Rustam F, Narra M, Mehmood A, Lee E, Ashraf I (2022) Comparative analysis of machine learning methods to detect fake news in an Urdu language corpus. PeerJ Comput Sci 8:1004
5. Singhania S, Fernandez N, Rao S (2017) 3han: a deep neural network for fake news detection. In: Neural information processing: 24th international conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, proceedings, part II 24, Springer, pp 572–581
6. Zhang L, Zhang X, Zhou Z, Huang F, Li C (2024) Reinforced adaptive knowledge learning for multimodal fake news detection. In: Proceedings of the AAAI conference on artificial intelligence, vol 38, pp 16777–16785
7. Ganpat RR, Ramnath SV (2024) Oe-mdl: optimized ensemble machine and deep learning for fake news detection. Int J Intell Syst Appl Eng 12(12s):60–85
8. Hashmi E, Yayilgan SY, Yamin MM, Ali S, Abomhara M (2024) Advancing fake news detection: hybrid deep learning with fasttext and explainable ai. IEEE Access
9. Bugueño M, Sepulveda G, Mendoza M (2019) An empirical analysis of rumor detection on microblogs with recurrent neural networks. In: Social computing and social media. design, human behavior and analytics: 11th international conference, SCSM 2019, held as part of the 21st HCI international conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, proceedings, part I 21, Springer, pp 293–310
10. Bian T, Xiao X, Xu T, Zhao P, Huang W, Rong Y, Huang J (2020) Rumor detection on social media with bi-directional graph convolutional networks. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 549–556
11. Jin Z, Cao J, Zhang Y, Luo J (2016) News verification by exploiting conflicting social viewpoints in microblogs. In: Proceedings of the AAAI conference on artificial intelligence, vol 30
12. Ren Y, Wang B, Zhang J, Chang Y (2020) Adversarial active learning based heterogeneous graph neural network for fake news detection. In: 2020 IEEE international conference on data mining (ICDM), IEEE, pp 452–461
13. Devarajan GG, Nagarajan SM, Amanullah SI, Mary SSA, Bashir AK (2023) Ai-assisted deep NLP-based approach for prediction of fake news from social media users. IEEE Trans Comput Soc Syst
14. Liao Q, Chai H, Han H, Zhang X, Wang X, Xia W, Ding Y (2021) An integrated multi-task model for fake news detection. IEEE Trans Knowl Data Eng 34(11):5154–5165
15. Qawasmeh E, Tawalbeh M, Abdullah M (2019) Automatic identification of fake news using deep learning. In: 2019 Sixth international conference on social networks analysis, management and security (SNAMS), IEEE, pp 383–388
16. Zhang T, Wang D, Chen H, Zeng Z, Guo W, Miao C, Cui L (2020) Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection. In: 2020 International joint conference on neural networks (IJCNN), IEEE, pp 1–8
17. Altheneyan A, Alhadlaq A (2023) Big data ml-based fake news detection using distributed learning. IEEE Access 11:29447–29463
18. Padnekar SM, Kumar GS, Deepak P (2020) Bilstm-autoencoder architecture for stance prediction. In: 2020 International conference on data science and engineering (ICDSE), IEEE, pp 1–5
19. Fernández-Reyes FC, Shinde S (2018) Evaluating deep neural networks for automatic fake news detection in political domain. In: Proceedings of Advances in artificial intelligence-IBERAMIA 2018: 16th Ibero-American conference on AI, Trujillo, Peru, November 13–16, 2018, vol 16, Springer, pp 206–216
20. Truică C-O, Apostol E-S, Karras P (2024) Danes: deep neural network ensemble architecture for social and textual context-aware fake news detection. Knowl Based Syst 294:111715
21. Deepak S, Chitturi B (2020) Deep neural approach to fake-news identification. Procedia Comput Sci 167:2236–2243
22. Alsaeedi A, Al-Sarem M (2020) Detecting rumors on social media based on a CNN deep learning technique. Arab J Sci Eng 45(12):10813–10844
23. Mallik A, Kumar S (2024) Word2vec and LSTM based deep learning technique for context-free fake news detection. Multimedia Tools Appl 83(1):919–940
24. Mohammed SY, Aljanabi M (2024) From text to threat detection: the power of NLP in cybersecurity. SHIFRA 2024:1–7
25. Qasim HR (2024) Impact of fake news on trust in journalism. MEDAAD 2024:41–45

