Received: 30 April 2025 | Revised: 11 June 2025 | Accepted: 7 July 2025 | Published online: 28 August 2025

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2025, Vol. 4(4) 603-619
DOI: 10.47852/bonviewJCCE52026051

BON VIEW PUBLISHING

Fake News Detection with Deep Learning: Insights from Multi-dimensional Model Analysis

Qiuping Li[1], Fen Fu[1], Yinjuan Li[1], Bhunnisa Wisassinthu[1], Wirapong Chansanam[1,*] and Tossapon Boongoen[2]

[1]*Faculty of Humanities and Social Sciences, Khon Kaen University, Thailand*

[2]*Advanced Reasoning Research Group, Aberystwyth University, UK*

Abstract: This study aims to systematically evaluate and compare various deep learning models in terms of accuracy, efficiency, and interpretability for fake news detection. Leveraging recent advancements in pretrained models (e.g., BERT, RoBERTa) and lightweight frameworks (e.g., TextCNN), we implemented and optimized multiple detection models. Comparative analysis was conducted on a dataset containing approximately 40,000 news texts. Results revealed that BERT Large significantly outperformed other models, achieving an accuracy of 99.33%, attributed to its extensive semantic understanding capabilities. Conversely, TextCNN, despite its simpler architecture, achieved competitive accuracy (98.77%), demonstrating substantial practical value for resource-limited environments. Interpretability analysis via attention visualization highlighted distinct cognitive strategies of pretrained models when classifying real versus fake news. While the study addresses critical technical challenges in fake news detection, limitations related to potential dataset biases and domain specificity were acknowledged, suggesting opportunities for future research on multimodal and cross-domain adaptations. This research contributes substantially by providing practical benchmarks and interpretability insights, significantly enhancing real-world fake news detection systems, thus aiding platforms in combating misinformation effectively.

Keywords: fake news detection, deep learning, BERT, TextCNN, model interpretability

1. Introduction

With the rapid development of social media and online information platforms, the mode and efficiency of news dissemination have undergone profound changes, allowing users around the world to receive and share massive amounts of information from various sources for the first time. However, while this information ecosystem facilitates the public's access to information, it also provides an unprecedented environment for the growth and dissemination of fake news [1]. Fake news often employs exaggerated language, evokes strong emotions, or presents misleading multimedia content to capture attention and clicks, thereby compromising the public's ability to form accurate judgments about facts and potentially undermining political opinion, economic decision-making, and social stability [2].

The decentralized communication mechanism of social media platforms has reconstructed the information ecological pattern. According to statistics, Twitter adds approximately 500 million tweets per day, of which about 15% involve news-like content. Meanwhile, the average detection lag time for fake news is more than 20 h [3]. This time lag leads to an exponential increase in the reach of false information; for example, during the 2020 U.S. election, retweets of false political news were 70% higher than those of real news [4]. More grimly, breakthroughs in generative AI technologies (e.g., GPT-4, DALL-E 3) have brought the marginal cost of faking multimodal content close to zero, and MIT experiments have confirmed that the accuracy of human recognition of AI-generated fake news is only 48.7% [5]. In this context, building automatic detection systems with strong generalization capabilities has become a core topic in the field of cybersecurity.

In the face of the social harm caused by the spread of fake news, both academia and industry have invested considerable effort in developing automatic detection and identification mechanisms. Effective fake news detection systems can help platforms identify and flag suspicious content promptly, reducing the scope and impact of misinformation. At the same time, such technologies can also improve the public's media literacy, cultivate critical thinking, and enhance the ability to recognize online information [6].

At the technical level, fake news detection involves cutting-edge technologies in multiple fields, such as natural language processing, machine learning, and social network analysis. Its research progress not only promotes the development of these fields but also provides new ideas for solving the universal problem of information authenticity verification [7]. In particular, research comparing the computational efficiency and performance balance of different detection models can provide an important reference for practical deployment scenarios and address the detection needs in resource-constrained environments [8].

The following major challenges are currently facing the field of fake news detection:

Model generalization and robustness: The expression of fake news is constantly evolving, and existing models often exhibit insufficient generalization ability when encountering new domains or types of fake content [9].

Data scarcity and category imbalance: High-quality labeled data is challenging to obtain, and there is often an imbalance in the amount of real and fake news, which affects the effectiveness of model training [10].

***Corresponding author:** Wirapong Chansanam, Faculty of Humanities and Social Sciences, Khon Kaen University, Thailand. Email: wirach@kku.ac.th

Multimodal fusion: In reality, fake news is often accompanied by multimedia content, such as pictures and videos, and it is difficult to comprehensively capture the information characteristics solely by relying on a single modal analysis [11].

Computational efficiency and deployment cost: While large-scale pre-training models exhibit superior performance, their vast number of parameters and high computational requirements limit their practical application in resource-constrained environments [12].

