scientific reports

OPEN

A fake news detection model using the integration of multimodal attention mechanism and residual convolutional network

Ying Lu[1]✉ & Naiwei Yao[2]

To improve the accuracy and efficiency of fake news detection, this study proposes a deep learning model that integrates residual networks with attention mechanisms. Building on traditional convolutional neural networks, the model incorporates multi-head attention mechanisms to enhance the extraction of key features from multimodal data such as text, images, and videos. Additionally, residual connections are introduced to deepen the network architecture, mitigate the vanishing gradient problem, and improve the model's learning depth and stability. Compared with existing approaches, this study introduces several key innovations. First, it constructs a multimodal feature fusion module that integrates text, image, and video data. Second, it designs a cross-modal alignment mechanism to better connect information across different data types. Third, it optimizes the feature fusion structure for more effective integration. Finally, the study employs attention mechanisms to highlight and enhance the representation of salient features. Experiments were conducted using three representative datasets: the LIAR dataset for political short texts, the FakeNewsNet dataset for English multimodal news, and the Weibo dataset from a Chinese social media platform. These were selected to comprehensively evaluate the model's performance across different scenarios. Baseline models used for comparison include Bidirectional Encoder Representations from Transformers (BERT), Robustly Optimized Bidirectional Encoder Representations from Transformers Approach (RoBERTa), Generalized Autoregressive Pretraining for Language Understanding (XLNet), Enhanced Representation through Knowledge Integration (ERNIE), and Generative Pre-trained Transformer 3.5 (GPT-3.5). In terms of four key performance metrics—accuracy, precision, recall, and F1 score—the proposed model achieved best-case values of 0.977, 0.986, 0.969, and 0.924, respectively, outperforming the aforementioned baseline models overall. Furthermore, simulated experiments were conducted to evaluate the model's real-world applicability from four dimensions: robustness, generalization ability, response time, and resource consumption. The results demonstrate that the model maintains strong stability and adaptability under data perturbations and diverse input conditions, with a response time controllable within 0.02 s. The model also shows significant computational advantages when handling large-scale datasets. Therefore, this study presents a high-performance and deployment-friendly solution for fake news detection in multimodal contexts. The study also offers valuable theoretical insights and practical guidance for applying deep learning to public opinion governance and text classification.

Keywords Fake news detection, Residual attention networks, Deep learning, Performance evaluation, Resource consumption

In today's era of information explosion, the internet has become the primary channel for the public to access information[1]. However, the rapid spread of fake news has caused significant harm to society. It not only misguides public perception but also has the potential to trigger social panic, which can have a profound impact on the country's political stability, economic operations, and international public opinion environment[2,3]. Therefore, how to efficiently and accurately identify fake news has become a critical social issue that needs to be addressed. In recent years, with the rapid development of deep learning technologies, natural language

[1]Xi'an International University, Xi'an City 710000, China. [2]Northwest Electric Power Design Institute Co., Ltd., Xi'an City 710000, China. ✉email: 245972639@qq.com

processing (NLP) and text classification techniques have become mainstream methods for fake news detection. Traditional convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have shown good performance in feature extraction and sequence modeling. However, they still face significant limitations when handling complex language structures, capturing contextual relationships, and semantic features in long texts, especially when dealing with large-scale and diverse data. In such cases, the model's generalization ability and semantic understanding capability need to be further improved[4,5]. To address these issues, the introduction of attention mechanisms has become a breakthrough in optimizing deep learning models. This mechanism enables the model to focus on key information during text processing, significantly enhancing the model's ability to judge and represent feature importance. However, the simple combination of attention mechanisms with traditional CNNs still faces problems such as information loss and feature redundancy when handling long texts and complex semantic structures. The model's depth and ability to aggregate information are also limited.

Therefore, this study proposes a fusion of the deep feature propagation capability of Residual Network (ResNet) with the multi-head attention mechanism. This fusion constructs a fake news detection model with strong deep learning capability, semantic sensitivity, and adaptability to multimodal data structures. Compared with existing studies, the innovations in this study lie in the construction of a residual attention detection structure. This structure integrates multimodal features, enabling it to handle text, image, and video data simultaneously. Additionally, the design of feature alignment and weighted fusion mechanisms enhances the collaborative expression of cross-modal features. The feature extraction and residual propagation paths are also optimized, strengthening the model's deep learning ability and resistance to interference.

Based on the above research design, this study focuses on the following three specific research questions:

(1) How does the attention mechanism enhance the performance of ResNet in text feature extraction, especially in terms of handling long texts and complex semantic structures?
(2) How does the proposed multimodal feature alignment and fusion strategy significantly improve the accuracy and stability of fake news detection after integrating text, image, and video information?
(3) Does the model show significant advantages in robustness (resistance to interference) and generalization across different datasets and real-world application scenarios, and is it feasible for deployment?

To answer these questions, this study designs and implements a fake news detection model based on a multimodal residual attention mechanism. Comparative experiments and simulation evaluations are conducted on multiple public datasets to validate its performance advantages and application potential. The goal of this study is to provide an efficient, reliable, and multi-scenario intelligent solution for fake news detection. This solution aims to contribute to the purification of the public opinion environment and the healthy development of the social information ecosystem.

Literature review

The proliferation of fake news poses a mounting challenge, prompting researchers to devise various technological strategies to counteract this issue. The current research mainly focuses on text feature extraction, data fusion, and multi-modal information processing, using a variety of datasets and methods. For example, Hamed et al.[6] proposed that CNNs could effectively extract local features of text data, thereby improving the accuracy of fake news detection. They used the Liar Dataset and compared the CNN model with different depths to verify its superior performance in processing news text[6]. Guo et al.[7] introduced a fake news detection model based on a long short-term memory (LSTM), which could capture the remote dependency relationship in text and solve the limitations of traditional machine learning methods in processing long text. They trained and tested on the FakeNewsNet dataset, showing that the LSTM model was excellent at recognizing complex text patterns[7]. Qu et al.[8] used the Bidirectional Encoder Representations from Transformers (BERT) model to classify fake news. The advantages of pre-trained language models in understanding context and capturing subtle semantic changes were emphasized. Experiments on the Kaggle Fake News dataset showed that BERT performed well in fake news detection[8]. Nadeem et al.[9] presented a multi-modal fake news detection model, combining text and image information. Robustly Optimized Bidirectional Encoder Representations from Transformers Approach (RoBERTa) and ResNet methods were combined. Training on Microsoft Common Objects in Context (COCO) and FakeNewsNet datasets denoted that multi-modal information fusion could improve detection accuracy[9]. Al-tai et al.[10] discussed the application of an Enhanced Representation through Knowledge Integration (ERNIE) model in fake news detection, which combined external knowledge graphs to enhance text feature representation. Experiments on the PolitiFact dataset indicated that the ERNIE model performed particularly well in dealing with subtle semantic fake news[10].

Khan et al.[11] proposed the Deep Dual Patch Attention Mechanism (D2PAM) model. Their study proposed that by introducing an adversarial training strategy and a dual-patch attention mechanism, the model could capture local key signals in temporal data and improve the accuracy of epilepsy seizure prediction. The design of this local attention block provided a paradigm for high-robustness feature representation in time-series tasks[11]. Subsequently, Khan et al.[12] further proposed the Dual 3D Mixed Multi-Transformer with Alzheimer's Diagnosis (Dual-3DM3-AD) model. They pointed out that this model achieved significant results in early multi-class Alzheimer's diagnosis by combining a mixed Transformer structure with a triplet preprocessing mechanism. The experiment emphasized that the attention mechanism not only enhanced sensitivity to brain structure information but also optimized the consistency of semantic features in spatial distribution, which was critical for the fine segmentation of medical images[12]. In addition, Perumal et al.[13] proposed the Triple Multi-Modality Multi-Transformer (Tri-M2MT) model. Their study showed that the method, which used multimodal data inputs from neonatal magnetic resonance imaging and integrates multiple Transformer structures, effectively improved the diagnostic accuracy for acute bilirubin encephalopathy. Their study highlighted that the multi-

Architecture	Adjustment
The convolutional layer	This layer uses multiple filters (kernels) of different sizes to extract local features by sliding on the text matrix. For example, a 3 × N filter can cover all embedding dimensions of three words and can capture the association between words, similar to the n-gram model
The activation layer	Convolutional layers are usually followed by an activation layer, such as ReLU, which adds nonlinear characteristics to the network and helps capture complex patterns[17]
The pooling layer	This layer (usually max pooling) in text processing is used to reduce feature dimensions and extract the most significant features, which helps reduce the need for computational resources and improves the model's generalization ability

Table 1. Adaptation of CNNs architecture.

Dimension	Challenges
Context dependency	Traditional CNN models are less effective than RNNs or Transformer series models in capturing long-distance dependencies[20]
Hyperparameter tuning	The choice of convolutional kernel size, quantity, and network depth directly affects the model's performance, requiring extensive experimentation to determine the optimal configuration
Data insufficiency and overfitting	CNN models may require massive training data to achieve optimal performance. For tasks with small datasets, the model may face overfitting issues

Table 2. Model optimization and challenges.

path attention mechanism played a positive role in filtering redundant information between modalities and focusing on key regions[13]. These studies further validate the advantages of attention mechanisms in multimodal, high-dimensional data, particularly in discriminating complex structures and feature fusion.