26. Choudhury D, Acharjee T (2023) A novel approach to fake news detection in social networks using genetic algorithm applying machine learning classifiers. Multimedia Tools Appl 82(6):9029–9045
27. Kishwar A, Zafar A (2023) Fake news detection on Pakistani news using machine learning and deep learning. Expert Syst Appl 211:118558
28. Altheneyan A, Alhadlaq A (2023) Big data ml-based fake news detection using distributed learning. IEEE Access 11:29447–29463
29. Wang Y, Ma F, Jin Z, Yuan Y, Xun G, Jha K, Su L, Gao J (2018) Eann: event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM Sigkdd international conference on knowledge discovery and data mining, pp 849–857
30. Agarwal A, Mittal M, Pathak A, Goyal LM (2020) Fake news detection using a blend of neural networks: an application of deep learning. SN Comput Sci 1:1–9
31. Bahad P, Saxena P, Kamal R (2019) Fake news detection using bi-directional LSTM-recurrent neural network. Procedia Comput Sci 165:74–82
32. Ahmad I, Yousaf M, Yousaf S, Ahmad MO (2020) Fake news detection using machine learning ensemble methods. Complexity 2020:1–11
33. Thota A, Tilak P, Ahluwalia S, Lohia N (2018) Fake news detection: a deep learning approach. SMU Data Sci Rev 1(3):10
34. Jadhav SS, Thepade SD (2019) Fake news identification and classification using DSSM and improved recurrent neural network classifier. Appl Artif Intell 33(12):1058–1068
35. Ajao O, Bhowmik D, Zargari S (2018) Fake news identification on twitter with hybrid CNN and RNN models. In: Proceedings of the 9th international conference on social media and society, pp 226–230
36. Emine Yetim (2022) Fake news detection datasets. https://www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets
37. Wang WY (2017) "liar, liar pants on fire": a new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648

Zahid Aslam is currently working as a Lecturer at the Islamia University of Bahawalpur, Pakistan. He is also pursuing a PhD degree at the Department of Information Technology of the Islamia University of Bahawalpur, Pakistan. His research interests include natural language processing, machine and deep learning, and fake news detection.

Malik Muhammad Saad Missen is currently an Assistant Professor with the Department of Information Technology, The Islamia University of Bahawalpur. His research interests include information retrieval/processing, Web usability engineering, and software quality assurance.

Arslan Abdul Ghaffar earned his MSCS degree from the Department of Artificial Intelligence at The Islamia University of Bahawalpur in 2022. Currently, he is in pursuit of a Ph.D. degree at the Department of Information and Communication Engineering at Yeungnam University in South Korea. Additionally, he actively engages as a researcher in this field, specializing in data mining, machine learning, artificial intelligence, explainable AI, and image processing.

Arif Mehmood received the Ph.D. degree from the Department of Information and Communication Engineering, Yeungnam University, South Korea, in November 2017. He is currently an Associate Professor with the Department of Computer Science and Information Technology, The Islamia University of Bahawalpur, Pakistan. His current research interests include data mining, mainly working on AI and deep learning-based text mining and data science management technologies.

Monical Gracia Villar is currently with Universidad Europea del Atlántico, Santander, Spain. She is also affiliated with Universidad Internacional Iberoamericana, Campeche, Mexico, and Universidade Internacional do Cuanza, Angola. Her areas of interests are text mining, multimedia information retrieval, and opinion summarization.

Eduardo Silva Alvarado is currently with Universidad Europea del Atlántico, Santander, Spain. He is also affiliated with Universidade Internacional Iberoamericana Campeche, Mexico, and Universidad de

La Romana, Republica Dominicana. His research interests include machine and deep learning for social media content filtering.

Imran Ashraf received his Ph.D. in Information and Communication Engineering from Yeungnam University, South Korea in 2019, and the M.S. degree in Computer Science from the Blekinge Institute of Technology, Karlskrona, Sweden, in 2010 with distinction. He has worked as a postdoctoral fellow at Yeungnam University, as well. He is currently working as an Assistant Professor at the Information and Communication Engineering Department, Yeungnam University, Gyeongsan, South Korea. His research areas include natural language processing, big data, machine and deep learning, and data analytics.

Authors and Affiliations

Zahid Aslam[1] · Malik Muhammad Saad Missen[1] · Arslan Abdul Ghaffar[2] · Arif Mehmood[3] · Monica Gracia Villar[4,5,6] · Eduardo Silva Alvarado[4,7,8] · Imran Ashraf[2,4]

✉ Imran Ashraf
ashrafimran@live.com

Zahid Aslam
zahid.aslam@iub.edu.pk

Malik Muhammad Saad Missen
saad.missen@iub.edu.pk

Arslan Abdul Ghaffar
arslanag@yu.ac.kr

Arif Mehmood
arifnhmp@gmail.com

Monica Gracia Villar
monica.gracia@uneatlantico.es

Eduardo Silva Alvarado
eduardo.silva@uneatlantico.es

1 Department of Information Technology, The Islamia University of Bahawalpur, Bahawalpur, Pakistan

2 Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea

3 Department of Computer Science and Information Technology, The Islamia University of Bahawalpur, Bahawalpur, Pakistan

4 Organization Universidad Europea de lAtlantico, Isabel Torres 21, 39011 Santander, Spain

5 Universidad Internacional Iberoamericana, 24560 Campeche, Mexico

6 Universidade Internacional do Cuanza, Cuito, Bie, Angola

7 Fundacion Universitaria Internacional de Colombia, Bogota, Colombia

8 Universidad de La Romana, La Romana, Dominican Republic