To address the above challenges, this study proposes a systematic framework for comparing and optimizing fake news detection methods, with the following main innovations and contributions:

Multimodel comparison and performance boundary exploration: This study systematically implements and compares five typical models (BERT Large, RoBERTa, TextCNN, BERT + XGBoost and logistic regression) from simple to complex on the same dataset, establishes the performance benchmarks and upper bounds for the fake news detection task, and provides empirical evidence for model selection.

Analysis of the trade-off between computational efficiency and detection accuracy: By recording and analyzing in detail the relationship between the training time, the number of parameters, and the detection accuracy of each model, this study reveals the optimal balance between resource efficiency and performance of the TextCNN model, and provides an optimization strategy for practical deployment scenarios.

Optimization of BERT large model parameters: In this study, the key hyperparameters (learning rate, batch size, classification threshold, etc.) are systematically optimized for the BERT Large model, so that it achieves an accuracy of 99.10% and a high macro-averaged F1-score on the test set, which significantly improves the detection performance.

Decoupling method of feature extraction and classification: The hybrid BERT + XGBoost model implemented in this study provides an efficient and practical solution for resource-constrained scenarios, significantly reducing computational requirements while maintaining reasonable performance by combining deep feature extraction with traditional classifiers.

Category imbalance problem-solving strategy: Through threshold optimization techniques, this study effectively addresses the common category imbalance problem in false news detection, enabling each model to achieve a more balanced performance across the two categories and thereby improving the overall detection effect.

Previous studies on fake news detection, while valuable, have faced significant limitations, particularly in their inability to simultaneously address critical dimensions such as accuracy, computational efficiency, and model interpretability. Most research has either concentrated on the superior accuracy provided by complex pretrained models or highlighted computationally efficient yet less powerful lightweight models, thereby overlooking the critical balance required among these dimensions. Consequently, a comprehensive evaluation that integrates these factors remains scarce, resulting in fragmented insights and limited practical applicability. To address this notable gap, this study proposes a systematic, multi-dimensional evaluation approach utilizing both pretrained models (e.g., BERT, RoBERTa) and lightweight architectures (e.g., TextCNN). By concurrently examining accuracy, efficiency, and interpretability, this research aims to provide a holistic understanding of fake news detection models, offering robust benchmarks and guiding principles for real-world implementation and future methodological advancements.

2. Literature Review

Fake news detection, as an interdisciplinary research topic, has attracted widespread attention in recent years. In this section, we will systematically review the existing research methods and their progress, from traditional machine learning methods to deep learning techniques to the latest pre-trained language models and hybrid architectures, and explore the characteristics and limitations of each type of methods.

2.1. Fake news detection based on traditional machine learning

Early fake news detection methods mainly used traditional machine learning models to classify text features by manually extracting them. For example, Altunbey Özbay and Alatas [13] compared 23 supervised learning algorithms, including Support Vector Machine (SVM) and Decision Tree, for social media data by extracting features such as a word frequency matrix. The results showed that the Decision Tree achieved optimal performance. Overall, such traditional methods can recognize fake news to a certain extent, but they are highly dependent on manual feature engineering and cannot capture deep semantic information [14].

2.2. Application of deep learning in fake news detection

With the development of deep learning, researchers have begun to utilize models such as Text Convolutional Neural Networks (TextCNN) and Recurrent Neural Networks (e.g., Bidirectional Long Short-Term Memory Networks, BiLSTM) to learn text features for fake news detection automatically. Deep neural networks can learn contextual semantic features from data and typically achieve higher accuracy rates compared to traditional methods [15]. For example, Ouassil et al. [15] constructed a hybrid model that combines convolutional and bidirectional LSTMs, and fused pre-trained word vectors to achieve significant improvements in classification accuracy and precision compared to traditional machine learning algorithms.

2.3. Application of pre-trained language models in fake news detection

In recent years, pre-trained language models have demonstrated impressive results in detecting fake news. Large pre-trained models, such as BERT, RoBERTa, XLNet, etc., can be fine-tuned to provide deep semantic understanding for the detection task, which dramatically improves the classification performance [16]. The FakeBERT model proposed by Kaliyar et al. [16] combines BERT with a convolutional network and achieves an accuracy of 98.9% on the fake news dataset, significantly outperforming previous models. In addition, specialized pre-trained models designed for detecting neural network-generated news have emerged, such as Grover. Grover, a generative Transformer-based model studied by Gagiano et al. [17], has been shown to outperform baselines such as GPT-2 and regular BERT in automatically distinguishing between machine-generated fake news and authentic news.

2.4. Mixed modeling and integrated learning

To further improve detection, researchers have explored hybrid models and integrated learning methods. For example, Essa et al. [18] proposed a hybrid architecture that incorporates BERT with LightGBM, exhibiting better performance than a single model. These studies combine semantic vectors extracted from pre-trained BERT models with gradient-boosting decision trees to achieve a "BERT + XGBoost"-style fusion classification, leveraging the complementary advantages of deep semantic features and traditional models [18]. In addition, Zhou and Zafarani [7] found that hybrid models combining

deep feature extraction with traditional classifiers can achieve balanced performance in resource-constrained environments by comprehensively evaluating multiple integration approaches. These hybrid and integrated approaches achieved state-of-the-art performance in the fake news detection task by synthesizing the strengths of different models.

