scientific reports

Check for updates

OPEN

Transfer learning driven fake news detection and classification using large language models

Basma S. Alqadi[1], Suliman A. Alsuhibany[2], Samia Nawaz Yousafzai[3], Sharf Alzu'bi[4], Deema Mohammed Alsekait[5]✉ & Diaa Salama AbdElminaam[6,7]

Today, the problem of using social media to spread false information is not only widespread but also quite serious. The extensive dissemination of fake news, regardless of whether it is produced by human beings or computer programs, has a negative impact not only on society but also on individuals in terms of politics and society. Currently of social networks, the quick dissemination of news provides a challenge when it comes to establishing the reliability of the information in a satisfactory manner. Because of this, the requirement for automated technologies that can identify fake news has become of the utmost importance. Existing fake news detection methods often suffer from challenges such as limited labeled data, inability to fully capture complex linguistic nuances, and inadequate integration of different embedding techniques, which restrict their effectiveness and generalizability. In this work, we propose a novel multi-stage transfer learning framework that leverages the strengths of pre-trained large language models, particularly RoBERTa, tailored specifically for fake news detection in limited data scenarios. Unlike prior studies which primarily rely on standard fine-tuning, our approach introduces a systematic comparison of word embedding techniques such as Word2Vec and one-hot encoding, combined with a refined fine-tuning process to enhance model performance and interpretability. The experimental results on two real-world benchmark datasets demonstrate that our method achieves a significant accuracy improvement of at least 3.9% over existing state-of-the-art models, while also providing insights into the role of embedding techniques in fake news classification. To address these limitations, our approach fills the gap by combining multi-stage transfer learning with embedding comparisons and task-specific optimizations, enabling more robust and accurate detection on small datasets. Based on the findings of our experiments conducted on two datasets derived from the real world, we have determined that the transfer learning-based strategy that we have developed can outperform the most advanced approaches by a minimum of 3.9% in terms of accuracy and offering a rational explanation.

Keywords Transfer learning, Large language models, Fake news detection, Deep learning, RoBERTa, Word embedding

Despite the fact that social media has only been growing and developing for a relatively short period of time, individuals have become unable to imagine their lives without it as it not only connects people and allows them to participate in shared activities, but also provides an opportunity to witness significant events in the global community[1,2]. Nevertheless, a significant proportion of users (approximately 57%) consume news through social networks, where most of the posts are inaccurate or fake. The consequences of such misrepresentation are more deleterious than supposed in most cases[3]. The problem is that now even when connected to the internet anyone can spread information and rumors spread quickly. Such rumors, as a rule, appear in the form of comments, which are among the most important indicators when it comes to determining the authenticity of trending information[4]. Public opinion, therefore, plays a crucial role in ascertaining whether a piece of information is

[1]Computer Science Department,College of Computer and Information Science, Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia. [2]Department of Computer Science,College of Computer, Qassim University, 51452 Buraydah, Saudi Arabia. [3]Department of Computer Science, HITEC University Taxila, Taxila, Punjab 47080, Pakistan. [4]Department of Information Technology,College of Engineering and Technology, Royal University for Women, West Riffa, Bahrain. [5]Department of Information Technology,College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, 11671 Riyadh, Saudi Arabia. [6]Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt. [7]Jadara University, Jadara Research Center, Irbid 21110, Jordan. ✉email: Dmalsekait@pnu.edu.sa

authentic or a fake one[5,6]. However, inadequate opinions and comments are usually made later in the early stages of rumor dissemination to further substantiate the authenticity of the message[7]. Therefore, there is increasing demand in the creation of an intelligently automated system capable of determining the veracity of rumors without using these preliminary signs[8,9]. Promising new developments in machine learning, including transfer learning and language models, offer a solid basis for creating strong frameworks for early identification of more fake news. While several recent works employ transfer learning and fine-tuning of pre-trained language models like BERT and RoBERTa for fake news detection, these methods often overlook the impact of different word embedding techniques and lack adaptation strategies tailored for limited data scenarios.

Fake news is posted on social media with the intention to deceive the public, and sometimes greatly resembles real news articles. In today's society where the general public relies both on social interface and sources for news updates, such fake information may go viral in the social networks. However, there are several drawbacks to getting news through social media sites It is possible to get news through social media instantaneously and free of charge while on the go, especially local news. Among such challenges, the foremost one is the inability to check the veracity of the published news once they have been released, and this has a destructive effect on society. The overload of information, along with the constraints concerning the knowledge, time, or ability to access high-quality sources, enables an unrelated person to separate truths from myths. This has underlined the importance of the development of automated and supportive systems that can detect false news at an early stage. As the field of artificial intelligence (AI) progresses, researchers are applying AI technologies to develop tools that can detect and mitigate fake news on their own[10]. In this study, we propose a novel multi-stage transfer learning framework that not only fine-tunes pre-trained language models but also systematically compares embedding techniques such as Word2Vec and one-hot encoding to improve fake news detection performance, particularly in datasets with limited labeled samples.

Previous approaches in the automatic detection of fake news primarily focuses on the synthesis of efficient features derived from various domains including textual content[11,12], publisher information[13], and communications ?,[14]. While these approaches provided useful outcomes, they were not very efficient because they were very heavy in terms of time and the use of manpower. Also, the quality of the both models mostly depended on the quality of the manually designed features. Thus, while these approaches were reasonable in academic settings, their efficacy in real-world situations was somewhat limited due to differences in the quality and uniformity of the features being extracted.Following the rapid advancement and effectiveness of deep neural networks, many recent studies[15–17] have proposed various types of neural networks for fake news classification. For example, there are recurrent neural networks (RNNs)[9,18] that can be used to capture the representation of the tweet text through the posting timeline. The propagation paths were formulated as multivariant time series and combined RNNs with convolutional neural networks (CNNs) to learn the dynamic and temporal patterns of users throughout the propagation process. The major limitation of such approaches is that they require a huge amount of labeled data that can sometimes be very difficult to gather. Further, these methods are designed to deal with sequentially arranged data, while in the case of the propagation structure, other types of data are more suitable, so the results of using these methods are approximate. It is within this background that this study aims at identifying techniques that achieve high accuracy even when working with a small amount of data.

This study analyzed a comprehensive dataset of fabricated news. Initially, several preprocessing approaches were employed to clean the text. We refined multiple large language models (LLMs), enhancing their functionalities expressly for the task of false news categorization. This entailed modifying the models to enhance their comprehension of the intricate language patterns inherent in misleading or incorrect information, while also utilizing their capacity to discern complicated linkages between words and settings in the news. In this study, we presented two types of comparisons regarding our proposed work: the first involves word embedding techniques, including one-hot encoding and Word2Vec, while the second compares various machine learning (ML) and deep learning (DL) algorithms to ascertain the optimal performance of the proposed model. To address limitations in previous approaches, our methodology introduces a multi-stage transfer learning framework that strategically fine-tunes RoBERTa in two phases for enhanced domain adaptation. Unlike existing works, we incorporate a comparative analysis of embedding methods Word2Vec and One-Hot Encoding within the same framework to evaluate their effectiveness in low-resource settings. Furthermore, our study investigates the impact of controlled layer freezing and adaptive fine-tuning schedules, offering a deeper understanding of model behavior under constrained data conditions. These design choices represent a methodological advancement that balances generalization and task-specific adaptation in fake news detection.

The motivation behind our proposed model stems from the limitations observed in existing fake news detection methods, particularly their reliance on large labeled datasets and lack of flexibility in embedding strategies. Unlike previous studies that primarily focus on single-model fine-tuning, our framework prioritizes adaptability and performance on low-resource datasets. This ensures practicality in real-world scenarios where data is often scarce or noisy, addressing a critical gap in current research.

The main objectives of this paper are as follows: (1) to develop a multi-stage transfer learning framework tailored for effective fake news detection on limited datasets; (2) to systematically compare word embedding techniques such as Word2Vec and one-hot encoding within this framework to evaluate their impact on model performance; (3) to conduct comprehensive experiments on real-world benchmark datasets to validate the proposed method's accuracy and robustness; and (4) to provide insights into embedding roles and fine-tuning strategies to guide future research in this domain.

Rest of the paper is organized as follows. In Sect. 2, we review recent work on detecting fake news. Section 3 outlines the proposed transfer learning technique. In Sect. 4, experiments and findings are presented. Finally, we analyze the results and conclude the proposed work in Sect. 5.

Literature review

The rapid spread of misinformation and fake news on social media platforms has created significant challenges, prompting extensive research into automated detection methods. Various approaches employing machine learning, deep learning, and natural language processing techniques have been developed to accurately identify fake news across diverse datasets and domains. These methods range from domain-invariant feature extraction models to hybrid neural networks, ensemble learning frameworks, and transformer-based architectures. The evolving landscape of fake news detection reflects the ongoing effort to improve accuracy, adaptability, and real-time performance of detection systems.

In fake news detection, different models have been developed with method as ML, DL, and natural language processing. Domain Adversarial and Graph Attention Neural Network (DAGANN) was introduced for fake news detection. The DAGANN model allows to detect fake news in different domains and events through domain-invariant feature extraction. In the fake news detection scenario, results from the simulation indicated that DAGANN offered very high performance in different domains of Weibo and twitter datasets[19]. Similarly,[20] proposed an improved approach using RNN and CNN. The hybrid CNN-RNN approach outperformed non-hybrid baselines, according to their results with the FA-KES and ISOT datasets. Furthermore, various feature extraction techniques are presented in the context of the automated identification of false information on social networks. They classified their work as consisting of analyzing the behavior of the Facebook accounts employing a deep learning-based analyzer and adopting other features associated with the accounts to enhance identification[21].