Despite the remarkable progress in fake news detection, some research gaps and challenges still exist. First, most studies focus on single-modal data processing, ignoring the comprehensive utilization of multi-modal information. Second, existing models still face challenges in handling long-distance dependencies and subtle semantic changes. Moreover, these models' robustness and generalization ability in practical applications are also insufficient. The innovation of this study is that a new model of fake news detection integrating attention mechanism and residual convolutional network is proposed. This model can extract text features effectively and improve the detection accuracy through multi-modal information fusion technology. Experiments on multiple datasets verify the model's superior performance in accuracy, precision, recall, and F1 score, offering new research ideas and application prospects for fake news detection.

Optimization of fake news detection models

Application of convolutional neural networks in text processing

CNNs, as a deep learning architecture, are widely used in the field of image processing and are highly favored due to their powerful feature extraction capabilities. In recent years, CNNs have also made breakthrough progress in text-processing tasks in NLP. Their applications in text processing cover multiple tasks such as text classification, sentiment analysis, and topic recognition, mainly by capturing local features in the text to parse and process language data[14–16].

In the application of text processing, the architecture of CNNs needs to be appropriately adjusted according to the characteristics of the text data. Traditional CNNs used for images typically include convolutional, pooling, and fully connected layers. When processing text data, the input is usually word vectors or character vectors, organized into a two-dimensional data structure similar to an image, where each row represents an embedded vector of a word or character. This adaptation of CNN architecture for processing text data is illustrated in Table 1.

In text classification tasks, CNNs effectively categorize articles, comments, social media posts, etc., by learning local features from the text. For instance, CNNs can recognize key phrases or sentence structures that influence the sentiment orientation of comments, and classify sentiment based on these features. After multiple layers of convolution and pooling, CNNs generate a global feature representation that can be used for classification[18,19]. Convolutional kernels of different sizes can capture dependency relationships of varying lengths, from individual words to phrases, and even entire sentences. Despite CNNs' excellent performance in text processing, there are also some challenges and areas for optimization, as outlined in Table 2.

In practical applications, CNNs are used in various fields such as news classification, sentiment analysis of user comments, social media monitoring, etc. By deploying CNN-based models, enterprises and research institutions can automatically process and analyze large volumes of text data, thereby improving efficiency and decision quality[21].

CNNs provide an effective method for text-processing tasks due to their powerful feature extraction capabilities. With technological advancements and deeper applications, the role of CNNs in the NLP field is expected to become increasingly important[22].

Structure and characteristics of residual attention networks

Residual attention networks blend the robust feature propagation capabilities of the deep ResNet with the information filtering functions of attention mechanisms, creating an efficient network structure to enhance the learning efficiency and accuracy of complex data features. This section offers an in-depth exploration of the architecture, characteristics, and unique attributes of this network upon integration. The basic architecture of residual attention networks includes several core components, as shown in Table 3.

In each residual unit, this study embeds the attention module between two convolutional layers. The module is designed to dynamically adjust the weights of different features. Before information enters the next convolutional computation, it prioritizes the key information that contributes to the judgment, suppressing noise and redundant features, thereby enhancing the effectiveness of the feature representation[23,24]. The attention module can be deployed at either the input or output end of the residual unit. It can perform pre-filtering on the input data or post-enhancement on the output results. This helps the model better understand the global semantic structure[25]. By introducing residual connections, the model can bypass nonlinear transformations during deep network training, directly passing the original features to deeper layers. This effectively alleviates the vanishing gradient problem and ensures that the model maintains stable learning ability as the network deepens. The combination of attention mechanisms and residual connections allows the model to both "retain complete information" and "highlight key features." For example, in the BERT model, the attention mechanism is used to model the dependency relationships between words in a sentence, thus improving semantic understanding. In ResNet, the residual connection allows the input to bypass convolutional layers and directly add to the output, effectively improving training depth and model convergence efficiency. The model proposed in this study combines both approaches. First, it retains the original feature information through the residual connection in each feature extraction unit. Then, it enhances key positional features using the attention mechanism. This achieves dual optimization for both information transmission and semantic filtering.

In summary, residual connections provide a stable information channel. This ensures that features are not lost in deep networks. The attention mechanism, on the other hand, prioritizes features. It allows the model to focus more on content that is crucial for the judgment result. The synergistic effect of both not only enhances the model's understanding of complex data but also significantly improves its prediction performance and generalization ability in practical scenarios.

Design of fake news detection model

Designing an effective fake news detection model involves several key steps, including data preprocessing, feature extraction, model construction, loss function selection, and optimized algorithm application in the training process. The optimization strategy is illustrated in Fig. 1.

In the data preprocessing phase, text cleaning operations are performed to remove noisy data, such as extra spaces, punctuation, and special characters, thereby standardizing the text. Tokenization and stemming are then applied to split the text into words or lexical units, converting them into their base forms to reduce the vocabulary size. Simultaneously, stopwords are removed to highlight key information. In feature extraction, techniques such as word vector representation, part-of-speech tagging, and named entity recognition are used to capture the semantic and syntactic structure of the text. For multimodal data, features such as color, texture, and shape are extracted from images; Inter-frame differences and motion information are extracted from videos, all of which are then transformed into formats suitable for model input. The model is constructed based on residual attention networks and CNNs. The network hierarchy, convolutional kernel size, number of layers, and other parameters are carefully designed to determine the location and type of attention mechanism, and give full play to the advantages of residual attention network and attention mechanism. Meanwhile, the model parameters are reasonably initialized to provide a good start for training. The loss function is selected according to the characteristics of the fake news detection task. For example, the cross-entropy loss function measures the difference between the model prediction and the real label, determining the optimization goal. Finally, the optimization algorithm is utilized to update the model parameters according to the gradient information of the loss function. This promotes the model to converge to the optimal solution and improves the model's accuracy and generalization ability, to effectively process and analyze the text data and accurately identify the fake news content. The optimized model's architecture is illustrated in Fig. 2.

It can be found the input layer receives preprocessed text data and converts each word into a pre-trained word vector representation using word embedding techniques, providing the foundation for subsequent network processing. The data then enters the convolutional layers, where the model uses multiple convolution kernels of varying sizes to capture local dependencies and features of the text at different granularities, extracting rich contextual information. The output of each convolutional layer is passed through a ReLU activation function, introducing a non-linear transformation that further enhances the network's feature learning ability. Based on the features extracted from the convolutional layers, the model incorporates a multi-head self-attention mechanism to improve the model's focus on critical information within the text. The self-attention mechanism captures long-range dependencies between different positions in the text, optimizing the expression of features. To address the

Component	Description
Residual unit	The residual unit is the basic building block of the ResNet. Each residual unit consists of two or more convolutional layers and a skip connection, which allows the input to be directly connected to the output of the next layer. This design helps address the problem of vanishing gradients in deep neural networks, allowing the network to deepen without losing input information
Attention module	Attention modules are typically inserted between or within residual units to weight important features and suppress unimportant information. These modules use soft attention mechanisms, such as multi-head attention based on Scaled Dot-Product Attention, to process data in parallel and focus on different representation subspaces
Fusion mechanism	At the input or output of residual units, the fusion of attention-weighted and original input features enhances the model's comprehensive processing capabilities of information. Depending on specific application requirements, this fusion is usually achieved through addition or concatenation

Table 3. Core components of residual attention networks.