2.5. Current challenges and research motivations

Despite the significant progress of existing methods, there are still some shortcomings in the field of fake news detection that require further research [14]. First, different studies often use their own datasets and evaluation metrics, lacking unified benchmarks and systematic comparisons, which makes it challenging to compare the effectiveness of various methods [8] directly. Second, although large pre-trained models have high accuracy, the computational resource overhead is huge, and the inference speed and deployment efficiency of the models become bottlenecks in practical applications [19]. Again, the distribution of dataset categories in real scenarios is usually unbalanced (e.g., there is far more real news than fake news), and this category imbalance leads to a model bias toward the majority category, which weakens the recognition of fake news in the minority category [20]. Finally, different models are more sensitive to hyperparameter settings, and the current systematic research on hyperparameter tuning is still relatively insufficient, which also affects the further improvement of model performance to some extent [21]. In the future, more in-depth research is needed on the unification of evaluation standards, model efficiency improvement, data imbalance treatment, and parameter optimization, among other areas, to further enhance the practicality and robustness of the fake news detection system.

3. Research Methodology

In this study, three categories of five different models are used in the fake news detection task, including a pre-trained language model based on Transformer, a recurrent neural network model, and a traditional machine learning model. First, we constructed a dataset comprising approximately 40,000 news texts derived from the publicly available dataset on Hugging Face [22], and divided it into a training set (approximately 32,500 texts) and a testing set (approximately 7,800,000 texts). Before model training, we performed preprocessing operations on the text data, e.g., standardizing the text length to 128 words (truncating excessively long texts and padding excessively short ones), and chose appropriate text representations according to the model type. For example, for the BERT-based model, we utilized its own WordPiece splitter to split the text into subword sequences. In contrast, for the BiLSTM and logistic regression models, we constructed a vocabulary list of approximately 140,000 words and converted the text into word frequency vectors or word embedding sequences, respectively.

3.1. BERT large model fine-tuning

BERT (Bidirectional Encoder Representations from Transformers) is a deep bi-directional Transformer language model proposed by Devlin et al. [12]. It is pre-trained on a large-scale corpus to obtain rich semantics by combining the two tasks of language modeling and next-sentence prediction. The adopted Transformer structure efficiently models global dependencies within sequences through a multi-head self-attention mechanism, thereby avoiding the limitations of traditional recurrent or convolutional neural networks [23]. The BERT Large Uncased model (HuggingFace implementation) was chosen for this study. The model architecture comprises a 24-layer Transformer encoder, 1024-dimensional hidden units, and 16 attention heads, totaling approximately 335 million parameters (335,143,938). BERT Large, with its powerful feature representation capability, is can effectively recognize semantic and contextual information in fake news.

In the fine-tuning phase, we add a fully connected classification layer on top of the BERT pre-trained model and optimize the whole model for end-to-end training using the training set data. We used the BERT tokenizer as the disambiguator and set the maximum sequence length to 128. For the training parameters, AdamW was chosen as the optimizer; the learning rate was set to 1e-5, the batch size to 16, and the total number of training rounds to 6, with early stopping at round 5 as the model had reached its optimal performance. Model training was done on Tesla V100-SXM2 (32 GB) hardware. The model was tested and achieved an accuracy and macro F1 value of 0.9933 on the test set, indicating excellent performance on the fake news detection task. Table 1 details the complete parameter configuration of the BERT Large model.

3.2. TextCNN (FakeNews-CNN)

To further capture the sequential features of text, this study employs FakeNews-CNN, a text categorization model based on convolutional neural networks. The model is based on the TextCNN architecture proposed by Kim [24], which captures n-gram features in the text through multi-scale convolutional kernels to extract key semantic information from short texts efficiently. Our implementation utilizes three different sizes of convolutional kernels (3, 4, and 5), each with 128 filters, to capture text patterns of varying lengths simultaneously.

The TextCNN model has a vocabulary size of 138,087, a fixed sequence length of 128, and an embedding dimension of 100. After the convolutional layer, we extract the most salient features using a global maximum pooling strategy and introduce a dropout rate of 0.5 to prevent overfitting. The model's output layer utilizes a Dense layer with a Sigmoid activation function, resulting in a total of 13,963,069 parameters. For optimization, we use the Adam optimizer, with a learning rate of 0.001, a batch size of 16, and a binary cross-entropy loss function.