An innovative ensemble learning model was developed for fake news detection, integrating four classifiers: These models include N-gram CNN, LSTM, depth-LSTM, and linguistic inquiry and word count. Self adaptive harmony search is used to optimize the weights in the ensemble model to check the classification of fake news[22]. A two-step method was presented for the fake news detection on social media. Firstly, they used different data preprocessing methodologies like, document-term matrices and term frequency weighting, to transform raw data into a structure form. In the second phase, they used around 23 Artificial Intelligence techniques on the preprocessed data for fake news detection[23]. To provide an example, authors in[24] proposed Elementary Discourse Unit that works at word-sentence level, with the goal of immediately identifying fake news after their release. Based on textual and visual information of the news articles, was introduced using FND-CLIP framework. The framework combines deep learning features of text with the BERT based encoder and images with the ResNet based encoder[25].

The fusion model involves the merger of DeepCnnLstm and DeepCnnBilstm as these two structures provided the highest accuracy in FA-KES fake news detection[26]. Another system described in[27] makes use of vote classifiers, features extraction, and feature selection. This system proves to have high accuracy and precision in the classification of fake and real news when tested on the ISOT dataset. The OPCNN-FAKE model which is an improved version of CNN, outperforms previous approaches in identifying fake news on various datasets[28]. An efficient model was suggested utilizing BERT and BiLSTM-GRU architecture to extract spatial and temporal features while global features were extracted through TF-IDF. The model was tested on three benchmark datasets and achieved promising results[17]. A novel model with a modern strategy for positioning and hyperparameter selection known as HyproBert, developed to address the fake news detection problem. Real-world experiments on the ISOT and FA-KES datasets clearly establish the proposed model as superior to the baseline and other sophisticated methods in terms of its effectiveness in the field[29].Recent research has explored ensemble learning techniques to improve fake news classification accuracy. One notable approach utilizes an optimized weighted-voting-based ensemble that combines multiple classifiers to enhance prediction performance. The model integrates diverse classifiers, such as CNN, LSTM variants, and linguistic feature-based methods, and optimizes the voting weights using self-adaptive harmony search. This strategy effectively balances the contributions of each classifier, leading to improved accuracy and robustness on benchmark datasets. The optimized ensemble approach demonstrates that carefully weighted model fusion can significantly boost fake news detection capabilities compared to individual classifiers[30].

Despite substantial progress in fake news detection, existing methods often face challenges related to domain adaptability, feature representation, and computational efficiency. Many models require extensive labeled data or struggle to maintain performance across different social networks and event types. Moreover, balancing detection accuracy with interpretability remains an open issue. Building upon these insights, the present study proposes an enhanced framework that integrates robust feature extraction, adaptive learning strategies, and efficient model architectures to overcome these limitations. This approach aims to provide a scalable, accurate, and interpretable solution for real-world fake news detection challenges.

Proposed methodology

We provide a comprehensive explanation of our primary framework and its supporting components in this section. The architecture of the proposed model is depicted in Fig. 1.

Pre-processing

In this work, various text pre-processing strategies were exercised with a view of cleaning the data under analysis in order to obtain improved results on modeling. The first process performed was a tokenization process which entails breaking the text into individual work or tokens to enable further processing. Lower casing was then done to ensure that the words were in the same form because the visualization software takes a form that will not differentiate between small and capital letters (e.g., Data and data). Any word that does not add value to the analysis was removed by stop-word removal; these tends to include words like“the”, “is”, and “and”. Further, stemming and lemmatization was used to convert word to their base form to ensure that the algorithm

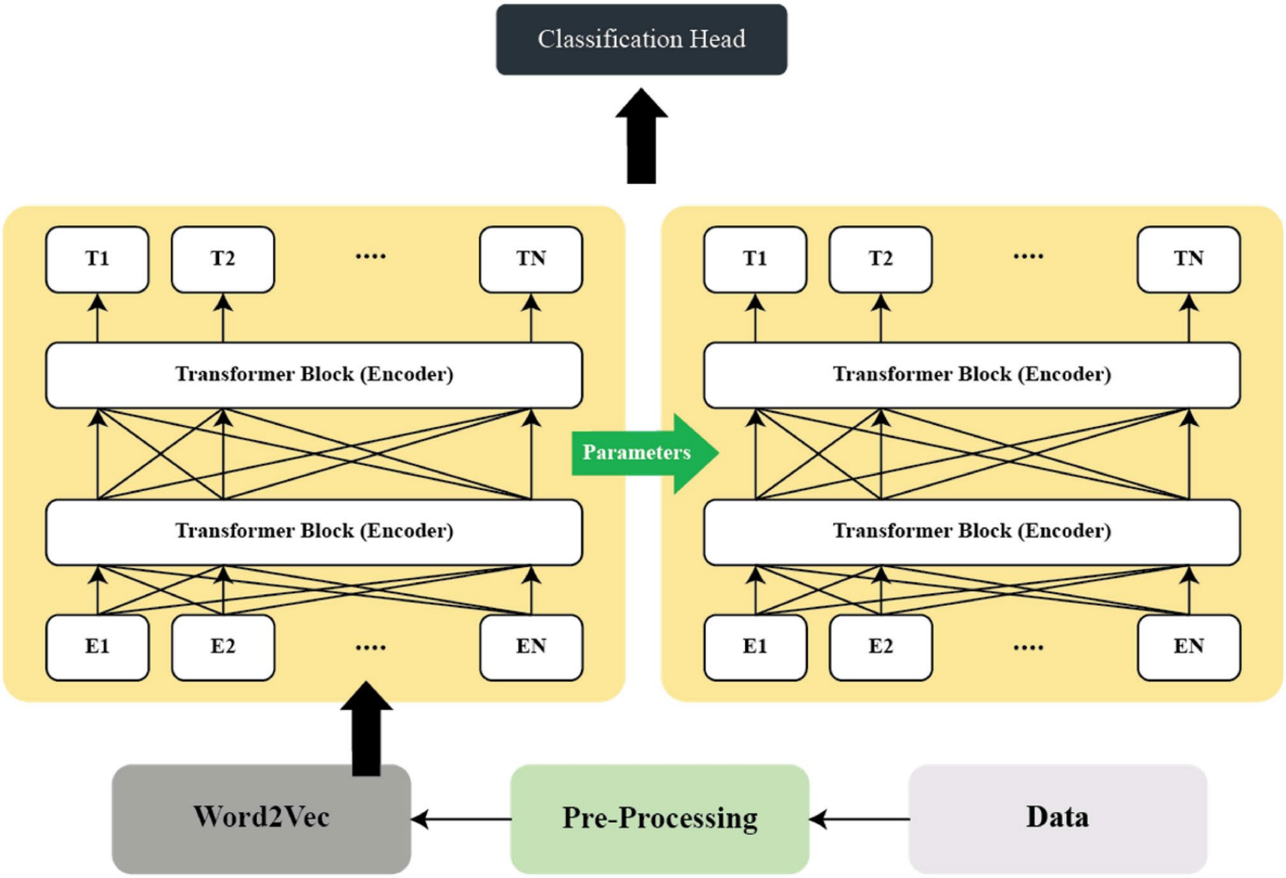
Classification Head
T1
T2
····
TN
Transformer Block (Encoder)
Transformer Block (Encoder)
E1
E2
····
EN
Parameters
T1
T2
····
TN
Transformer Block (Encoder)
Transformer Block (Encoder)
E1
E2
····
EN
Word2Vec
Pre-Processing
Data

Fig. 1. The architecture of proposed model.

distinguished between variations of the same verb such as 'running' and 'ran' to a whichever form of the word 'run'. All characters, numerals and punctuation marks were excluded to minimize on noise and any extraneous feature from the text. Additionally, Part of Speech (POS) tagging was applied which is the process of assigning one of the twelve grammatical categories to each word in a text based on the syntactic analysis. This process requires a consideration of the context in which each term is used in order to establish whether the term is a noun, verb, adjective or other component of grammar. The method can be defined mathematically as follows: For every document, t_i, obtain the POS tags for the document words to get a sequence of POS tags.

$$\mathrm{POS}(w_i) = \arg\max_{t_i} P(t_i \mid w_i) \tag{1}$$

where w_i represents words in a sentence, t_i is the POS tags assigned to w_i, and $P(t_i|w_i)$ is the probability of the tag t_i given the word w_i. By doing all this it will reduce noise and enhance the semantic quality of textual inputs. POS tagging plays a crucial role by identifying the grammatical structure of sentences, enabling the model to understand how different parts of speech (e.g., verbs, nouns, adjectives) are used in deceptive versus non-deceptive contexts. Fake news often exhibits distinctive syntactic patterns, such as exaggerated adjectives or passive constructions, which POS tagging helps capture. These steps are especially essential for small datasets, where syntactic cues become valuable features for improving classification accuracy.

Word embedding

One hot encoding

Due to the numerous characters in the English language, mapping each word using high dimension vectors can be cumbersome. A better procedure involves converting each word into vector of a certain dimensionality, which is called word embedding. Word embedding serves as a matrix transformation wherein the original English words are translated into new unique vectors. This technique is actually a feature learning process and the parameters necessary for mapping are acquired when the given model is trained.