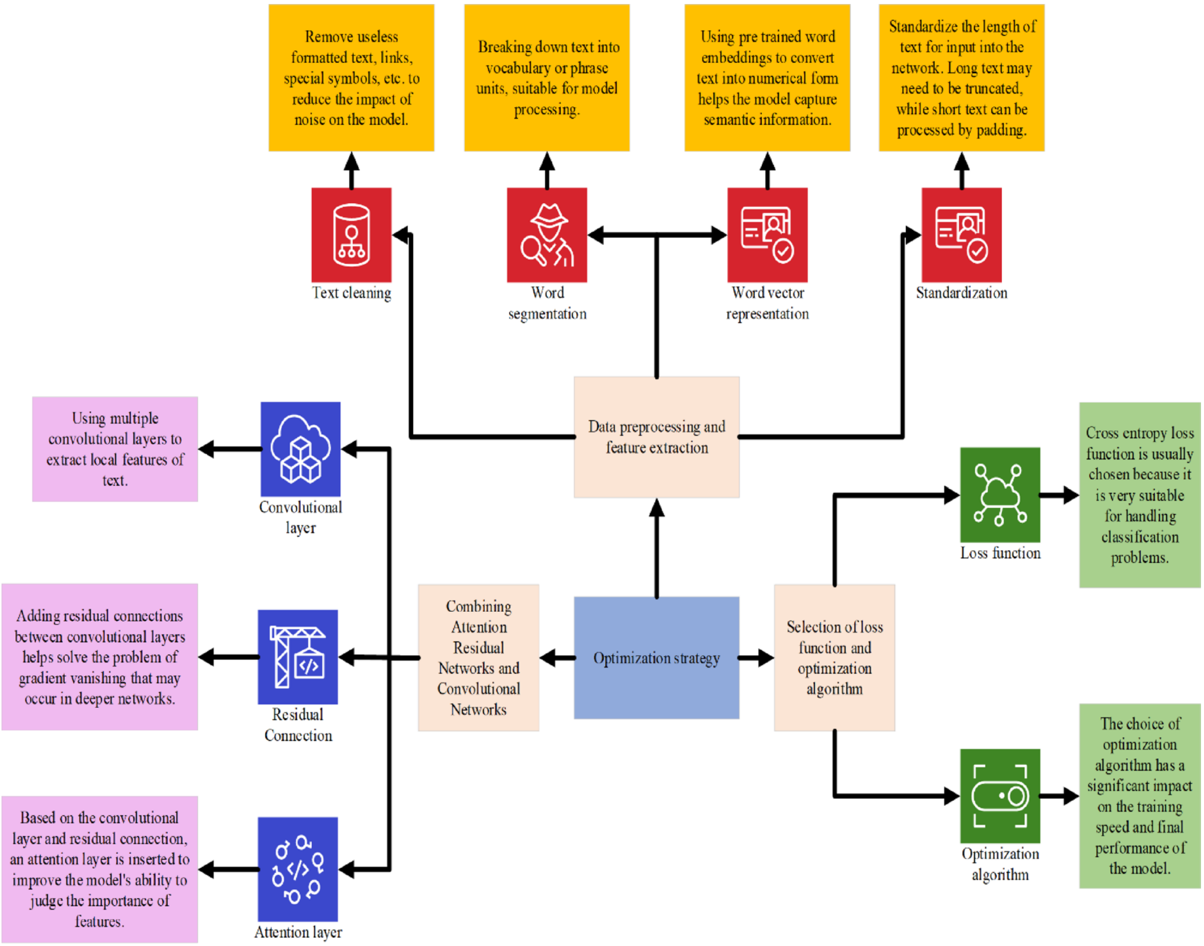

Remove useless formatted text, links, special symbols, etc. to reduce the impact of noise on the model.
Breaking down text into vocabulary or phrase units, suitable for model processing.
Using pre trained word embeddings to convert text into numerical form helps the model capture semantic information.
Standardize the length of text for input into the network. Long text may need to be truncated, while short text can be processed by padding.
Text cleaning
Word segmentation
Word vector representation
Standardization
Using multiple convolutional layers to extract local features of text.
Convolutional layer
Data preprocessing and feature extraction
Cross entropy loss function is usually chosen because it is very suitable for handling classification problems.
Loss function
Adding residual connections between convolutional layers helps solve the problem of gradient vanishing that may occur in deeper networks.
Residual Connection
Combining Attention Residual Networks and Convolutional Networks
Optimization strategy
Selection of loss function and optimization algorithm
The choice of optimization algorithm has a significant impact on the training speed and final performance of the model.
Optimization algorithm
Based on the convolutional layer and residual connection, an attention layer is inserted to improve the model's ability to judge the importance of features.
Attention layer

Fig. 1. Model optimization strategy.

potential vanishing gradient problem in deeper networks, the model employs residual connections, allowing information to be passed directly between the convolutional layers and the attention mechanism, improving training efficiency and network stability. Next, global average pooling is applied to simplify the dimensionality of the network's output, reducing the computational load on the fully connected layers. Through global average pooling, the average values of the feature maps are aggregated to produce a simplified feature representation. Finally, the processed data is fed into the classifier to complete the text classification task, with the classification results displayed at the output layer. Overall, the model effectively combines the local feature extraction capability of the convolutional layers with the long-range dependency capture capability of the multi-head self-attention mechanism. Meanwhile, it optimizes training and stability through residual connections, resulting in efficient and accurate fake news detection. This optimized CNN model architecture significantly improves the accuracy and efficiency of fake news detection and enhances the model's ability to learn and generalize different text features. To further improve the detection performance, the new model architecture incorporates multi-modal data processing techniques, covering text, video, and image processing. The designed multi-modal data processing model is plotted in Fig. 3:

Figure 3 illustrates the multimodal fake news detection model proposed in this study. The model accepts inputs from three data modalities—text, image, and video—and integrates feature extraction, alignment, and fusion mechanisms to achieve efficient and accurate fake news identification. In the data input phase, the model receives three types of information from news content: text, image, and video, which represent sources of linguistic, visual, and temporal-dynamic information, respectively. During the preprocessing and feature extraction stage: Text features are extracted using the BERT model, which captures contextual dependencies and generates high-dimensional semantic vectors. Image features are extracted by ResNet, whose deep convolutional structure and residual connections effectively capture spatial structure and visual details. Video features are obtained using the SlowFast network, a dual-path temporal modeling approach. The Slow path processes video frames at a low frame rate. This allows it to capture long-term motion information. In contrast, the Fast path processes frames at a high frame rate to capture fine-grained and short-term dynamic features. These two paths interact midstream and output a unified representation that encodes both temporal dynamics and local details. Before being input into the network, video data are decomposed into frame sequences and fed into the Slow and Fast paths separately, then merged into a unified video feature vector. In the multimodal feature fusion stage, the model first performs dimensional normalization and semantic projection on features from

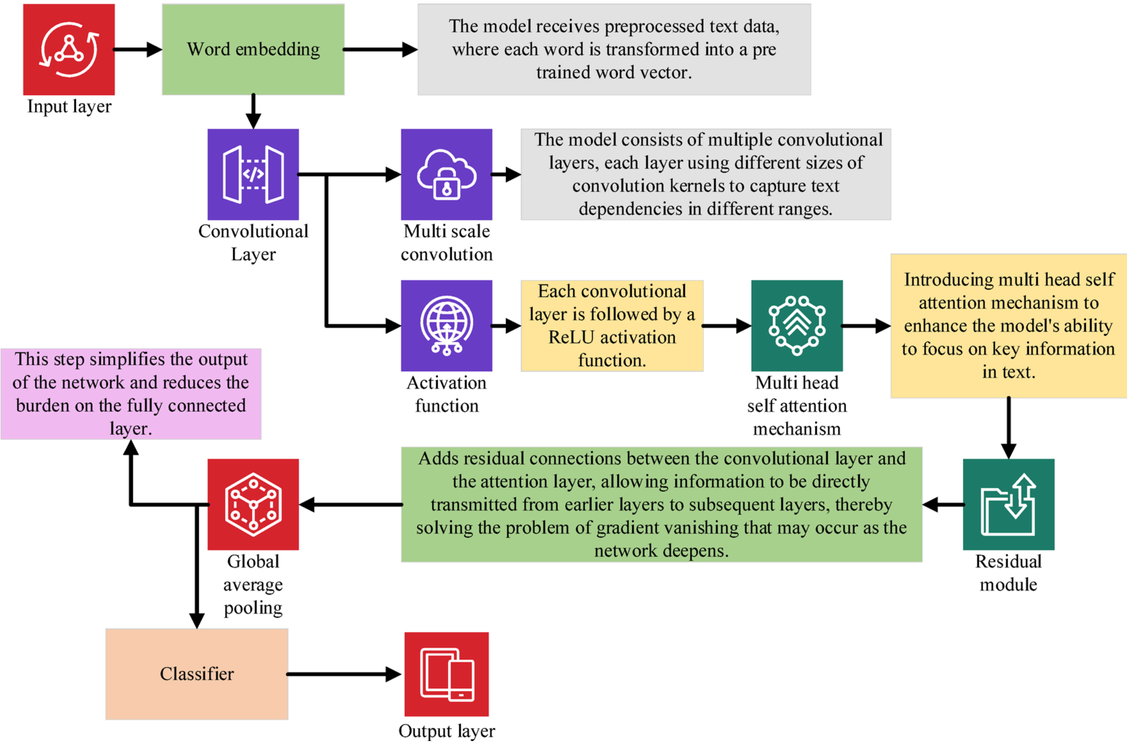

Input layer
Word embedding
The model receives preprocessed text data, where each word is transformed into a pre trained word vector.
Convolutional Layer
Multi scale convolution
The model consists of multiple convolutional layers, each layer using different sizes of convolution kernels to capture text dependencies in different ranges.
Activation function
Each convolutional layer is followed by a ReLU activation function.
Multi head self attention mechanism
Introducing multi head self attention mechanism to enhance the model's ability to focus on key information in text.
This step simplifies the output of the network and reduces the burden on the fully connected layer.
Global average pooling
Adds residual connections between the convolutional layer and the attention layer, allowing information to be directly transmitted from earlier layers to subsequent layers, thereby solving the problem of gradient vanishing that may occur as the network deepens.
Residual module
Classifier
Output layer

Fig. 2. Architecture of fake news detection model.