The training process was conducted under the TensorFlow 2.9 framework, utilizing NVIDIA Tesla V100-SXM2 32 GB GPUs as the training hardware. We set the maximum number of training rounds to 5, while using an early stopping strategy (patience value of 2 to reach the best performance in round 3). The dataset is divided into a training set (29,222 samples), a validation set (3247 samples), and a test set (8118 samples). With an optimal threshold of 0.43, the model achieved an accuracy of 0.9877 and a macro F1 score of 0.9866 on the test set, demonstrating

Table 1
Fine-tuning configuration of BERT Large model

Parameter	Setting
Model Name	BERT Large Uncased (HuggingFace)
Architecture	24 Transformer layers, 1024 hidden units, 16 heads
Parameters	335,143,938 (~335 M)
Tokenizer	BertTokenizer
Max Sequence Length	128
Optimizer	AdamW (lr = 1e-5)
Batch Size	16
Epochs	6 (EarlyStopped at epoch 5)
Hardware	Tesla V100-SXM2 (32 GB)
Test Accuracy	0.9933
Macro F1-Score	0.9933

excellent capabilities in detecting fake news. Table 2 presents the complete parameter configuration details of the FakeNews-CNN model.

3.3. BERT base fine-tuning

In this study, we conducted fine-tuning experiments on the BERT Base and RoBERTa models to evaluate the performance of different pre-trained Transformer architectures in the task of detecting fake news. Both models used the same training parameters and data partitioning, so the following discussion will use BERT Base as a representative example. The RoBERTa model differs only in the source of its pre-trained weights and underlying encoding structure, with all other settings remaining consistent.

We used the BERT Base Uncased and RoBERTa Base models, both of which include a 12-layer Transformer encoder, 768-dimensional hidden units, and 12 attention heads, with approximately 102 million parameters. The main differences between the two models are as follows: BERT uses the WordPiece tokenizer and the "[CLS] Text [SEP]" input format, while RoBERTa uses the BPE tokenizer and the "<s> Text </s>" input format.

During training, the Adam optimizer was used with a learning rate of 1e-5, a batch size of 16, and a maximum of 4 training epochs. Early stopping (with a patience value of 1) was enabled to prevent overfitting. The dropout rate was set to 0.1, and the loss function was sparse classification cross-entropy. Training was performed using a NVIDIA Tesla V100-SXM2 32 GB GPU in the TensorFlow 2.9 framework.

We evaluated the model on 29,222 training samples, 3247 validation samples, and 8118 test samples. Under the default threshold of 0.53, the BERT Base model achieved a test accuracy of 0.9808 and a macro F1 score of 0.9807. The RoBERTa Base model achieved an accuracy of 0.9812 and a macro F1 score of 0.9810 under the same conditions, with a ROC AUC of 0.9980, showing slightly better performance. Further, by adjusting the threshold to 0.60, the RoBERTa model's accuracy improved to 0.9832, and its macro F1 score reached 0.9831, demonstrating stronger fake news detection capabilities. Table 3 lists the detailed configuration information for both models.

Table 2
Model configuration of TextCNN (FakeNews-CNN)

Parameter	Setting
Model Name	FakeNews-CNN (TextCNN-based)
Vocabulary Size	138,087
Sequence Length	128
Embedding Dimension	100
Convolution Kernels	[3, 4, 5] (128 filters each)
Pooling Strategy	Global Max Pooling
Dropout Rate	0.5
Output Layer	Dense (1), Sigmoid activation
Total Parameters	13,963,069
Optimizer	Adam
Learning Rate	0.001
Batch Size	16
Max Epochs	5 (EarlyStopping at epoch 3, patience = 2)
Loss Function	Binary Crossentropy
Framework & Version	TensorFlow 2.9
Training Hardware	NVIDIA Tesla V100-SXM2 32 GB GPU
Train/Validation/Test Size	29,222/3247/8118
Test Accuracy*	0.9877 (with best threshold = 0.43)
Macro F1-Score*	0.9876

3.4. BERT(CLS) + XGBoost hybrid model

In this study, a hybrid feature extraction and classification method is employed, where the pre-trained BERT model is first utilized to extract the deep context-embedded features of the text, and then XGBoost is applied for classification. Jane [25] pointed out in a performance evaluation study of BERT, XGBoost, and the hybrid model that the hybrid model combines the deep semantic representation capability of the BERT model and the efficient decision tree classification performance of the XGBoost model, effectively improving classification accuracy and computational efficiency.