Unlike the word segmentation approach, where text is segmented into more coherent words, in this study text will be segmented by characters. First, we will represent each word as one-hot vector $v_{n,m}$ with B dimensions, where B represents the total number of possible characters, and the element corresponding to the certain character equals 1, and for other elements 0 value is assigned. Below is a depiction of $v_{n,m}$ as defined below.

$$v_{n,m} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ B & -1 & & & \end{bmatrix} \tag{2}$$

Next, the one-hot vector $v_{n,m}$ is converted into a different representative vector $V_{n,m}$ through the following process:

$$V_{n,m} = \sigma(W_V \cdot v_{n,m} + b_V) \tag{3}$$

here, W_V and b_V represent parameters that need to be optimized, while denotes the activation function, which is defined as follows:

$$\sigma(\zeta) = \left\{ \begin{array}{ll} \zeta, & \zeta \geq 0 \\ 0, & \zeta < 0 \end{array} \right. \tag{4}$$

Equation (2) indicates that the dimension of $V_{n,m}$ is primarily dictated by the dimensions of W_V and b_V. Consequently, various vectors of $V_{n,m}$ can be converted into vectors of a designated dimension. By enumerating m from $1toM$, the representative vectors for all words in news x_i may be derived. While the encoding outcomes for individual words are satisfactory, the encoding results for each news do not merely represent a straightforward amalgamation of those words. News is influenced by complex elements, and there are underlying relationships among individual words. Specific representation models must be employed or created for the semantics of sentence-level news.

Word2Vec

Word2Vec is a common method of creating word vectors; it uses neural networks to predict the context of a word and establish the latent semantic connections between them[31]. It operates through two contrasting architectures: The two models are Continuous Bag of Words (CBOW) and Skip-Gram. Skip-Gram is an unsupervised learning framework based on word occurrence context to extract semantic meanings. It uses the log probability defined in Eq. (5) to attend to the input while its goal is to maximize the average log probability.

$$E = -\frac{1}{V}\sum_{v=1}^{V} \sum_{\substack{-c \leq m \leq c \\ m \neq 0}} \log\log \left[p(w_{v+m} \mid w_v)\right] \tag{5}$$

Using the given training data $w_1, w_2, w_3, \cdots, w_N$, the following variable *c* is a context size, which is also called the window size. The symbol *E* represents the embedding dimension. The probability $p(w_{v+m} \mid w_v)$ may be computed using Eq. (6):

$$p(o) = \frac{\exp(u_i^T \cdot u_o')}{\sum_{v \in V} \exp(u_v^T \cdot u_o')} \tag{6}$$

here, *V* stands for the vocabulary and u for the 'input' vector representation of *i* and u_o for the 'output' vector representation of *o*. CBOW predicts the target word based on the co-occurrence of surrounding words within a specific text body[32]. It takes advantage of distributed continuous contextual representations. CBOW uses a fixed window of words within a sequence of words with the middle word being predicted by a log-linear classifier that is trained on both the previous and future words in the window. The higher a value in Eq. (7), the greater the likelihood of predicting the word w_v.

$$\frac{1}{V}\sum_{v \in V} \log\left(p(w_{v-c}, \ldots, w_{v-2}, w_{v-1}, w_{v+1}, w_{v+2}, \ldots, w_{v+c})\right) \tag{7}$$

here, *V* and *c* correspond to the parameters of the Skip-Gram model. Figure 2 shows both models.

Although these models serve the same fundamental purpose of transforming words into vector representations, they capture different linguistic properties. Word2Vec focuses on capturing semantic similarity based on co-occurrence, FastText improves generalization by incorporating subword information, while BERT and DistilBERT provide contextual embeddings that understand word usage in different syntactic structures. In our framework, these embeddings are not fused together in a single model; rather, they are applied independently to identical classification pipelines to evaluate their individual contribution to model performance. This design allows us to assess which embedding approach is more robust, especially in low-resource scenarios, and helps guide future embedding selection for similar tasks. To investigate the influence of word representation on fake news detection, we systematically evaluate two word embedding methods: One-Hot Encoding and Word2Vec. These embeddings serve as input features for various ML and DL models, including RoBERTa, enabling us to assess how semantic information encoded in embeddings affects model performance. Our methodology uniquely integrates these embedding comparisons within the transfer learning framework, offering insights into the complementary roles of embeddings and transformer-based fine-tuning.

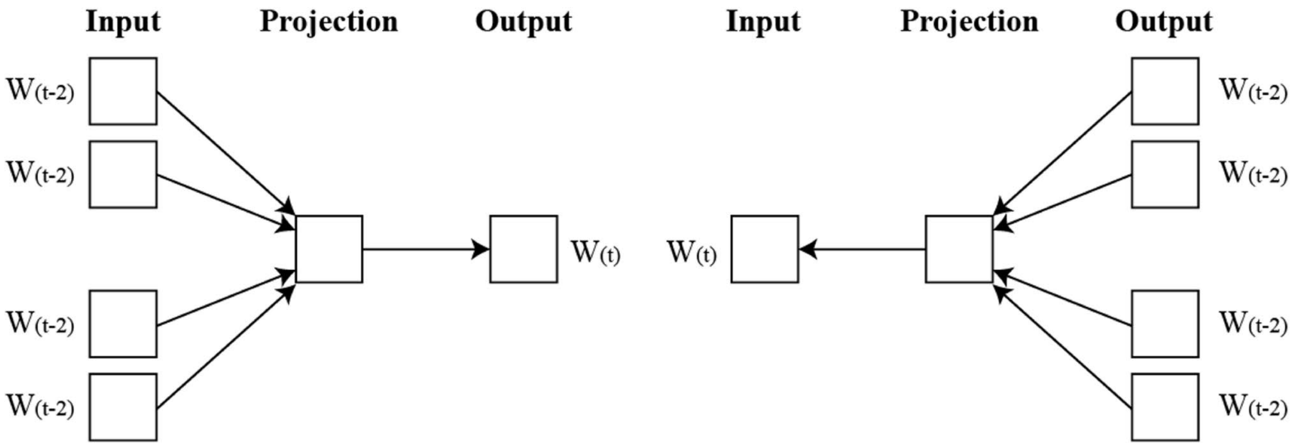

Fig. 2. Representation of word2vec model (CBOW and Skip Gram).

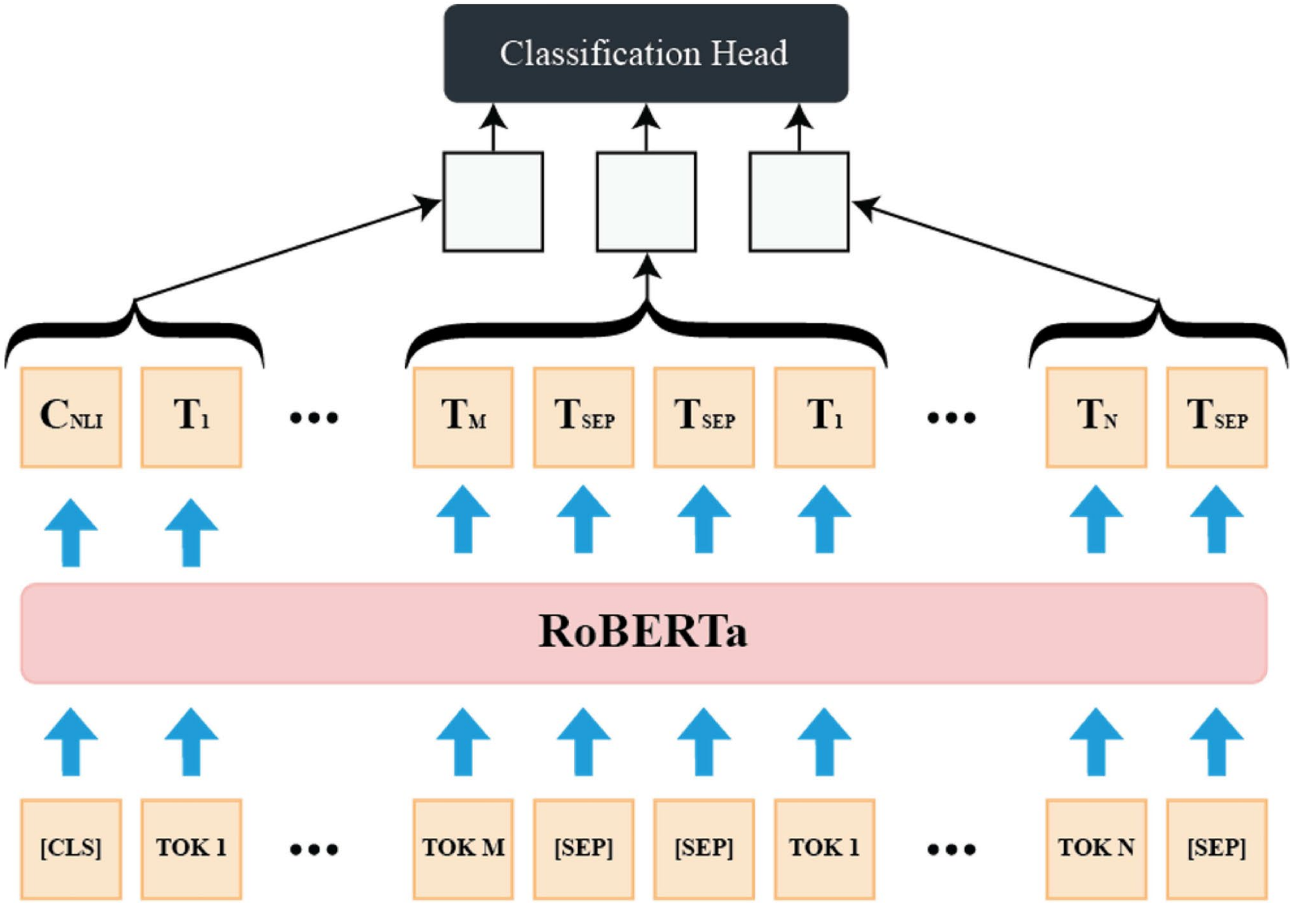

Fig. 3. Architecture of RoBERTa model.