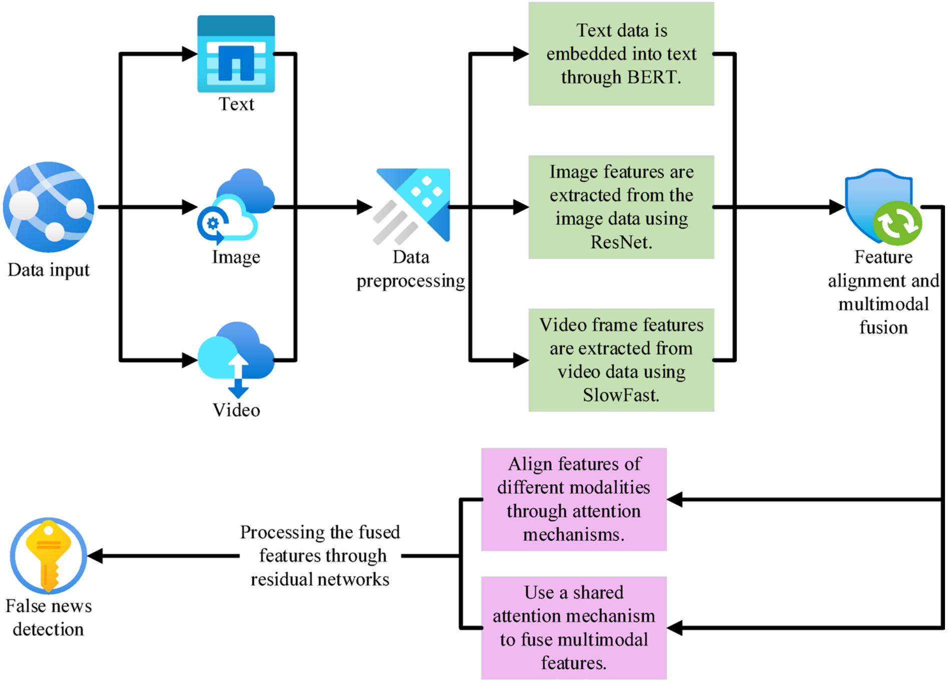

Text
Data input
Image
Video
Data preprocessing
Text data is embedded into text through BERT.
Image features are extracted from the image data using ResNet.
Video frame features are extracted from video data using SlowFast.
Feature alignment and multimodal fusion
Align features of different modalities through attention mechanisms.
Use a shared attention mechanism to fuse multimodal features.
Processing the fused features through residual networks
False news detection

Fig. 3. Multi-modal data processing model.

different modalities, mapping them into a unified feature space to address differences in modality expression. Then, a weighted attention mechanism is introduced to calculate the similarity between text, image, and video features and dynamically assign fusion weights to each modality. For example, when action scenes in a video align with the semantics of the text, the model automatically enhances the contribution of the video modality. Next, a shared attention mechanism further fuses the multimodal features by calculating their inter-modal associations to achieve complementary integration at the semantic level. Fusion can be conducted via concatenation or weighted averaging to generate the final fused feature vector. In the fusion optimization stage, the fused multimodal vector is fed into ResNet again for deep feature refinement. This process enhances the representational power of the fused features, reduces inter-modal interference, and improves the stability and generalization ability of network training. Finally, the optimized fused feature vector is input into a classifier to perform precise fake news detection.

Overall, the model fully leverages the strengths of BERT, ResNet, and SlowFast for extracting textual, visual, and temporal features, respectively. Coupled with attention-driven alignment and fusion strategies, it significantly improves the model's capacity to understand complex multisource information, demonstrating enhanced robustness and discriminative power in fake news detection tasks.

Experimental design

The datasets used in this study include the Liar, FakeNewsNet, and Weibo datasets. They are used to support the research of the false news detection model and strictly follow the terms and conditions of data collection and analysis of each dataset.

(1) The Liar dataset: The dataset was developed by Professor Lei Wang's team at the University of Texas at Austin and contains over 12,800 politically related short phrases. The data primarily comes from political debates, press conferences, and political interviews, with each entry labeled for authenticity. The categories include: "True", "Mostly True", "Half True", "Mostly False", and "False." Additionally, the text in the dataset is relatively short, focusing on the political domain, making it suitable for NLP tasks such as text classification, authenticity assessment, and fake news detection. Due to its clear labeling and wide coverage, this dataset is widely used in political text authenticity analysis research. The dataset can be downloaded from the official website (https://archive.ics.uci.edu/ml/datasets/LIAR).

(2) The FakeNewsNet dataset: The FakeNewsNet dataset was collected and organized by the FakeNewsNet project team. It is a multimodal fake news dataset containing 24,000 news articles, 13,000 images, and several videos. Each news article is accompanied by an authenticity label and linked with corresponding images and videos, enabling researchers to conduct fake news detection experiments across text, image, and video modalities. The multimodal nature of the FakeNewsNet dataset makes it particularly suitable for research on multimodal feature fusion, covering complex scenarios involving the interaction of visual and textual information. It has been widely used in academic research, including image detection, cross-modal feature fusion, and fake news detection, and the dataset can be downloaded from the official website (https://github.com/KaiDMML/FakeNewsNet).

(3) The Weibo dataset: The Weibo dataset was obtained from the Chinese social media platform Sina Weibo, specifically designed for multimodal fake news detection research. It contains 6000 images, 3000 videos, and their associated textual content. The data originates from user-posted Weibo content and focuses on the dissemination and detection of fake news in social media contexts, making it suitable for studying multimodal feature recognition and dissemination mechanisms of fake news in the social media environment. Unlike other datasets, the Weibo dataset combines Chinese text with visual information, providing a more realistic application scenario for Chinese fake news detection. Moreover, the videos and images in the dataset are labeled and closely linked to the text content, facilitating cross-modal collaborative detection research. It can be downloaded from the official website (https://github.com/yaqingwang/EANN-KDD18).

The three datasets each have distinct characteristics. The Liar dataset focuses on political text authenticity analysis and is suitable for unimodal text detection; The FakeNewsNet dataset provides a research platform for multimodal detection, involving interactions among text, images, and videos; The Weibo dataset is tailored for the Chinese social media context, emphasizing the combination of text and visual information. These datasets offer rich data support for feature extraction and fusion verification in the experiments presented in this study. Before model training, all data underwent a standardized preprocessing procedure. For the text data, cleaning was first performed to remove HTML tags, special characters, and extra spaces. The cleaned text was then tokenized using the tokenizer corresponding to the BERT pre-trained model and converted into word vector sequences. To ensure consistent input length, the maximum sequence length was set to 400. Short texts were padded, while long texts were truncated. For the image data, all images were resized to 224 × 224 pixels and normalized to match the input requirements of the ResNet model. Video data were first decomposed into frame sequences. Frame sampling followed the SlowFast network protocol—for example, sampling 8 frames per second. The sampled frames were then input into the Slow and Fast paths to extract frame-level features. The resulting feature vectors were normalized before being used in the model. The dataset was divided into training, validation, and test sets in an 8:1:1 ratio. This division ensured the scientific rigor of the evaluation process and the proper assessment of the model's generalization ability. To enhance the reproducibility of the experiments, a fixed random seed was applied during both data splitting and model training. The seed value was uniformly set to 42, ensuring consistency and repeatability of results.

Statement

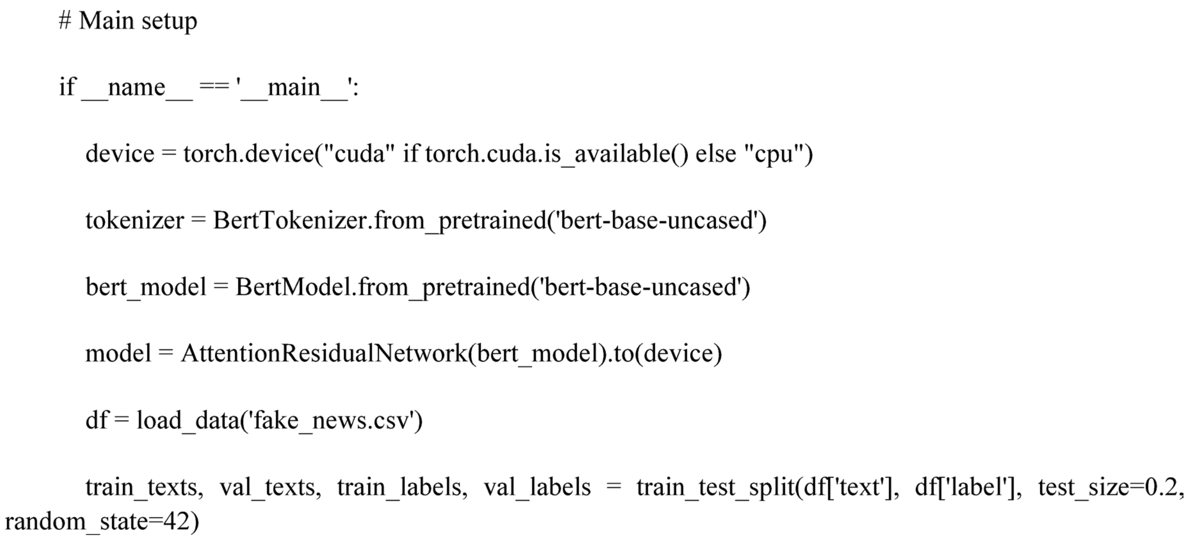
In this study, the collection, analysis, and use of all data comply with the terms and conditions of the data sources, ensuring the legality of the data and the compliance of research methods. The study team strictly adheres to relevant privacy policies and ethical standards when processing data, ensuring the fairness and scientific nature of the data analysis process.

The hardware and software configuration parameters required for the experiments are shown in Table 4.