Table 3
Configuration of BERT base and RoBERTa base models

Parameter	BERT Base	RoBERTa Base
Model Name	BERT Base Uncased	RoBERTa Base
Architecture	12-layer Transformer, 768 hidden units, 12 attention heads	12-layer Transformer, 768 hidden units, 12 attention heads
Total Parameters	~110 million	~125 million
Tokenizer	BertTokenizer (HuggingFace)	RobertaTokenizer (HuggingFace)
Max Sequence Length	128	128
Input Format	[CLS] Text [SEP]	<s> Text </s>
Pre-trained Source	Roberta-base-uncased	Roberta-base
Optimizer	Adam	Adam
Learning Rate	1e-5	1e-5
Batch Size	16	16
Max Epochs	4 (early stopped at epoch 2)	4 (completed all epochs)
Early Stopping	Patience = 1 (based on validation loss)	Same as left
Loss Function	Sparse categorical cross-entropy	Sparse categorical cross-entropy
Dropout Rate	0.1	0.1
Framework & Version	TensorFlow 2.9	TensorFlow 2.9
Training Hardware	NVIDIA Tesla V100-SXM2 32 GB GPU	NVIDIA Tesla V100-SXM2 32 GB GPU
Train/Val/Test Size	29,222/3247/8118	29,222/3247/8118
Test Accuracy*	0.9808 (threshold = 0.53)	0.9812 (threshold = 0.53); 0.9832 (threshold = 0.60)
Macro F1-Score*	0.9807	0.9810 (default); 0.9831 (optimized)

In our implementation, BERT Base Uncased is utilized as a feature extractor (using TFBertModel and freezing all layers) to extract the CLS token embeddings from the news text, resulting in a 768-dimensional vector representation. The sequence length is set to 128, and the text is processed using the BertTokenizer from the HuggingFace Transformers library. Approximately 110 million parameters of the BERT model are frozen during the feature extraction process and are used solely as feature extractors.

For the classification task, we employ the XGBoost classifier, explicitly using the Gradient Boosted Decision Tree (GBDT) algorithm. The key hyperparameters include the following: the number of estimators (n_estimators) is set to 200, the maximum tree depth (max_depth) is 6, the learning rate (learning_rate) is 0.01, the sample subsampling rate (subsample) is 0.8, and the feature column sampling rate (colsample_bytree) is 1.0. Due to the XGBoost API limitation, the early stopping strategy cannot be realized. The feature extraction batch size was set to 16, and the model was trained on NVIDIA Tesla V100-SXM2 32 GB GPUs using the TensorFlow 2.9 and XGBoost 3.0.0 frameworks.

The time efficiency of the entire processing flow is excellent: the feature extraction phase took 361.07 s, while XGBoost training took only 29.32 s. As with the previous model, we used 29,222 training samples, 3247 validation samples, and 8118 test samples. With an optimization threshold of 0.55, this hybrid model achieves a test accuracy of 0.9261 and a macro F1 score of 0.9258. Through feature importance analysis, we found that the three most influential features are #720 (importance 0.0659), #512 (importance 0.0510), and #414 (importance 0.0226).

Table 4
Configuration of BERT + XGBoost hybrid model (fake news detection task)

Parameter	Setting
Model Name	BERT + XGBoost Hybrid
Feature Extractor	BERT Base Uncased (TFBertModel, frozen layers)
Feature Type	CLS Token Embedding (768-dimensional)
Sequence Length	128
BERT Parameters	~110 million (frozen during training)
Tokenizer	BertTokenizer (HuggingFace Transformers)
Classifier	XGBoost Classifier
XGBoost Algorithm	Gradient Boosting Decision Trees (GBDT)
Key Hyperparameters	n_estimators = 200, max_depth = 6, learning_rate = 0.01, subsample = 0.8, colsample_bytree = 1.0
Early Stopping	Not available due to API limitation
Batch Size	16 (for feature extraction)
Framework & Version	TensorFlow 2.9 + XGBoost 3.0.0
Training Hardware	NVIDIA Tesla V100-SXM2 32 GB GPU
Feature Extraction Time	361.07 s
XGBoost Training Time	29.32 s
Train/Val/Test Size	29,222/3247/8118
Test Accuracy*	0.9261 (optimized threshold = 0.55)
Macro F1-Score*	0.9258

Table 4 presents the complete parameter configuration details of the BERT + XGBoost hybrid model.

3.5. Logistic regression (word frequency features)

As a baseline model of traditional machine learning methods, this study first employs the logistic regression algorithm for text authenticity classification. During the experimental process, we constructed a large-scale vocabulary list with approximately tens of thousands of dimensions based on the training corpus. We extracted the TF (Term Frequency) and TF-IDF (Term Frequency-Inverse Document Frequency) feature vectors of each news text as input representations, respectively. To alleviate the overfitting problem that may result from the high-dimensional feature space, we implement appropriate dimensionality reduction on the feature vectors and introduce L2 regularization terms to constrain the model complexity. The logistic regression model quantitatively evaluates the degree of contribution of each word to the authenticity of the text by learning a set of linear weighting coefficients and then maps the weighted sum to the predicted probability in the interval [0,1] with the help of a sigmoid activation function and finally makes a binary classification decision with a discriminant threshold of 0.5. This method has the advantages of high computational efficiency and strong model interpretability. Still, at the same time, it has the inherent limitation of limited expressive ability, i.e., it can only capture simple linear relationships between words and tags, and it is difficult to effectively model the structured information and deep semantic features of text as deep learning models do. Nevertheless, it has been demonstrated that combining logistic regression with optimized feature representations can still achieve a satisfactory level of performance in specific text classification tasks [26]. Table 5 presents the parameter configuration details and evaluation results of the logistic regression model.