Transfer learning using RoBERTa

Transfer learning with RoBERTa applies pre-trained language models to improve performance on specific NLP tasks by adapting previously learned knowledge. As depicted in Fig. 3, RoBERTa undergoes an initial pre-training phase on extensive corpora, allowing it to capture nuanced language patterns. This is followed by a fine-tuning phase tailored to particular tasks, where RoBERTa's pre-trained representations are optimized for task-specific requirements. Multi-stage transfer learning strategy to better adapt the pre-trained RoBERTa model for fake news detection in limited-data scenarios. Initially, the model undergoes fine-tuning on a related large corpus to learn domain-specific features. Subsequently, a second fine-tuning phase is performed on the smaller target datasets (Politifact and GossipCop), with carefully controlled learning rates and layer freezing to prevent overfitting. This two-stage process allows the model to retain generalized language understanding while gradually specializing in the fake news detection task, improving accuracy and robustness compared to single-step fine-tuning approaches.

Token and positional embeddings

RoBERTa begins by transforming the input text sequence into a series of token embeddings. Each token x_i is mapped to a high-dimensional vector representation through a learnable embedding matrix. Specifically, for each input token x_i, its corresponding embedding $E(x_i)$ is calculated as:

$$E(x_i) = W_e \cdot x_i \tag{8}$$

where x_i is a one-hot encoded representation of the token and W_e is an embedding matrix which is learnt through the training process. This transformation permits RoBERTa to connect discrete tokens to a continuous vector space, implying that the relations among words are accessible.

Moreover, since the transformer model does not capture the relative location of the tokens in a sequence, the positional encoding is added to the token vectors. These surrounding embeddings assist in keeping in touch with the location of the tokens and their significance within the sequence to assist towards the full comprehension of the input's architecture in RoBERTa. The positional embedding for each of the tokens at position i is computed as follows:

$$P(x_i) = W_p \cdot i \tag{9}$$

where W_p is a trainable embedding matrix, similar to token embeddings, and i is the token index within the sequence. In the final input representation $H_0(x_i)$ of each token x_i, such as the position embedding, embedding for each token, embedding for each token, is associated with linear transformation.

$$H_0(x_i) = E(x_i) + P(x_i) \tag{10}$$

Self-attention mechanism

RoBERTa consists of a multi-headed self-attention layer, which enables the model to evaluate the importance of each token to other tokens in the input sequence. For each token x_i, three vectors are computed: query *Q*, key *K*, and value *V*, from the token's input representation $H_0(x_i)$. These vectors are obtained through linear transformations using learnable matrices:

$$Q = H_0(x_i) \cdot W_Q \tag{11}$$

$$K = H_0(x_i) \cdot W_K \tag{12}$$

$$V = H_0(x_i) \cdot W_V \tag{13}$$

where W_Q, W_K, and W_V are learned projection matrices. The self-attention mechanism calculates the attention score between each token i and every other token j in the sequence by taking the dot product of the query and key vectors, followed by a scaling factor proportional to the square root of the dimensionality d_k:

$$A_{ij} = \frac{Q_i \cdot K_j^\top}{\sqrt{d_k}} \tag{14}$$

These attention scores indicate the level of significance that token *i* attributes to token *j*. In order to normalize the attention scores, a softmax function is utilized, which transforms the raw scores into a probability distribution:

$$a_{ij} = \mathrm{softmax}(A_{ij}) = \frac{\exp(A_{ij})}{\sum_{j=1}^{n} \exp(A_{ij})} \tag{15}$$

The weighted aggregate of all value vectors V_j, where the weights are the normalized attention scores a_{ij}, is the final output of the self-attention mechanism for token i.

$$O_i = \sum_{j=1}^{n} a_{ij} \cdot V_j \tag{16}$$

Multi-head attention and layer normalization

RoBERTa employs multi-head attention to strengthen the model's capabilities for capturing various contextual concepts encountered in the input sequence. In this form of attention, the self-attention technique is performed in parallel h number of times with each attention head having its distinct attention parameters. All the projection led outputs of attention heads concatenated and were minimal compression transformation through the projection matrix W_O.

$$O_{\text{multi-head}} = \mathrm{Concat}(O_1, O_2, \ldots, O_h) \cdot W_O \tag{17}$$

The multi-head attention mechanism enables the model to focus on multiple segments of the input sequence simultaneously, improving the understanding of each token in its context. RoBERTa further applies residual connections and layer normalizations after the multi-head attention layer to promote stability during the training process as well as convergence of the model. The output of the multi-head attention layer is also conveyed as input to the previous layer through a residual connection and thus the input is added to the layer output.

$$H_l(x_i) = \text{LayerNorm}\left(O_{\text{multi-head}}(x_i) + H_{l-1}(x_i)\right) \tag{18}$$

$H_l(x_i)$ represent the output representation of token x_i after layer *l*.

Feed-forward neural network
Every layer of the RoBERTa contains a position wise feed forward neural network (FFN), which operates on each token independently. The FFN contains two linear layers separated by a ReLU. Token outputs are computed as follows:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{19}$$

where W_1 and W_2 are the weight matrices, b_1 and b_2 are biases. In addition, the layer provides the model with non-linearity, improving the ability to process token representations.

Pre-training objective: masked language modelling (MLM)
RoBERTa is pre-trained with the masked language modelling (MLM) aim. In this setting, some of the tokens in the input sequence are retrieved at random, whereas the model must learn to predict what those hidden tokens are from the context provided by the other visible tokens. The loss function corresponding to the MLM is described as follows:

$$L_{\text{MLM}} = -\sum_{m=1}^{M} \log P(x_m \mid x_{\backslash m}) \tag{20}$$

M represents the quantity of masked tokens, while $P(x_{\backslash m})$ denotes the probability attributed to the accurate token x_m, contingent upon the unmasked tokens in the sequence $(x_{\backslash m})$. This pre-training exercise enables RoBERTa to acquire profound contextual representations of words.

Fine tuning
Fine-tuning RoBERTa involves adapting a pre-trained language model for a specific task by updating its weights using a labeled dataset, allowing the model to capture task-specific patterns while leveraging its rich pre-trained knowledge. This process typically involves adding a task-specific output layer (e.g., a classification head) and optimizing the model end-to-end with a loss function, such as cross-entropy for classification tasks. The objective during fine-tuning can be expressed as minimizing the task-specific loss $L(\theta)$ where θ represents the model parameters being updated:

$$\theta^* = \arg\min_{\theta} L(\theta) \tag{21}$$

Fine-tuning of RoBERTa was performed with a learning rate of 0.001 using the Adam optimizer and a batch size of 16. Early stopping based on validation loss was applied to avoid overfitting. We experimented with freezing initial transformer layers to evaluate the impact on model generalization, ultimately selecting the configuration yielding the best validation performance. Our methodology combines domain-tailored preprocessing, embedding technique evaluation, and a novel multi-stage transfer learning framework. This integrated design addresses the challenges of fake news detection on small datasets, leading to improved classification accuracy and robustness over conventional fine-tuning methods.

Experimental results

This section delineates the dataset description and additional implementation setup. We evaluated the performance of the proposed model against that of the other ML and DL models to assess the efficacy of the method. We analyzed the results in conjunction with recent studies to evaluate the efficacy of the proposed approach.

Dataset

Two sources of multimodal fake news datasets are utilized, namely GossipCop[33] and Politifact (https://github.com/KaiDMML/FakeNewsNet, accessed on September 10, 2024), both derived from the Twitter portal. Each dataset contains both text and images, collected from fact-checking websites that provide news articles with labels and social context information. The textual content of the news and their labels are used in this study, excluding social interactions and images.

GossipCop dataset consists of approximately 10,000 news articles, with a larger proportion of fake news samples, while Politifact contains around 500 news articles. The data are split into training, testing, and validation sets as shown in Table 1. Before training, standard preprocessing steps such as tokenization, lowercasing, removal of punctuation, and stop-word removal were applied to clean the textual data.

	Training set		Testing set		Validation set	
Datasets	Real	Fake	Real	Fake	Real	Fake
Gossip	1832	7177	615	385	204	797
Politifact	122	221	70	30	13	25

Table 1. Description of selected datasets.

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
BERT	84.47	82.51	81.98	81.44	83.11
FastText	82.36	79.86	79.43	79.07	81.98
DistilBERT	93.12	92.04	92.86	92.54	92.86
ALBERT	91.98	90.11	91.23	90.78	90.99
T5	89.76	88.43	89.01	88.57	89.56
RoBERTa	97.03	95.85	96.87	96.39	96.87

Table 2. Performance of models using Word2Vec on Politifact dataset.