Some relevant codes in the calculations of this study are as follows:

Equipment type	Parameter configuration
Processor	Inter(R) Xeon(R) Center Processing Unit (CPU) E5-2620 v4 @ 2.10 GHz
Graphics processing unit (GPU)	NVIDIA Titan Xp 12 GB
Memory	64 GB
Programming language	Python 3.6
Technical framework	Py Torch 1.7.0 deep learning framework

Table 4. Configuration of experimental hardware and software.



```
# Main setup

if __name__ == '__main__':

    device = torch.device("cuda" if torch.cuda.is_available() else "cpu")

    tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')

    bert_model = BertModel.from_pretrained('bert-base-uncased')

    model = AttentionResidualNetwork(bert_model).to(device)

    df = load_data('fake_news.csv')

    train_texts, val_texts, train_labels, val_labels = train_test_split(df['text'], df['label'], test_size=0.2,
random_state=42)
```

To ensure the accuracy of experimental data and the stability of model training, all algorithm parameters were configured uniformly throughout this study. For input features, the word vector dimension was set to 200, and the input sequence length was fixed at 400. In the convolutional layer, 3 × 3 kernels were used, with a total of 64 filters and a stride of 1, to capture local n-gram features from the text. The hidden layer dimension was set to 64, and the ReLU activation function was employed to enhance the network's nonlinear expression capability. The network included one fully connected layer with 512 neurons, which was directly connected to the output classifier. During training, the model adopted the cross-entropy loss function as the optimization objective. The Adam optimizer was used to balance convergence speed and accuracy. Key hyperparameters were configured as follows: the initial learning rate was set to 0.001, batch size to 32, and the number of training epochs to 30. An early stopping mechanism was employed to monitor validation loss and prevent overfitting. To further improve model stability and generalization, a Dropout mechanism with a rate of 0.5 was applied to the fully connected layer. This helped prevent neurons from becoming overly dependent on local features. The dataset was split into training and validation sets in an 8:2 ratio. After each training epoch, validation performance was evaluated to ensure the reliability and generalization capability of the model throughout training. All parameter settings were kept consistent across the LIAR, FakeNewsNet, and Weibo datasets. This uniformity ensured the fairness and reproducibility of experimental comparisons and provided a solid foundation for subsequent model performance evaluations. The algorithms chosen for comparison in the experiment include BERT, RoBERTa, Generalized Autoregressive Pretraining for Language Understanding (XLNet), ERNIE, and Generative Pre-trained Transformer 3.5 (GPT-3.5).

BERT is one of the earliest bidirectional pre-trained language models, known for its strong text representation capability. It is trained through the masked language model and next-sentence prediction tasks. The configuration chosen for this study is the BERT-base model, which has a moderate number of parameters, making it suitable for basic text feature representation. The reason for selecting BERT is its status as a milestone model in the NLP field, providing effective contextual semantic capturing capabilities, making it an ideal baseline for performance comparison. RoBERTa is an optimized version of BERT, which removes the next-sentence prediction task, increases the training data volume, and uses a larger batch size. In the experiment, the RoBERTa-base configuration is selected for its stronger robustness and generalization ability. RoBERTa is chosen because it has outperformed BERT in several NLP tasks, making it a better model for evaluating the optimized model's performance in deep semantic understanding. XLNet adopts a pre-training method combining autoregression and autoencoding. By using a bidirectional permutation mechanism, it overcomes the limitations of BERT's masked pre-training, further improving text modeling ability. The experiment uses the XLNet-base configuration, which excels in text context dependency representation. XLNet is chosen for its superior performance in handling long texts and capturing global dependencies, making it ideal for comparing multimodal text feature extraction. ERNIE is a Chinese-enhanced model based on BERT that optimizes language representation by incorporating entity knowledge and semantic information. The experiment uses the ERNIE 2.0 version, which excels in knowledge and semantic fusion. ERNIE is chosen for its adaptability to Chinese datasets, making it suitable for detecting fake news in Chinese contexts. GPT-3.5 is a generative pre-trained language model with an extremely large parameter scale and powerful text generation capabilities. A simplified version of GPT-3.5 is used for comparison in the experiment, primarily to evaluate its generalization ability in text classification tasks. GPT-3.5 is chosen because, as a next-generation large model, it represents the cutting edge of current NLP, making it an ideal reference model for top-tier performance comparison.

In summary, the models selected for the experiment represent different stages and technological approaches in the development of NLP models, including bidirectional encoding, knowledge enhancement, and generative pre-training. These models cover a wide range of methodologies, providing a solid comparative foundation for evaluating the proposed improved model in this study. Additionally, by systematically comparing the performance of BERT, RoBERTa, XLNet, ERNIE, and GPT-3.5, the performance advantages of the model in terms of accuracy, precision, recall, and F1 score can be more clearly demonstrated.

Performance and simulation experiment comparison of hybrid algorithm in tourist route recommendation system

Experimental comparison of hybrid algorithm performance

The performance metrics compared in the experiment encompass accuracy, precision, recall, and F1 score. Accuracy is a fundamental metric for evaluating the overall performance of a classification model, representing the proportion of correctly predicted samples out of the total samples. The reason for selecting accuracy is that it provides an intuitive reflection of the model's overall performance in the fake news detection task. Especially, when the dataset is evenly distributed, accuracy serves as a reliable indicator. Precision measures the proportion of true positive predictions for a specific class (e.g., fake news) among all instances predicted as positive. Precision is particularly important for fake news detection because false positives can lead to negative social impacts. A high precision enables the model to reduce the number of falsely identified fake news, thereby increasing the credibility of the detection results. Recall represents the proportion of true positives correctly identified by the model, i.e., the fraction of all fake news successfully detected. Recall is chosen because, in fake news detection tasks, it is critical to cover as much fake news as possible. In real-world applications, missing fake news (low recall) can lead to the spread of misinformation, causing severe consequences. The F1 score is the harmonic mean of precision and recall, used to balance the trade-off between the two. The reason for selecting the F1 score is that it offers a comprehensive evaluation of the model's performance in balancing precision and recall, making it especially useful in cases of imbalanced datasets. In fake news detection tasks, the F1 score effectively assesses the model's overall performance in reducing false positives and improving detection rates.

By comparing accuracy, precision, recall, and the F1 score, the model's overall performance, misjudgement control ability, coverage capability, and comprehensive effectiveness in the fake news detection task can be thoroughly evaluated. These metrics complement each other and help reveal the model's strengths and weaknesses from different perspectives, providing valuable insights for model optimization and practical application. A detailed comparison is revealed in Fig. 4.

Figure 4 denotes that in terms of accuracy, the optimized model attains an accuracy of 0.92 with a data volume of 1000, increases to 0.951 with 2000 data volumes, and peaks at an accuracy of 0.977 with 3000. In contrast, other models like BERT have an accuracy of 0.842 with 1000 data volumes, RoBERTa achieves 0.855, and XLNet performs slightly higher at 0.861. The high accuracy of the optimized model is attributed to the introduction of multimodal feature fusion and attention mechanisms, which can more effectively capture features from text, images, and videos. Additionally, leveraging residual connections addresses the vanishing gradient problem associated with increasing model depth. The model also performs exceptionally well in feature alignment and multimodal data interaction, significantly enhancing overall classification performance. Regarding precision, the optimized model records a precision of 0.95 with 1000 data volumes, increases to 0.97 with 2000, and culminates in the highest precision of 0.986 with 3000. In comparison, other models such as RoBERTa and XLNet also show good performance, with RoBERTa achieving a precision of 0.932 with 3000 data volumes, and XLNet at 0.939. BERT and ERNIE exhibit slightly lower results, reaching 0.92 and 0.926, respectively, with 3000 data volumes. The increase in precision indicates that the optimized model has a higher correctness rate when predicting true positives. This is thanks to the attention mechanism's emphasis on key features and the deep integration of multimodal data, effectively reducing the occurrence of misjudgments. In contrast, the limitations of traditional models in single-modal feature extraction result in a slightly inferior precision performance. When examining recall, the optimized model secures a recall of 0.93 at 1000 data volumes, increases to 0.953 at 2000, and reaches the highest recall of 0.969 at 3000. In contrast, models such as RoBERTa and XLNet have recall rates of 0.873 and 0.882, respectively, with 3000 data volumes, while BERT's performance is slightly lower at 0.854. A high recall means that the optimized model can more comprehensively identify fake news, reducing the false negative rate. In the task of fake news detection, recall is crucial because the failure to detect fake news can lead to serious consequences. The optimized model, through feature alignment and multi-head attention mechanisms, enhances its ability to capture and integrate multimodal features, resulting in a recall that is significantly higher than that of the baseline model. Considering the F1 score, the optimized model outperforms across all data volumes, registering F1 scores of 0.892, 0.91, and 0.924, respectively. XLNet also performs well, achieving an F1 score of 0.897 with 3000 data volumes. Other models, including RoBERTa and BERT, yield slightly lower F1 scores at the same data volume, reaching 0.888 and 0.87, respectively. The results of the F1 score further demonstrate that the optimized model achieves a good balance between precision and recall, especially as the data volume increases, the model's advantages become more pronounced. This indicates that the optimized model can reduce misjudgments and effectively cover the range of fake news identification.