3.6. Cross-validation

To robustly evaluate the generalization ability of the proposed BERT Large model and address potential issues such as overfitting, dataset leakage, and data homogeneity bias, we adopted a rigorous stratified 5-fold cross-validation method. Specifically, we first divided

Table 5
Logistic regression model parameter configuration (fake news detection task)

Parameter	Configuration
Model Name	Logistic Regression (TF-IDF Features)
Feature Extractor	TfidfVectorizer
Input Text	Title + Body Concatenation
Regularization Type	L2 (Default)
Max Iterations	1000
Framework & Version	scikit-learn
Train/Test Split	32,469/8118
Random Seed	42
Test Accuracy	86.46%
Macro F1 Score	86.37%
Weighted F1 Score	86.46%
Class 0 Precision	85.37%
Class 0 Recall	85.16%
Class 1 Precision	87.39%
Class 1 Recall	87.57%

the dataset (40,587 samples) into a training-validation set (80%) and an independent test set (20%) through stratified sampling. During cross-validation, each fold maintained consistent category distribution, and information leakage was prevented through model re-initialization. Training used the same hyperparameters (learning rate 1×10^{-5}, batch size 16), and early stopping (patience value = 1) was employed to monitor validation accuracy and prevent overfitting.

4. Results

4.1. Experimental setting and data set description

All experiments in this study were conducted in an environment equipped with NVIDIA Tesla V100-SXM2 32 GB GPUs and implemented using TensorFlow and PyTorch frameworks. To ensure the reproducibility of the results, we have released the complete source code, pre-processed data, and other content. The dataset comprises approximately 40,587 news texts, including both real and fake news, with headlines and body text as input features. These texts were divided into a training set (32,469) and a test set (8118) after an 80%/20% split.

4.2. Comparison of model performance

We comprehensively compare the performance of the above five models in the fake news detection task using the following evaluation metrics: Accuracy, Precision, Recall, and the macro-averaged F1 score (F1-score). The performance metrics of each model are shown in Table 6.

From the results in the table, it is evident that the BERT Large fine-tuning model performs best, achieving accuracy and F1 score of approximately 99.3%, which is significantly better than the other models. Its precision and recall also reach about 99.3%, reflecting excellent classification ability and generalization performance.

Figure 1 shows the confusion matrix of the BERT Large model, and it can be visualized that the model performs very well on both positive and negative categories, with 99.2% True Negative and 99.4% True Positive.

4.3. Training process and optimization strategy analysis of BERT Large model

The exceptional performance of the BERT Large model mainly stems from its massive parameter scale (approximately 340 million parameters) and deep Transformer architecture, which enable it to capture rich linguistic patterns and semantic information during pre-training. Figures 2 and 3 display the model's accuracy and loss trends during training.

As shown in Figure 2, the training accuracy (solid blue line) increased rapidly in the first two epochs, rising from an initial ~95.5% to 98.9%. Although the growth rate slowed afterward, the accuracy continued to improve steadily, eventually surpassing 99.9% by the 5th epoch. The validation accuracy (dashed green line) started at ~98.9% in the first epoch, peaked at approximately 99.3% by the 3rd epoch, and then stabilized, indicating that the model had nearly reached its optimal generalization performance at this stage.

Figure 3 shows the loss curves, where the training loss (solid red line) consistently decreases from an initial value of 0.10 to near zero. In

Table 6
Performance comparison of different models on fake news detection task

Model No.	Model	Accuracy (%)	Weighted F1-score (%)
1	BERT Large (Fine-tuned)	99.33	99.33
2	TextCNN	98.77	98.76
3	RoBERTa (Fine-tuned)	98.00	98.00
4	BERT (CLS) + XGBoost	92.31	92.26
5	Logistic Regression (TF features)	86.46	86.46