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
BERT	76.98	73.33	74.88	68.45	75.62
FastText	74.81	71.23	73.79	66.67	72.33
DistilBERT	79.52	76.61	78.56	71.38	77.23
ALBERT	77.46	74.24	75.98	69.54	75.58
T5	79.98	78.03	78.21	70.32	78.12
RoBERTa	87.11	83.82	85.75	74.67	85.75

Table 3. Performance of models using one-hot encoding on Poltifact Dataset.

Experimental setup

To evaluate model effectiveness, the datasets were randomly split with 70% for training, 20% for testing, and 10% for validation, maintaining the class balance within each split. The models were implemented using the PyTorch framework in Python. All experiments were performed on a system with an Intel Xeon E5-2618L 2.3GHz CPU and an NVIDIA GeForce GTX 2080 Ti GPU with 11GB VRAM. We employed the Adam optimizer with a learning rate of 0.001 and a batch size of 32. The models were trained for 100 epochs with early stopping based on validation loss to prevent overfitting. Dropout with a rate of 0.3 was used in the network to enhance generalization. The standard cross-entropy loss function was utilized for all classification tasks.

Additionally, hyperparameters such as learning rate, batch size, and dropout rate were tuned empirically to optimize performance. Text preprocessing included tokenization, lowercasing, stop-word removal, and punctuation filtering to clean the dataset before training. These detailed implementation choices ensure reproducibility and provide clarity on the model training process.

Results on Politifact using different embedding techniques

The Table 2 displays the efficacy of various models (BERT, FastText, DistilBERT, ALBERT, T5, and RoBERTa) employing Word2Vec embeddings for the detection and classification of fake news. Word2Vec is a pre-trained word embedding method that encapsulates the semantic links among words, enabling models to comprehend word meanings in context. In this configuration, RoBERTa exhibits superior performance across all parameters, achieving an accuracy of 97.03%, precision of 95.85%, recall of 96.87%, and an F1-Score of 96.39%. The AUC 96.87% further corroborates the model's capacity to differentiate between true positives and false positives. Moreover, DistilBERT and ALBERT, also exhibit strong performance with Word2Vec embeddings. DistilBERT, a more efficient and expedited variant of BERT, attains an accuracy of 93.12%, illustrating its capability to sustain competitive performance with enhanced computational efficiency. ALBERT, recognized for its parameter efficiency, attains an accuracy of 91.98% alongside robust precision, recall, and F1-Score metrics, rendering it a suitable choice for applications necessitating high performance with reduced computational resources. T5 demonstrates moderate performance, achieving an accuracy of 89.76%, but BERT and FastText exhibit somewhat inferior accuracies of 84.47% and 82.36%, respectively.

The Table 3 presents the performance of the identical models when trained with One-Hot Encoding, a more simplistic and less contextually aware text representation technique. One-Hot Encoding depicts words as sparse vectors, disregarding their meaning links. This leads to a significant decline in performance relative to Word2Vec, as evidenced by the results across all models. RoBERTa, while remaining the highest-performing model, experiences a decline in accuracy to 87.11% (in contrast to 97.03% as shown in Table 2). This indicates that although RoBERTa is resilient, it substantially gains from more sophisticated embeddings such as Word2Vec. BERT, ALBERT, and DistilBERT similarly exhibit performance decline when subjected to One-Hot Encoding. BERT's accuracy declines to 76.98%, whilst ALBERT and DistilBERT achieve accuracies of 77.46% and 79.52%, respectively. T5 has comparable performance in both configurations, demonstrating a marginal decrease in accuracy to 79.98% with One-Hot Encoding, suggesting that although it can manage simpler representations, its efficacy is enhanced with more sophisticated embeddings such as Word2Vec. FastText, being less advanced than transformer-based models, has the lowest accuracy of 74.81% and encounters difficulties with One-Hot Encoding due to its lack of deeper contextual comprehension.

Classification report on Politifact using Word2Vec

As demonstrated in the previous section on the PolitiFact dataset RoBERTa indicates its outstanding efficacy in fake news identification. In Table 4, classification report for RoBERTa utilizing Word2Vec embeddings is illustrated. The model demonstrates strong accuracy and equitable classification scores for both the Fake and Real news categories. In the fake news classification, RoBERTa attains a precision of 93.33%, indicating that the occurrences it identified as fake news were indeed genuine cases of falsehood. The model achieves a recall of 96.55%, signifying its successful identification of 96.55% of genuine cases of bogus news. The F1-Score for the counterfeit class is 94.92%, indicating a robust equilibrium between precision and recall, and the AUC of 96.87% highlights the model's proficiency in differentiating between false and authentic news.

In the Real news class, RoBERTa exhibits superior performance, achieving a precision of 98.57% and a recall of 97.18%, indicating the model's high reliability in accurately detecting real news with minimal false positives. The F1-Score of 97.87% further validates the model's precision in accurately detecting genuine news. The AUC remains stable at 96.87%, demonstrating the model's superior discriminatory ability between the two groups. RoBERTa attains a remarkable accuracy of 97.03% on the PolitiFact dataset. The macro average for precision, recall, F1-Score, and AUC indicates a robust performance, with values of 95.%, 96.87%, 96.39%, and 96.87%, respectively. The weighted average for the identical measures is elevated, with scores of 97.05%, 97.00%, 97.01%, and 96.87%, signifying that RoBERTa exhibits consistent performance across both categories, with a minor inclination towards accurately detecting genuine news.

The confusion matrix as depicted in Fig. 5, demonstrates the model's classification accuracy in distinguishing between fake and true news. The algorithm accurately identified 28 out of 29 cases in the fake news class, resulting in a single misclassification as true news. The program correctly detected 69 out of 71 examples in the true news class, misclassifying only 2 instances as false. This indicates the model's robust performance, since most predictions are accurate with few misclassifications. Furthermore, as shown in Fig. 4, the accuracy and loss graphs demonstrate the model's convergence and high accuracy throughout the training phase, with minimum loss and constant performance increases (Fig. 5).

Table 5 indicates that RoBERTa attains the maximum accuracy of 95.90%, showcasing its remarkable classification proficiency. It also demonstrates superiority in several metrics, with precision at 95.55%, recall at 95.78%, F1-score at 95.66, and AUC at 95.78%, signifying a robust capacity to differentiate between classes. ALBERT exhibits an accuracy of 91.45%, precision of 89.52%, recall of 90.85%, resulting in an F1-Score of 89.54% and an AUC of 88.65%. DistilBERT demonstrates competitive performance with an accuracy of 89.26%, precision of 88.11%, recall of 86.46%, and a F1-score of 86.67%, whereas BERT and FastText exhibit lesser accuracies of 82.65% and 76.98%, respectively.

Furthermore, in Table 6 indicates a consistent trend in the performance metrics. RoBERTa achieves an Accuracy of 94.71%, showcasing its dependability across both encoding methods, with a precision of 93.16% and a recall of 93.54%, culminating in an F1-Score of 93.34%. DistilBERT demonstrates robust performance with an accuracy of 89.95%, precision of 88.26, and recall of 87.47%, highlighting its versatility across many encoding techniques. Nonetheless, BERT exhibits a performance reduction with an accuracy of 79.57%, whereas FastText demonstrates a diminished accuracy of 74.36%. ALBERT experiences a decline, attaining an accuracy of 86.42%, but T5 demonstrates a significant accuracy of 90.21%, reflecting good categorization ability.

Classification report on GossipCop using Word2Vec

The model's classification performance on the GossipCop dataset, employing Word2Vec embeddings for feature representation, is detailed in Table 7. RoBERTa with word2vec has excellent precision, attaining 94.03% for the fake class and 97.07% for the real class, signifying its efficacy in accurately identifying true positives. The recall values of 95.26% for Fake and 96.29% for real underscore the model's proficiency in identifying a substantial proportion of genuine cases from both categories. Moreover, F1-Score, indicating the equilibrium between precision and recall, demonstrates outstanding performance with scores of 94.64% for fake and 96.68% for real. The model's total accuracy is 95.90%, highlighting its robustness throughout the dataset. Figure 6 illustrates the

Class	Precision (%)	Recall (%)	F1-score (%)	AUC
Fake	93.33	96.55	94.92	96.87
Real	98.57	97.18	97.87	96.87
Accuracy	97.03			
Macro average	95.85	96.87	96.39	96.87
Weighted avg	97.05	97.00	97.01	96.87

Table 4. Classification report of RoBERTa using Word2Vec on Pilitifact dataset.