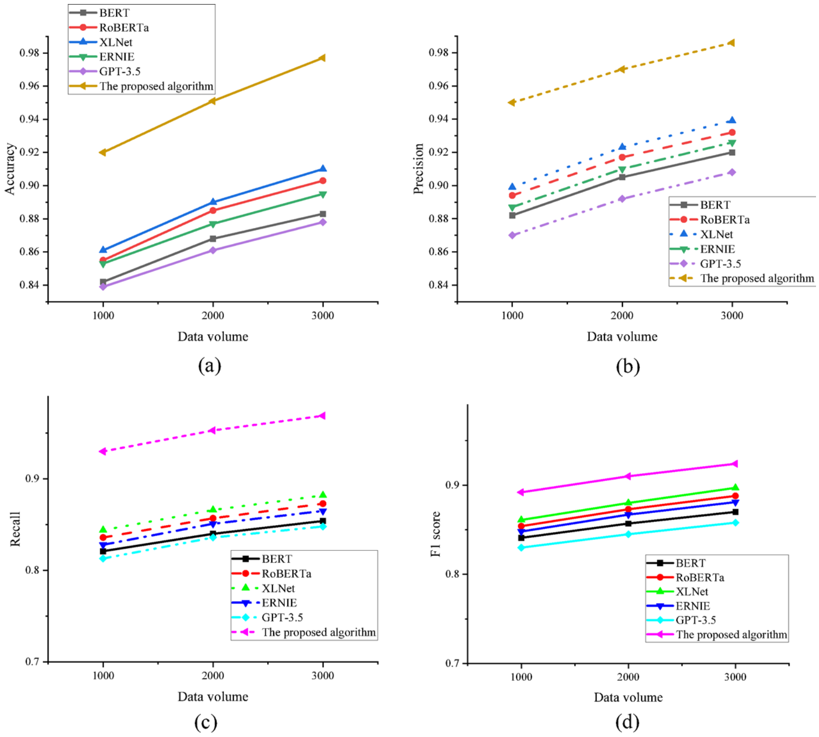


Fig. 4. Model performance comparison results (**a**) accuracy comparison results; (b) precision comparison results; (**c**) recall comparison results; (**d**) F1 score comparison results.

Analysis of simulation experiment results

To further verify the effectiveness of the optimized algorithm, simulations simulate the behavior and performance of the model when processing real-world data. The comparison metrics include robustness, generalization ability, response time, and resource consumption. Robustness is a key metric for assessing a model's performance stability when faced with data noise or perturbations. In the context of fake news detection, input data may contain errors, missing, or ambiguous information, such as grammatical mistakes, image blurring, or video frame omissions. A model with high robustness can maintain high detection accuracy under these adverse conditions, ensuring stable performance in real-world applications. Therefore, evaluating robustness helps verify the model's resistance to interference and its reliability in practical scenarios. Generalization ability refers to a model's performance on unseen data outside of the training dataset. Fake news detection tasks involve a vast array of dynamic news content, where features and patterns may change significantly across different contexts or periods. A model with good generalization ability can adapt to various data distributions while maintaining high performance, avoiding overfitting to the training set. Consequently, comparing generalization ability reflects how well the model can adapt to new data in real-world deployment. Response time is the metric for measuring the time required by the model to process a single input, reflecting its real-time processing capability. In the fake news detection task, particularly in fast-spreading information environments, the model needs to make quick decisions to mitigate the spread of fake news. Selecting response time as a metric allows for evaluating the model's processing speed under varying data volumes, ensuring its usability and efficiency in high-data-flow scenarios. Resource consumption refers to the computational resources (such as memory, CPU/GPU usage) and energy consumption required during training and inference. In practical applications, especially in low-resource devices or large-scale data processing scenarios, models with efficient resource usage offer greater practical value. By evaluating resource consumption, it is possible to visually demonstrate whether the model has lower computing overhead and higher deployable while ensuring high performance.

By evaluating four metrics—robustness, generalization ability, response time, and resource consumption—a comprehensive analysis of the model's performance and feasibility in practical applications can be conducted. These metrics focus on the model's performance while considering its stability, efficiency, and cost in complex, dynamic, and large-scale data environments. Hence, a scientific basis can be offered for the optimization and deployment of fake news detection models, as depicted in Fig. 5.

In Fig. 5, in the comparison of robustness, the optimized model exhibits significant advantages. Specifically, when the data volume is 1000, the robustness score of the optimized model is 0.890, and as the data volume increases to 2000 and 3000, the robustness scores improve to 0.910 and 0.921, respectively. In contrast, other models such as XLNet and RoBERTa have robustness scores of 0.850 and 0.838, respectively, with 3000 data volumes, while BERT and ERNIE have slightly lower scores of 0.815 and 0.825, respectively. The robustness advantage of the optimized model is primarily attributed to the deep integration of multimodal data and the effective introduction of attention mechanisms, enabling the model to better handle situations with data noise or

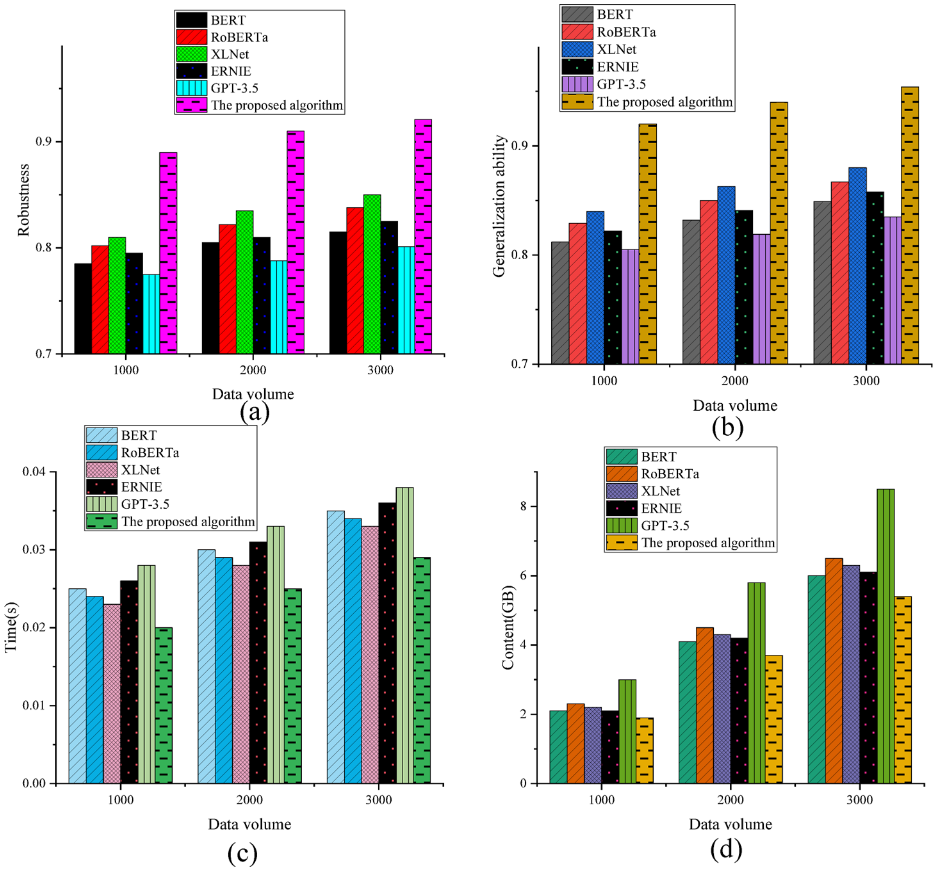


Fig. 5. Comparison results of simulation experiment (**a**) robustness comparison; (**b**) generalization ability comparison; (**c**) response time comparison; (**d**) resource consumption comparison.

incomplete information. Additionally, through residual connections, the optimized model enhances the stability of feature propagation, markedly reducing the vanishing gradient problem caused by deepening networks, and maintaining stable performance in complex data scenarios. In terms of generalization ability, the optimized model has a generalization ability of 0.920 with 1000 data volumes, which increases to 0.940 and 0.954 as the data volume rises to 2000 and 3000, respectively. Other models like XLNet and RoBERTa also demonstrate strong generalization ability, with XLNet reaching 0.880 and RoBERTa reaching 0.867 at 3000 data volumes. The optimized model's advantage in generalization ability is mainly reflected in its ability to effectively capture the correlations between multimodal data, avoiding the limitations of generalization performance on unseen data due to single-modal features. Furthermore, the attention mechanism, by highlighting key information, effectively enhances the model's adaptability to different data distributions, allowing it to perform well even in environments with differences between the training and test sets. Regarding response time, when the data volume reaches 1000, the optimized model's response time is only 0.02 s, and as it increases to 2000 and 3000, the response times are 0.025 s and 0.029 s, respectively. Compared to XLNet, the response times at the same data volume are 0.023, 0.028, and 0.033 s, respectively; RoBERTa has slightly longer response times of 0.024, 0.029, and 0.034 s. The advantage of the optimized model in response time is due to its efficient network structure and the introduction of residual modules, which make feature propagation more fluent. Moreover, the optimized attention mechanism reduces redundant computations, thus improving the model's processing speed. In real-time fake news detection tasks, reducing response time is particularly critical, enabling rapid response to the spread of false information and enhancing the model's practicality. For resource consumption, when processing 1000 data volumes, the optimized model's resource consumption is 1.9 GB, and as the data volume increases to 2000 and 3000, the resource consumption is 3.7 GB and 5.4 GB, respectively. In comparison, GPT-3.5 attains the highest resource consumption at 3000 data volumes, reaching 8.5 GB; XLNet and RoBERTa also have relatively high resource consumption, achieving 6.3 GB and 6.5 GB, respectively. The significant superiority of the optimized model in resource consumption stems from the efficient design of the network structure. Especially, optimizing residual connections and attention mechanisms can reduce computational burden while ensuring performance. In addition, the efficient implementation of feature alignment and multimodal fusion mitigates redundant computational overhead in the processing of multimodal data, thereby maximizing resource utilization.