Figure 1
Confusion matrix of BERT Large model

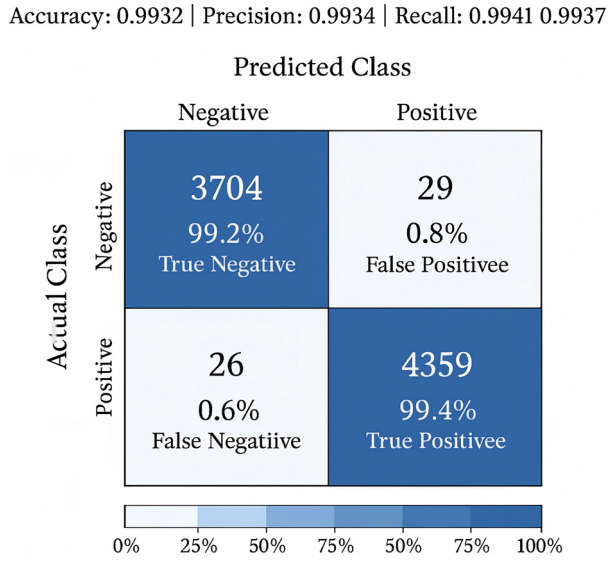


Figure 2
Accuracy trend of BERT Large model

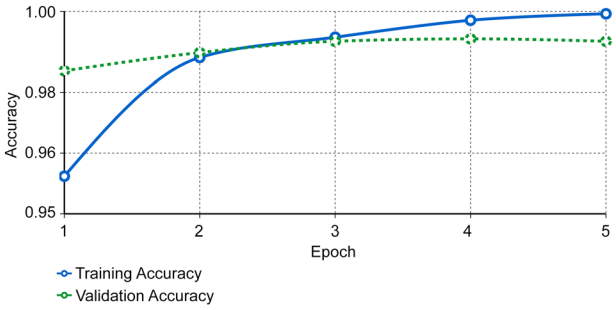
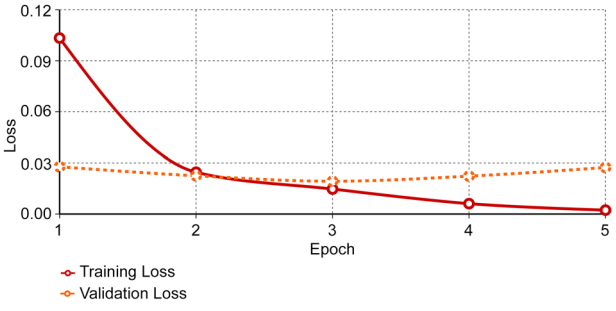


Figure 3
Loss trend of BERT Large model



contrast, the validation loss (dashed orange line) reaches its lowest point at the 3rd epoch before slightly increasing, suggesting mild overfitting in later training stages. However, given the stable validation accuracy, this slight overfitting had a negligible impact on model performance.

4.3.1. Hyperparameter optimization analysis for BERT Large model

To achieve optimal performance, we conducted detailed hyperparameter tuning for the BERT Large model. Through multiple rounds of experiments, we identified the following optimal configuration: the batch size of 16, the learning rate of 1e-5, and six training epochs. Additionally, we implemented these optimization strategies:

Learning rate warmup and decay: A linear warmup was applied for the first 10% of training steps, followed by a linear decay, which helps stabilize the model during initial training and avoids local optima.

Gradient clipping: A maximum gradient norm of 1.0 was set to prevent gradient explosion effectively.

Weight decay regularization: A weight decay parameter of 0.01 was applied to mitigate overfitting.

Early stopping: With patience set to 2, training was stopped when validation loss showed no improvement for two consecutive epochs. Table 7 presents the impact of various hyperparameter settings on the performance of the BERT Large model.

Figure 4 visualizes the impact of two key hyperparameters, learning rate and batch size, on the model performance. It can be observed that the combination of a learning rate of 1e-5 and a batch size of 16 yields the best performance in this task.

4.3.2. Decision threshold optimization analysis

By default, the binary classification model uses a decision threshold of 0.5. We fine-tuned the threshold for the BERT Large model, exploring the range from 0.40 to 0.60. As shown in Table 8, when the threshold was adjusted to 0.53, the model's accuracy on the test set improved slightly from 99.32% to 99.33%.

Figure 5 illustrates the curve of the impact of the decision threshold on the various performance metrics of the model. From the figure, it can be observed that as the threshold value increases from 0.40 to 0.60, the precision rate shows an increasing trend, while the recall rate gradually decreases, which is in line with the typical trade-off relationship between precision rate and recall rate. At a threshold value of 0.53, the precision rate and F1 score reach their optimum, indicating that the model has achieved the best balance between positive and negative category predictions at this point.

Table 7
The impact of hyperparameters on BERT Large model performance

Hyperparameter	Value	Validation accuracy (%)	Training time/ epoch
Learning Rate	1e-4	98.9	780 s
	1e-5	99.3	804 s
	1e-6	98.7	805 s
Batch Size	8	98.8	1135 s
	16	99.3	804 s
	32	99.0	625 s
Epochs	3	99.1	2412 s (total)
	6	99.3	4824 s (total)
	10	99.3	8040 s (total)