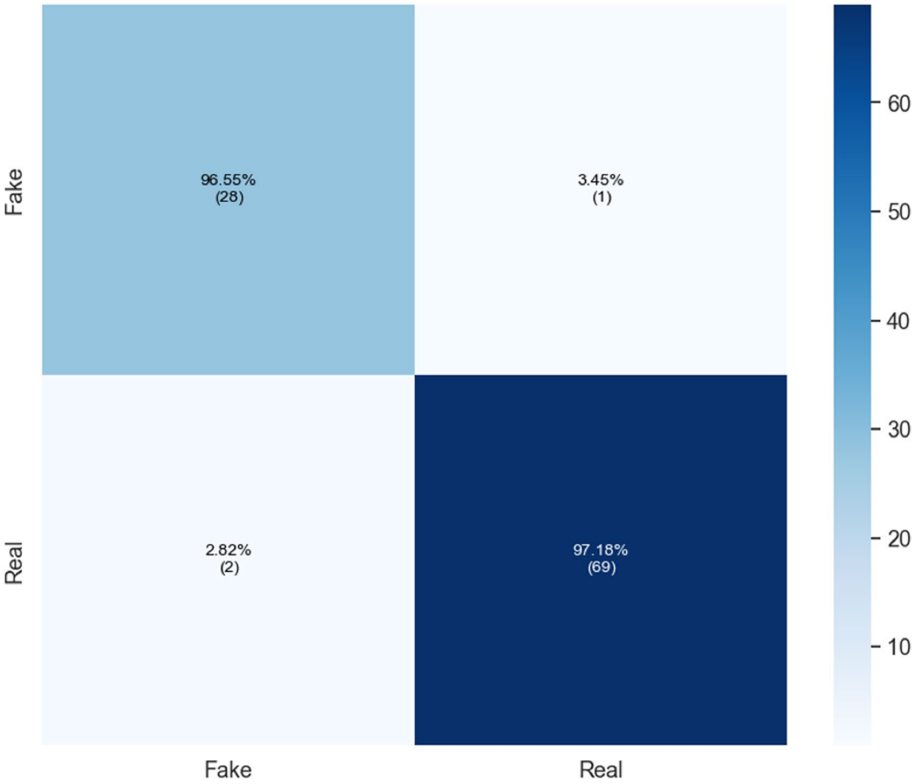

Fig. 5. Confusion matrix of proposed model on Politifact dataset.

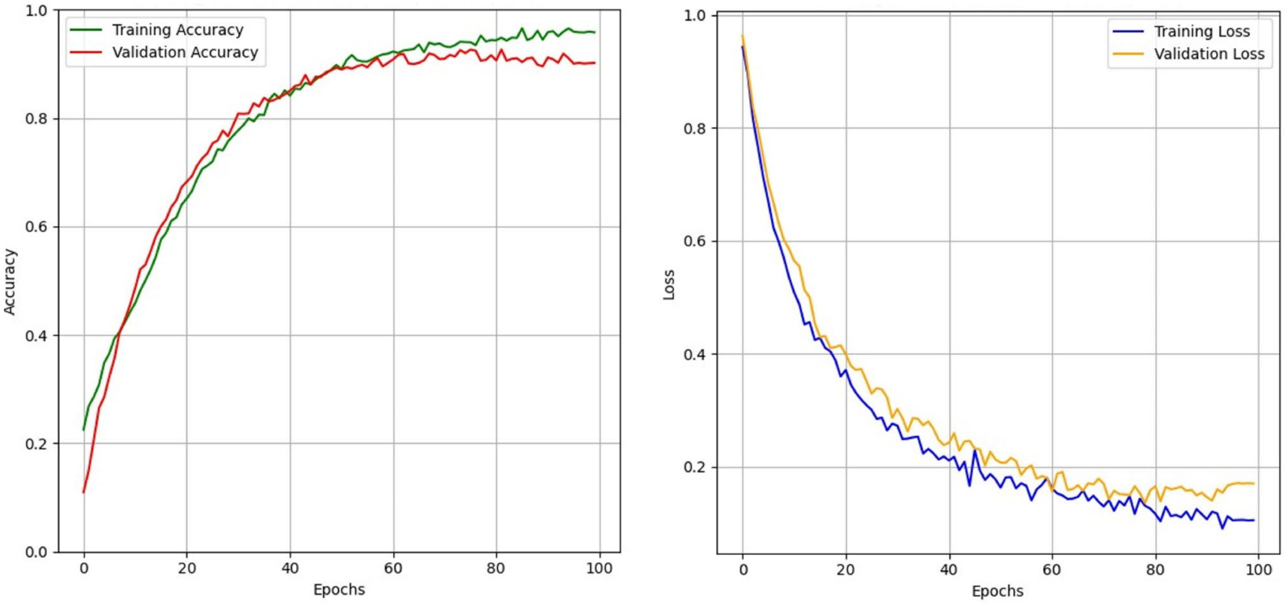

Fig. 4. Accuracy and loss graphs of proposed model for Politifact dataset.

confusion matrix of proposed model on GossipCop dataset. In addition, the accuracy and loss curves, as shown in Fig. 7, illustrate the model's high accuracy and minimal loss on GossipCop dataset.

Ablation studies

Table 8 compares the effectiveness of different embedding techniques on the Politifact and GossipCop datasets. Word2Vec embeddings consistently achieve the highest performance, with accuracies of 97.03% and 95.90% respectively, along with strong precision, recall, F1-score, and AUC metrics. One-Hot Encoding results in a

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
BERT	82.65	81.67	80.72	79.24	78.15
FastText	76.98	75.32	72.45	74.33	71.26
DistilBERT	89.26	88.11	86.46	86.67	85.21
ALBERT	91.45	89.52	90.85	89.54	88.65
T5	84.71	82.23	81.21	83.98	81.89
RoBERTa	95.90	95.55	95.78	95.66	95.78

Table 5. Performance of models using Word2Vec on GossipCop dataset.

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
BERT	79.57	76.89	78.91	77.44	77.67
FastText	74.36	72.71	73.23	71.26	71.23
DistilBERT	89.95	88.26	87.47	88.91	86.51
ALBERT	86.42	84.14	83.33	85.68	82.28
T5	90.21	88.67	89.41	88.25	87.59
RoBERTa	94.71	93.16	93.54	93.34	93.54

Table 6. Performance of models using one-hot encoding on GossipCop dataset.

Class	Precision (%)	Recall (%)	F1-Score (%)	AUC
Fake	94.03	95.26	94.64	95.78
Real	97.07	96.29	96.68	95.78
Accuracy	95.90			
Macro Average	95.55	95.78	95.66	95.78
Weighted Avg	95.92	95.90	95.91	95.78

Table 7. Classification report of RoBERTa using Word2Vec on GossipCop dataset.

significant performance drop, especially on Politifact where accuracy decreases to 87.11%, though it performs comparatively better on GossipCop with 94.71% accuracy. Models trained without any embedding perform poorly, with accuracies close to random guessing (around 54–55%), highlighting the crucial role of effective word representations in capturing semantic and contextual information necessary for accurate fake news detection. These findings demonstrate that advanced embedding methods like Word2Vec substantially improve classification outcomes across diverse datasets.

Table 9 presents the impact of layer freezing on RoBERTa's performance for fake news classification on both the Politifact and GossipCop datasets. When all layers are fine-tuned, the model achieves its highest accuracy and F1-score—97.03% and 96.39% on Politifact, and 95.90% and 95.66% on GossipCop. Freezing the bottom six layers causes a moderate decrease in performance, with accuracy dropping to 95.20% on Politifact and 93.45% on GossipCop. Further freezing of the bottom ten layers results in more significant declines, reducing accuracy to 93.81% and 91.80%, respectively. Similar trends are observed in precision, recall, and F1-score metrics across both datasets. These results indicate that while freezing lower layers can reduce computational cost, it comes at the expense of classification performance, highlighting the importance of fine-tuning deeper layers for optimal results.

Table 10 presents the comparative performance of RoBERTa models trained with a single fine-tuning stage versus the proposed multi-stage transfer learning approach on the Politifact and GossipCop datasets. On Politifact, the multi-stage method improves accuracy from 94.50 to 97.03%, with corresponding increases in precision (from 93.00 to 95.85%), recall (from 92.50 to 96.87%), and F1-score (from 92.75 to 96.39%). Similarly, for GossipCop, accuracy rises from 92.10 to 95.90%, and precision, recall, and F1-score show notable gains, emphasizing the effectiveness of multi-stage transfer learning. These results demonstrate that progressive fine-tuning across multiple stages significantly enhances the model's ability to generalize and classify fake news more accurately across diverse datasets.