Model	BERT	RoBERTa	XLNet	ERNIE	GPT-3.5	The optimization model of this study
Model complexity	6.2	7.1	7.5	6.8	8.2	5.9
Training time (h)	5.4	6.3	7.2	6	8.1	4.1
Number of model parameters (M)	110.5	125.4	128.3	113.2	170.7	98.6
Memory usage (MB)	2300	2450	2500	2350	2950	2100
Convergence speed (epochs)	10.2	11.8	13.5	11	15.7	7.5
Overfitting risk	0.38	0.41	0.45	0.39	0.52	0.36
Data requirement (GB)	2.6	3.1	3.4	2.9	3.8	2.1
Model interpretability	0.47	0.44	0.4	0.46	0.35	0.49

Table 5. Monitoring experiment of images and videos.

Image and video monitoring experiments

To validate the specific performance and advantages of the optimized model in a multi-modal environment, this study conducts monitoring experiments using the FakeNewsNet and Weibo datasets. Model complexity refers to the overall computational burden of the model during actual operation. It takes into account the number of parameters, network depth, the quantity of nonlinear structures, and the consumption of computational resources. While the number of parameters is one dimension of measurement, model complexity focuses more on the computational cost and actual workload per unit of time. A higher value indicates greater resource demands during the inference phase. Training time refers to the total time required for the model to train from initialization to a stable convergence state, measured in hours. This metric is assessed under consistent hardware conditions and reflects the model's training efficiency. A shorter training time suggests that the model can complete the learning process more quickly under the same conditions, making it suitable for applications requiring rapid deployment. Parameter count, measured in millions, represents the total number of trainable weights and biases in the model. This metric directly reflects the model's size and learning capacity. A larger number of parameters typically implies a more expressive model, but it also increases memory and computational time requirements for both training and inference. Memory usage indicates the peak memory consumption (in MB) during training or inference. This includes memory used by model parameters, activations, and intermediate computation caches. It is a critical metric for evaluating whether a model can be deployed on low-memory devices, such as edge devices. Convergence speed, expressed in training epochs, refers to the number of training cycles needed for the model to reach a stable performance level from its initial state. This metric indirectly reflects the model's "learning responsiveness" to the data. A lower value indicates that the model can more easily learn effective features, resulting in a more efficient training process. Overfitting risk is a composite metric based on the performance gap between the training and validation sets. It measures the likelihood that a model has learned too much from the training data, leading to decreased performance on unseen samples. A lower value suggests better generalization and adaptability to new data. Data requirement refers to the minimum dataset size needed for model training. Under consistent comparison conditions, a lower data requirement indicates that the model is better suited for small-sample learning scenarios and offers greater flexibility. Model interpretability evaluates how easily the model's output can be understood by humans. This metric is assessed based on factors such as model transparency and feature responsiveness. A higher value indicates that the model's decision logic is easier to explain through visualization or intermediate outputs, which is particularly beneficial for deployment in high-risk domains such as public communication or healthcare. The experimental results are presented in Table 5:

Table 5 shows that the optimized model performs excellently across multiple dimensions. Firstly, the training time for this model is 4.1 h, significantly shorter than other models; BERT requires 5.4 h, while GPT-3.5 takes 8.1 h. Additionally, the optimized model's memory consumption is 2100 MB, notably lower than GPT-3.5's 2950 MB. The optimized model has 98.6 M parameters, far fewer than GPT-3.5's 170.7 M. Regarding convergence speed, the optimized model requires only 7.5 epochs, whereas BERT needs 10.2 epochs. Finally, the data requirement for the optimized model is 2.1 GB, the lowest among all models. These data indicate that the optimized model has significant advantages in efficiency and resource consumption.

Discussion

In the performance comparison experiments, the accuracy of all models improved as the dataset size increased. This aligns with the general trend in machine learning, where larger datasets help models better learn complex feature distributions. However, the proposed optimized model demonstrated a significant performance leap, achieving an accuracy of 0.977 when the dataset was expanded to 3000 samples—substantially outperforming other models. This highlights its remarkable advantages in feature capture and learning capacity. The core reason behind this improvement lies in the structural synergy of residual connections and attention mechanisms. This combination ensures stable training of deep networks while enhancing the model's ability to focus on critical information. Specifically, residual connections allow shallow-layer input information to be directly propagated through deeper layers, effectively mitigating the vanishing gradient problem and enabling the model to learn more complex features through deeper architectures. Simultaneously, the attention mechanism enables dynamic assessment of feature importance. When handling multimodal data, this mechanism strengthens key modality information—such as image details or textual keywords—through a weighted strategy, thereby improving the model's overall semantic modeling capability. In contrast, although pre-trained models such as BERT, RoBERTa, and GPT-3.5 excel in textual understanding, they still face structural limitations in cross-modal processing, contextual reinforcement, and computational resource control. These constraints hinder their ability to achieve dynamic optimization in multimodal fusion and feature selection. From the perspective of precision and recall, the optimized model consistently exhibits stable and efficient performance across varying data volumes. As the dataset grows, the model becomes increasingly effective at identifying subtle semantic cues and cross-modal signals present in fake news, particularly in complex scenarios where textual and visual content are incongruent. This explains its consistently superior F1 score, as it achieves a well-balanced trade-off between false positives and false negatives. Nonetheless, some limitations remain. For instance, the current experiments did not include fine-grained evaluations across different categories of fake news, such as political, health-related, or disaster-related misinformation. As a result, the model's adaptability to specific domains has not yet been clearly validated. Moreover, while multimodal inputs generally enhance performance, the attention mechanism can be misled when image or video data is missing or of poor quality. This may negatively affect the model's final judgment. Future work could introduce modality confidence adjustment mechanisms or adaptive modality weighting strategies to further improve model robustness.

In simulation experiments, the optimized model demonstrated superiority across key indicators, including robustness, generalization, response time, and resource consumption. The improved robustness is primarily attributed to the attention mechanism's capacity to filter out noisy information, along with the redundancy in information pathways enabled by the residual structure. When faced with missing values, typographical errors, or blurred frames in videos, the model still maintained high prediction accuracy. Its generalization capability is evident in its adaptability to diverse data sources and formats—maintaining strong performance even in social media contexts such as Weibo. In comparison, Transformer-based models tend to perform consistently on unimodal tasks, but their limited structural flexibility in multimodal fusion and feature interaction reduces their adaptability to complex scenarios. While GPT-3.5 offers strong capabilities in text generation and comprehension, its large number of parameters results in high resource consumption during both training and inference. This leads to slower response times, making it less suitable for real-time news verification systems. The optimized model demonstrates distinct advantages in real-time performance and resource efficiency. As the data volume increased from 1000 to 3000 samples, its response time rose only marginally from 0.02 to 0.029 s—outperforming peer models. In terms of resource usage, optimizations in the attention module and residual path design significantly reduced redundant computations, enabling efficient operation even under limited GPU resources. This architectural design provides a practical foundation for real-world deployment and large-scale applications.

Compared to the study by Nadeem et al.[26], this study presents significant optimizations in feature extraction and multimodal fusion. Their fake news detection model primarily relied on single-modal text features, using traditional Term Frequency-Inverse Document Frequency (TF-IDF) and LSTM models for news text classification[26]. However, traditional methods exhibit limitations when handling long texts and complex semantics, making it difficult to capture deep features within the text. This study adopts the BERT model for text feature extraction, which leverages the advantages of pre-trained language models to capture contextual relationships and enhance text feature representation. Moreover, this study combines image and video features, using a multimodal fusion mechanism to strengthen the model's overall performance, addressing the issue of insufficient single-modal feature information. Experimental results show that the proposed model outperforms others in accuracy, precision, recall, and F1 score, particularly in handling multimodal data in complex scenarios. In comparison to Hashmi et al.[27], this study demonstrates stronger performance in feature alignment and fusion of multimodal data. They proposed a multimodal deep learning-based fake news detection method, utilizing CNN for image feature extraction, LSTM for text feature capture, and simple feature concatenation for data fusion[27]. However, the simple concatenation approach fails to fully exploit complementary information between different modalities, limiting the effectiveness of feature representation. This study introduces an attention mechanism to achieve feature alignment and fusion across diverse modalities, employing a weighted attention mechanism for deep interaction of multimodal data, thereby highlighting key information. Additionally, this study optimizes network training using residual connections, mitigating the vanishing gradient problem and further improving model stability and generalization ability. Compared to Hashmi et al.'s method, this study maintains lower response times and resource consumption with increasing data volumes, validating the model's efficiency and scalability. It illustrates that this study improves fake news detection accuracy, robustness, and generalization ability using advanced networks such as BERT, ResNet, and SlowFast for multimodal feature extraction. In addition, it employs the attention mechanism and ResNet for feature alignment and fusion, demonstrating significant advantages, especially when handling large-scale datasets.

Conclusion

Through in-depth theoretical research and experimental validation, this study successfully constructs and verifies a novel fake news detection model that integrates the attention mechanism and ResNet. The proposed model outperforms existing mainstream models, such as BERT, RoBERTa, XLNet, ERNIE, and GPT-3.5. It excels across multiple key performance metrics, including accuracy, precision, recall, F1 score, response time, and resource consumption. As the dataset size increases, the model demonstrates stronger stability and adaptability, proving its excellent scalability and practicality. By introducing the attention mechanism, the model enhances its ability to focus on key textual information. Meanwhile, the use of residual connections effectively alleviates the gradient vanishing problem in deep networks, ensuring that the model maintains efficient training and expressive power when learning from more complex data structures. This "deep structure + semantic focus" hybrid design offers an improved solution for fake news detection tasks. In the context of increasingly widespread and deceptive fake information, the proposed optimized model holds significant application value for media platforms and social networks that need to process information quickly and efficiently. It not only enhances the platform's ability to judge information authenticity but also provides technical support for building a healthy public opinion ecosystem. Future research can further advance in the following areas: First, evaluating the model's adaptability to different types of fake news, such as political, health, and disaster-related content, to implement more refined detection strategies. Second, in cases where image or video modalities are missing, introducing modality confidence mechanisms or designing adaptive fusion strategies could improve the model's robustness with incomplete input. Third, exploring the model's transferability in multi-language and multi-cultural contexts, combining transfer learning and domain adaptation techniques, could expand its application potential in international news and cross-platform environments. Finally, integrating edge computing or cloud service platforms to study the model's deployment in real content review systems could further verify its real-time capabilities and system integration feasibility.

In summary, this study advances the technical evolution of fake news detection models. It also provides new solutions to the challenges posed by the proliferation of false information in the information age. Ongoing optimization and application expansion will further enhance the model's intelligence, applicability, and societal impact. The model has the potential to play an important role in ensuring information authenticity and fostering a healthy communication environment.

Data availability

Data is provided within the manuscript or supplementary information files.

Received: 6 February 2025; Accepted: 4 June 2025
Published online: 01 July 2025

References

1. Amer, E., Kwak, K. S. & El-Sappagh, S. Context-based fake news detection model relying on deep learning models. *Electronics* **11**(8), 1255 (2022).
2. Razmjooy, N., Ramezani, M. & Ghadimi, N. Imperialist competitive algorithm-based optimization of neuro-fuzzy system parameters for automatic red-eye removal. *Int. J. Fuzzy Syst.* **19**, 1144–1156 (2017).
3. Ahmad, T. et al. Efficient fake news detection mechanism using enhanced deep learning model. *Appl. Sci.* **12**(3), 1743 (2022).
4. Zhang, L. et al. A deep learning outline aimed at prompt skin cancer detection utilizing gated recurrent unit networks and improved orca predation algorithm. *Biomed. Signal Process. Control* **90**, 105858 (2024).
5. Malla, S. J. & Alphonse, P. J. A. Fake or real news about COVID-19? Pretrained transformer model to detect potential misleading news. *Eur. Phys. J. Spec. Top.* **231**(18), 3347–3356 (2022).
6. Hamed, S. K., Ab Aziz, M. J. & Yaakub, M. R. A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. *Heliyon* **5**(1), 56–58 (2023).
7. Guo, Z., Zhang, Q., Ding, F., Zhu, X. & Yu, K. A novel fake news detection model for context of mixed languages through multiscale transformer. *IEEE Trans. Comput. Soc. Syst.* **11**(1), 70 (2023).
8. Qu, Z., Meng, Y., Muhammad, G. & Tiwari, P. QMFND: A quantum multimodal fusion-based fake news detection model for social media. *Inform. Fus.* **104**(32), 102172 (2024).
9. Nadeem, M. I., Mohsan, S. A. H., Ahmed, K. & Mostafa, S. M. HyproBert: A fake news detection model based on deep hypercontext. *Symmetry* **15**(2), 296 (2023).
10. Al-Tai, M. H., Nema, B. M. & Al-Sherbaz, A. Deep learning for fake news detection: Literature review. *Al-Mustansiriyah J. Sci.* **34**(2), 70–81 (2023).
11. Khan, A. A., Madendran, R. K., Thirunavukkarasu, U. & Faheem, M. D2PAM: Epileptic seizures prediction using adversarial deep dual patch attention mechanism. *CAAI Trans. Intell. Technol.* **8**(3), 755–769 (2023).
12. Khan, A. A., Mahendran, R. K., Perumal, K. & Faheem, M. Dual-3DM 3 AD: mixed transformer based semantic segmentation and triplet pre-processing for early multi-class Alzheimer's diagnosis. *IEEE Trans. Neural Syst. Rehabil. Eng.* **32**, 696–707 (2024).
13. Perumal, K., Mahendran, R. K., Khan, A. A. & Kadry, S. Tri-M2MT: Multi-modalities based effective acute bilirubin encephalopathy diagnosis through multi-transformer using neonatal Magnetic Resonance Imaging. *CAAI Trans. Intell. Technol.* **10**(2), 434–449. https://doi.org/10.1049/cit2.12409 (2025).
14. Ouassil, M. A. et al. A fake news detection system based on combination of word embedded techniques and hybrid deep learning model. *Int. J. Adv. Comput. Sci. Appl.* **13**(10), 10 (2022).
15. Liu, H. & Ghadimi, N. Hybrid convolutional neural network and Flexible Dwarf Mongoose Optimization Algorithm for strong kidney stone diagnosis. *Biomed. Signal Process. Control* **91**, 106024 (2024).
16. Han, M., Zhao, S., Yin, H., Hu, G. & Ghadimi, N. Timely detection of skin cancer: An AI-based approach on the basis of the integration of Echo State Network and adapted Seasons Optimization Algorithm. *Biomed. Signal Process. Control* **94**, 106324 (2024).
17. Cai, X., Li, X., Razmjooy, N. & Ghadimi, N. Breast cancer diagnosis by convolutional neural network and advanced thermal exchange optimization algorithm. *Comput. Math. Methods Med.* **2021**(1), 5595180 (2021).
18. Seddari, N. et al. A hybrid linguistic and knowledge-based analysis approach for fake news detection on social media. *IEEE Access* **10**(2), 62097–62109 (2022).
19. Whitehouse, C., Weyde, T., Madhyastha, P. & Komninos, N. Evaluation of fake news detection with knowledge-enhanced language models. *Proc. Int. AAAI Conf. Web Soc. Med.* **16**(4), 1425–1429 (2022).
20. Xu, Z., Sheykhahmad, F. R., Ghadimi, N. & Razmjooy, N. Computer-aided diagnosis of skin cancer based on soft computing techniques. *Open Med.* **15**(1), 860–871 (2020).
21. Das, S. D., Basak, A. & Dutta, S. A heuristic-driven uncertainty based ensemble framework for fake news detection in tweets and news articles. *Neurocomputing* **491**(92), 607–620 (2022).
22. Razmjooy, N., Sheykhahmad, F. R. & Ghadimi, N. A hybrid neural network–world cup optimization algorithm for melanoma detection. *Open Med.* **13**(1), 9–16 (2018).
23. Capuano, N., Fenza, G., Loia, V. & Nota, F. D. Content-based fake news detection with machine and deep learning: A systematic review. *Neurocomputing* **530**(15), 91–103 (2023).
24. Jain, V., Kaliyar, R. K., Goswami, A., Narang, P. & Sharma, Y. AENeT: An attention-enabled neural architecture for fake news detection using contextual features. *Neural Comput. Appl.* **34**(1), 771–782 (2022).
25. Choudhury, D. & Acharjee, T. A novel approach to fake news detection in social networks using genetic algorithm applying machine learning classifiers. *Multimed. Tools Appl.* **82**(6), 9029–9045 (2023).
26. Nadeem, M. I. et al. HyproBert: A fake news detection model based on deep hypercontext. *Symmetry* **15**(2), 296 (2023).
27. Hashmi, E., Yayilgan, S. Y., Yamin, M. M., Ali, S. & Abomhara, M. Advancing fake news detection: Hybrid deep learning with fasttext and explainable AI. *IEEE Access* **56**(11), 67 (2024).

Acknowledgement
This work was support by Scientific Research Program Funded by Education Department of ShaanxiProvincial Government (Program No.24JK0200).

Author contributions
Ying Lu contributed to conception and design of the study. Naiwei Yao organized the database. Ying Lu performed the statistical analysis. Naiwei Yao wrote the first draft of the manuscript. Ying Lu wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Declarations

Competing interests
The authors declare no competing interests.

Human participation statement
This study does not involve human participants.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-05702-w.

Correspondence and requests for materials should be addressed to Y.L.

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.