Table 11 illustrates how removing individual components affects model performance on the Politifact and GossipCop datasets. The full model attains the highest accuracy and F1-score — 97.03% and 96.39% on Politifact, and 95.90% and 95.66% on GossipCop. Excluding embedding comparison causes a noticeable drop, reducing Politifact accuracy to 94.72% (a 2.31% decline) and GossipCop accuracy to 93.20% (2.70% decline). Similarly, removing layer freezing decreases accuracy to 95.33% and 93.50% on Politifact and GossipCop respectively. The most significant performance degradation occurs when adaptive learning rates are omitted, with accuracy dropping to 93.45% (3.58% loss) and 91.10% (4.80% loss) on the two datasets, alongside corresponding declines

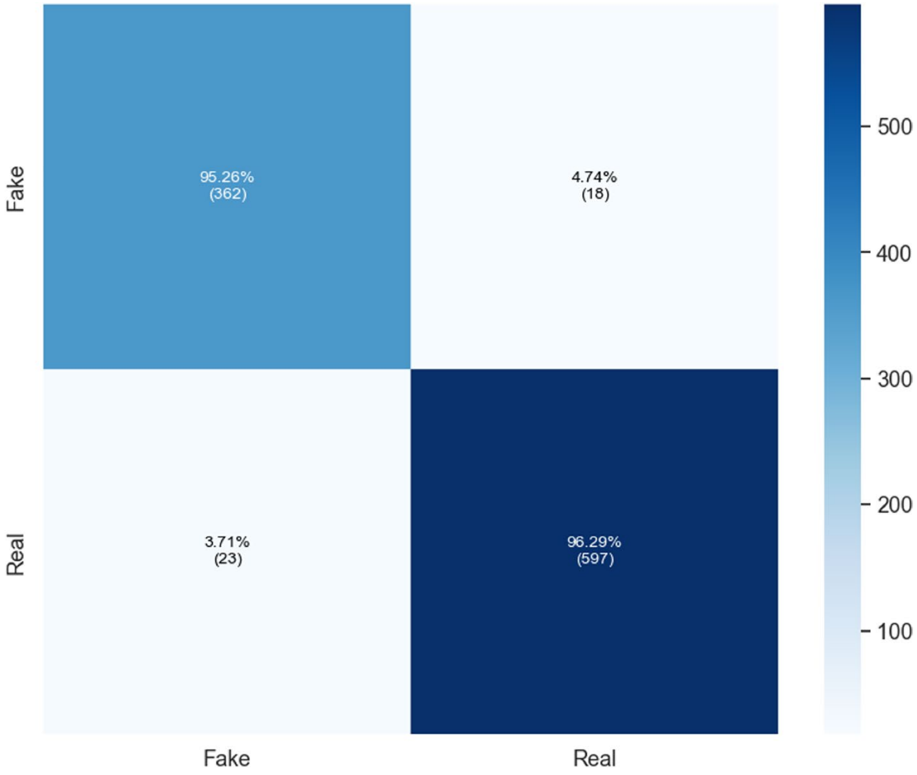

Fig. 6. Confusion matrix of proposed model on GossipCop dataset dataset.

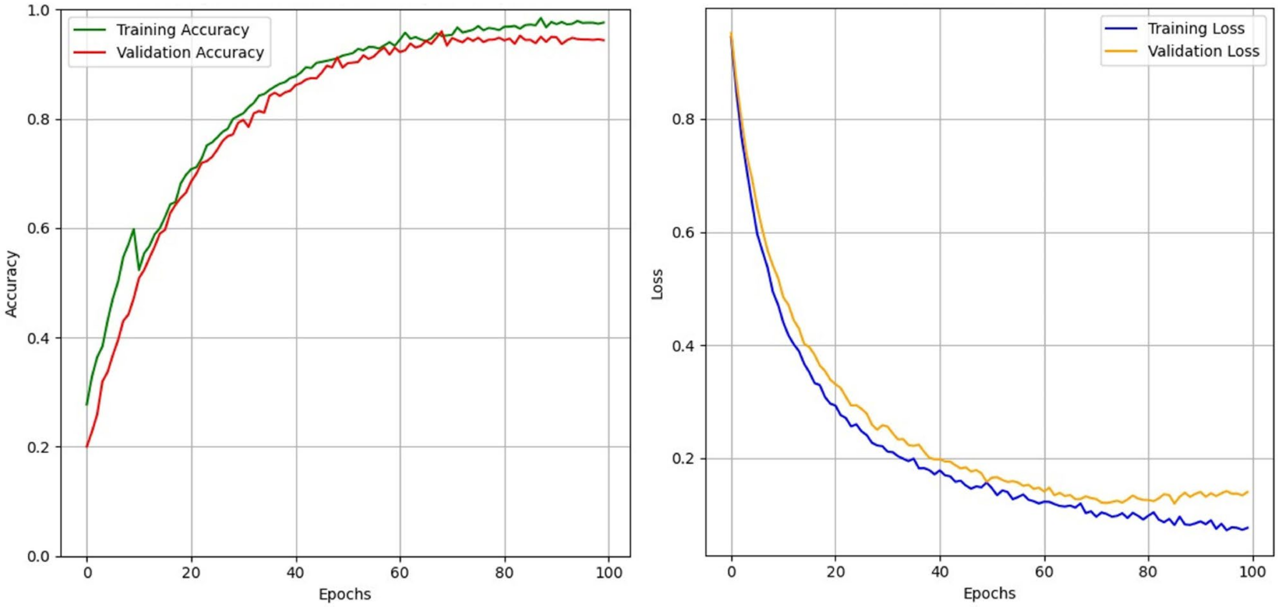

Fig. 7. Training and validation graph of proposed model for GossipCop dataset.

in F1-score. These results highlight the vital role of adaptive learning rates in model optimization and demonstrate that each component contributes substantially to maintaining high classification effectiveness.

Comparison with state of the art methods

The Table 12 illustrates comparison of proposed model with state-of-the-art methods. Compared to previous studies, the model in[34], for instance, obtained competitive accuracy and F1-scores, which included 90.4% and 92.8% on Politifact while falling short of the present proposed model. Other models such as those in[35] and[36]

Embedding technique	Politifact dataset					GossipCop dataset				
	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
Word2Vec	97.03	95.85	96.87	96.39	96.87	95.90	95.55	95.78	95.66	95.78
One-hot encoding	87.11	83.82	85.75	74.67	85.75	94.71	93.16	93.54	93.34	93.54
No embedding (Raw)	55.00	53.00	52.00	52.50	54.00	54.00	52.00	51.00	52.00	54.00

Table 8. Comprehensive performance metrics for different embedding techniques on Politifact and GossipCop datasets.

Frozen layers	Politifact dataset				GossipCop dataset			
	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
None (All tuned)	97.03	95.85	96.87	96.39	95.90	95.55	95.78	95.66
Freeze bottom 6 layers	95.20	93.12	94.23	93.67	93.45	92.12	92.89	92.40
Freeze bottom 10 layers	93.81	90.10	91.45	90.75	91.80	89.50	90.15	89.82

Table 9. Effect of layer freezing on RoBERTa performance: Politifact vs. GossipCop datasets.

Method	Politifact dataset				GossipCop dataset			
	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
RoBERTa fine-tuned once (Direct)	94.50	93.00	92.50	92.75	92.10	90.50	91.20	90.85
Multi-stage transfer learning (Proposed)	97.03	95.85	96.87	96.39	95.90	95.55	95.78	95.66

Table 10. Contribution of multi-stage transfer learning: performance comparison on Politifact and GossipCop datasets.

Configuration	Politifact			GossipCop		
	Accuracy (%)	F1-Score (%)	Drop (%)	Accuracy (%)	F1-score (%)	Drop (%)
Full model	97.03	96.39	–	95.90	95.66	–
Without embedding comparison	94.72	93.81	2.31	93.20	92.10	2.70
Without layer freezing	95.33	94.21	1.70	93.50	92.35	2.40
Without adaptive learning rates	93.45	92.05	3.58	91.10	90.00	4.80

Table 11. Component-wise ablation study: performance comparison on Politifact and GossipCop datasets.

References	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
35	Politifact	84.6	92.7	85.3	88.8
	GossipCop	87.9	89.9	95.8	92.8
37	Politifact	85.58	70.59	82.76	76.19
	GossipCop	85.43	66.19	52.37	58.47
36	Politifact	90.27	86.96	88.89	87.91
	GossipCop	75.18	74.54	76.34	75.43
34	Politifact	90.4	90.2	95.6	92.8
	GossipCop	80.8	72.9	78.2	75.5
38	Politifact	89.42	–	–	–
	GossipCop	88.62	–	–	–
39	Politifact	84.0	–	–	87.0
40	Politifact	93.91	85.19	88.46	86.79
Proposed	Politifact	97.03	95.85	96.87	96.39
	GossipCop	95.90	95.55	95.78	95.66

Table 12. Comparison with state of the art method on Politifact and GossipCop datasets.

whilst reasonable, lag significantly behind in terms of most of the metrics. Interestingly,[37] has the lowest F1-score of 58.47% in the GossipCop dataset. In general, the outcome of proposed model is superior to the previous methods in the aspects of accuracy and stability from both datasets, thus demonstrating its effectiveness in fake news detection.

Discussion

Interpretation of the main findings and comparison with state-of-the-art techniques

The results of our experiments demonstrate that the proposed multi-stage transfer learning framework significantly improves fake news detection accuracy, especially on datasets with limited labeled samples. This suggests that careful adaptation of pre-trained language models combined with embedding comparisons can effectively overcome data scarcity challenges. Our findings align with recent studies[17,29] that highlight the power of transfer learning in NLP tasks but extend them by systematically analyzing embedding techniques and fine-tuning strategies. This comprehensive approach provides deeper insights into model interpretability and performance optimization.

Implications

This study's findings have important implications for the development and deployment of fake news detection systems. By demonstrating that multi-stage transfer learning combined with embedding comparisons can improve detection accuracy on limited datasets, our approach provides a practical solution for real-world scenarios where labeled data is scarce. This can help social media platforms and fact-checking organizations implement more effective automated tools, enhancing the timeliness and reliability of misinformation identification. Moreover, the insights into embedding techniques and fine-tuning strategies offer guidance for designing adaptable models tailored to diverse data conditions, which is crucial for maintaining the integrity of information ecosystems.

Limitations and future work

However, this study has some limitations. The datasets used are limited to text-only data and relatively small sizes, which may affect the generalizability of the results to larger and multimodal datasets. Additionally, our method currently does not incorporate social network or user behavior features that could further enhance detection. Future research could focus on extending the multi-stage transfer learning approach to multimodal fake news datasets, integrating metadata such as user profiles and propagation patterns. Exploring more advanced embedding techniques or hybrid architectures may also yield further improvements.Overall, our study contributes a valuable framework and empirical insights that advance the state-of-the-art in fake news detection and open avenues for more robust and interpretable models in this critical domain.

Conclusion

This study suggested a false news stance detection model that was based on the headline and the body of the news, in contrast to previous studies that only examined individual sentences or phrases. The proposed transfer-learning framework is capable of identifying fake news with a dataset that is relatively modest in size. The proposed transfer-learning framework's efficacy is evaluated on two benchmark datasets, Politifact and GossipCop, using the f-measure, accuracy, recall, AUC, and precision in this study. The proposed transfer-learning framework achieved classification accuracy of 97.03% and 95.90% on the Politifact and GossipCop datasets, respectively, which is preferable to other comparative models. The results of this study have demonstrated that the multistage transferred learning approach can yield effective results, regardless of the small size of the dataset. The multi-stage transfer learning framework will be expanded in the future to incorporate meta-data, including user profiles, network structures, and web search engine results, to facilitate the early prediction of veracity without the need to wait for comments.

Data availability

Two publicly available datasets are used in this work, which are GossipCop [30] and Politifact (https://github.com/KaiDMML/FakeNewsNet, accessed on September 10, 2024). The implementation of this work is available at https://github.com/imashoodnasir/LLM-For-Fake-News-Detection.

Received: 15 April 2025; Accepted: 4 July 2025
Published online: 05 August 2025

References

1. Aker, A., Sliwa, A., Dalvi, F. & Bontcheva, K. Rumour verification through recurring information and an inner-attention mechanism. *Online Soc. Netw. Med.* **13**, 100045 (2019).
2. Singh, J. P., Kumar, A., Rana, N. P. & Dwivedi, Y. K. Attention-based lstm network for rumor veracity estimation of tweets. *Inf. Syst. Front.* **24**, 1–16 (2022).
3. Guo, M., Xu, Z., Liu, L., Guo, M. & Zhang, Y. An adaptive deep transfer learning model for rumor detection without sufficient identified rumors. *Math. Probl. Eng.* **2020**, 7562567 (2020).
4. Lv, Q., Wang, Y., Zhang, B. & Jin, Q. RV-ML: An effective rumor verification scheme based on multi-task learning model. *IEEE Commun. Lett.* **24**, 2527–2531 (2020).
5. Bai, N., Wang, Z. & Meng, F. A stochastic attention CNN model for rumor stance classification. *IEEE Access* **8**, 80771–80778 (2020).
6. Yousafzai, S. N. et al. X-news dataset for online news categorization. *Int. J. Intell. Comput. Cybern.* **17**, 737 (2024).
7. kumari Mukiri, R. & Babu, B.V. Withdrawn: Prediction of rumour source identification through spam detection on social networks-a survey (2021).

8. Bondielli, A. & Marcelloni, F. A survey on fake news and rumour detection techniques. *Inf. Sci.* **497**, 38–55 (2019).
9. Sadr, H. et al. Enhancing brain tumor classification in MRI images: A deep learning-based approach for accurate classification and diagnosis. *Image Vis. Comput.* https://doi.org/10.1016/j.imavis.2025.105555 (2025).
10. Castillo, C., Mendoza, M. & Poblete, B. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, 675–684 (2011).
11. Popat, K. Assessing the credibility of claims on the web. In *Proceedings of the 26th International Conference On World Wide Web Companion*, 735–739 (2017).
12. Nazari, M. et al. Design and analysis of a telemonitoring system for high-risk pregnant women in need of special care or attention. *BMC Pregnancy Childbirth* **24**, 817. https://doi.org/10.1186/s12884-024-07019-4 (2024).
13. Yang, F., Liu, Y., Yu, X. & Yang, M. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, 1–7 (2012).
14. Sadr, H., Khodaverdian, Z., Nazari, M. & Yamaghani, M. R. A shallow convolutional neural network for cerebral neoplasm detection from magnetic resonance imaging. *Big Data Comput. Vis.* **4**, 95–109. https://doi.org/10.22105/bdcv.2024.474574.1182 (2024).
15. Liu, Y. & Wu, Y.-F. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32 (2018).
16. Ali, A., Shahbaz, H. & Damaševičius, R. Xcvit: Improved vision transformer network with fusion of CNN and Xception for skin disease recognition with explainable AI. *Comput. Mater. Contin.* **83**, 1367 (2025).
17. Alzaidi, M. S. A. et al. An efficient fusion network for fake news classification. *Mathematics* **12**, 3294 (2024).
18. Song, C. et al. CED: Credible early detection of social media rumors. *IEEE Trans. Knowl. Data Eng.* **33**, 3035–3047 (2019).
19. Yuan, H., Zheng, J., Ye, Q., Qian, Y. & Zhang, Y. Improving fake news detection with domain-adversarial and graph-attention neural network. *Decis. Support Syst.* **151**, 113633 (2021).
20. Nasir, J. A., Khan, O. S. & Varlamis, I. Fake news detection: A hybrid CNN-RNN based deep learning approach. *Int. J. Inf. Manag. Data Insights* **1**, 100007 (2021).
21. Sahoo, S. R. & Gupta, B. B. Multiple features based approach for automatic fake news detection on social networks using deep learning. *Appl. Soft Comput.* **100**, 106983 (2021).
22. Huang, Y.-F. & Chen, P.-H. Fake news detection using an ensemble learning model based on self-adaptive harmony search algorithms. *Expert Syst. Appl.* **159**, 113584 (2020).
23. Ozbay, F. A. & Alatas, B. Fake news detection within online social media using supervised artificial intelligence algorithms. *Phys. A Stat. Mech. Appl.* **540**, 123174 (2020).
24. Wang, Y., Wang, L., Yang, Y. & Zhang, Y. Detecting fake news by enhanced text representation with multi-edu-structure awareness. *Expert Syst. Appl.* **206**, 117781 (2022).
25. Zhou, Y., Yang, Y., Ying, Q., Qian, Z. & Zhang, X. Multimodal fake news detection via clip-guided learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2825–2830 (IEEE, 2023).
26. Tokpa, F. W. R., Kamagaté, B. H., Monsan, V. & Oumtanaga, S. Fake news detection in social media: Hybrid deep learning approaches. *J. Adv. Inf. Technol.* **14**, 606 (2023).
27. Elsaeed, E., Ouda, O., Elmogy, M. M., Atwan, A. & El-Daydamony, E. Detecting fake news in social media using voting classifier. *IEEE Access* **9**, 161909–161925 (2021).
28. Saleh, H., Alharbi, A. & Alsamhi, S. H. Opcnn-fake: Optimized convolutional neural network for fake news detection. *IEEE Access* **9**, 129471–129489 (2021).
29. Nadeem, M. I. et al. Hyprobert: A fake news detection model based on deep hypercontext. *Symmetry* **15**, 296 (2023).
30. Toor, M. S. et al. An optimized weighted-voting-based ensemble learning approach for fake news classification. *Mathematics* **13**, 449 (2025).
31. Grohe, M. word2vec, node2vec, graph2vec, x2vec: Towards a theory of vector embeddings of structured data. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 1–16 (2020).
32. Choudhari, P. & Veenadhari, S. Sentiment classification of online mobile reviews using combination of word2vec and bag-of-centroids. In *Machine Learning and Information Processing: Proceedings of ICMLIP 2019*, 69–80 (Springer, 2020).
33. Shu, K., Mahudeswaran, D., Wang, S., Lee, D. & Liu, H. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data* **8**, 171–188 (2020).
34. Shu, K., Cui, L., Wang, S., Lee, D. & Liu, H. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 395–405 (2019).
35. Qu, Z., Meng, Y., Muhammad, G. & Tiwari, P. QMFND: A quantum multimodal fusion-based fake news detection model for social media. *Inf. Fusion* **104**, 102172 (2024).
36. Choudhary, M., Chouhan, S. S., Pilli, E. S. & Vipparthi, S. K. Berconvonet: A deep learning framework for fake news classification. *Appl. Soft Comput.* **110**, 107614 (2021).
37. Al Obaid, A., Khotanlou, H., Mansoorizadeh, M. & Zabihzadeh, D. Multimodal fake-news recognition using ensemble of deep learners. *Entropy* **24**, 1242 (2022).
38. Campus, P. Fakeexpose: Uncovering the falsity of news by targeting the multimodality via transfer learning. *J. Inf. Optim. Sci.* **44**, 301–314 (2023).
39. Shishah, W. Fake news detection using BERT model with joint learning. *Arab. J. Sci. Eng.* **46**, 9115–9127 (2021).
40. Lai, J. et al. RumorLLM: A rumor large language model-based fake-news-detection data-augmentation approach. *Appl. Sci.* **14**, 532 (2024).

Acknowledgements

The Researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2025). The authors would like to acknowledge the support of Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R435), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author contributions

B.S.A. and S.A.A. conceived the idea. B.S.A. developed the theoretical framework and carried out the computations. S.N.Y. and S.A. verified the analytical methods and validated the results. D.M.A. supervised the project, encouraged B.S.A. to explore the specific aspects, and contributed to the interpretation of the results. D.S.A. contributed to software development, visualization, and formal analysis. All authors contributed equally to manuscript preparation, review, editing, and approved the final version.

Funding

This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R435), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.M.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.