Figure 4
Impact of learning rate and batch size on BERT model performance

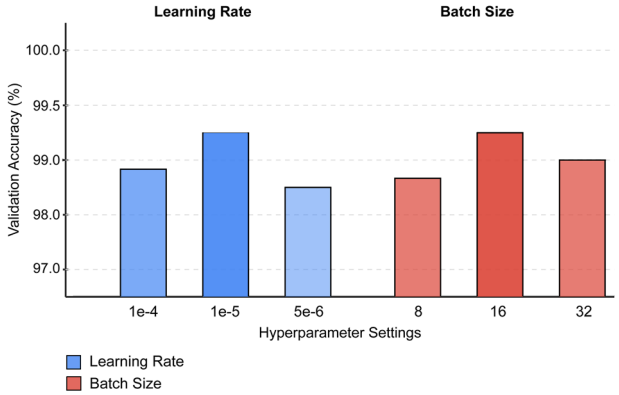
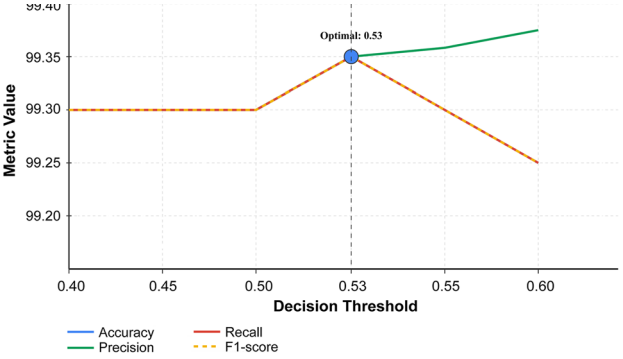


Table 8
Effect of decision threshold on BERT performance

Threshold	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
0.40	99.32	99.30	99.32	99.32
0.45	99.32	99.31	99.32	99.32
0.50	99.32	99.32	99.32	99.32
0.53	99.33	99.33	99.33	99.33
0.55	99.31	99.34	99.30	99.31
0.60	99.28	99.37	99.26	99.28

Figure 5
Performance metrics vs. decision threshold for BERT Large model



4.3.3. Cross-validation result analysis

The results of the stratified 5-fold cross-validation are summarized in Table 9. The model demonstrated high and consistent performance across all folds, with an average accuracy of 99.16% (±0.09%), an average F1-score of 99.16% (±0.09%), and an average ROC-AUC of 99.97% (±0.01%). Additionally, an independent test set evaluation yielded an accuracy of 99.22%, affirming the robustness and reliability of our BERT Large model.

4.4. Model interpretability analysis

To gain a deeper understanding of the decision-making mechanism of the BERT Large model in fake news detection, this study employs

Table 9
Performance of BERT Large on stratified 5-fold cross-validation

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC (%)
1	99.03	99.03	99.03	99.03	99.97
2	99.29	99.29	99.29	99.29	99.98
3	99.08	99.08	99.08	99.08	99.98
4	99.20	99.20	99.20	99.20	99.98
5	99.18	99.18	99.18	99.18	99.96
Mean (±SD)	99.16 (±0.09)	99.16 (±0.09)	99.16 (±0.09)	99.16 (±0.09)	99.97 (±0.01)

the attention visualization technique to analyze the model's internal representation systematically. This interpretive analysis not only helps to reveal the "black box" characteristics of the deep learning model but also provides a theoretical basis for improving the performance and credibility of the model.

4.4.1. Attention visualization methods

In this study, we visualize the multi-level attentional distribution of one real news item and one fake news item, both of which were correctly classified by the model in the test set. We focus on the interaction between [CLS] tokens and other lexical elements in the text at the last level of the model's self-attention mechanism, as the representation of [CLS] tokens directly determine the final classification results. Specifically, we employ three complementary visualizations:

Multi-head attention heatmap: Independently displaying the weight distribution of 16 attention heads, revealing the specialized division of labor among different attention heads in information extraction;

Average attention heatmap: aggregates the weight distribution of all attention heads and highlights word elements with weight >0.01 in red boxes for quantitative comparison;

Attention word clouds: linearly mapping average weights with word metafonts to quickly present model attention focus.

Figure 6 illustrates the comparison of attention distribution between real news and fake news, highlighting the significant difference in weight allocation between the two. It can be observed that the attention distribution of real news is more balanced, while fake news shows extreme concentration.

4.4.2. Analysis of differences in semantic features

1) The real news attention model

The model is characterized by "multi-point synergy" in real news. As shown in Figure 7, the attention weights are distributed over multiple semantic key points, indicating that the model evaluates multiple factual clues to determine the truthfulness of the news. By systematically analyzing these attention distributions, we summarize the following semantic features that receive significant attention, as shown in Table 10.

This distribution of attention shows that the model spreads its attention over multiple verifiable factual markers, forming a discriminative mechanism that corroborates each other. Particularly noteworthy is that the authoritative source marker receives the highest attentional weight, suggesting that the model has learned to use authoritative sources as an indicator of truthfulness.

Figure 8 further visualizes the attention distribution of real news in the form of word clouds, in which the prominence of word elements such as “reuters”, “votes”, and “madrid” verifies that the model pays high attention to authoritative sources and specific event descriptions.

2) Fake news attention model

In contrast to real news, the fake news samples exhibit the characteristic of “single-point over-concentration”. As shown in

Figure 6
Real vs. fake news attention distribution comparison

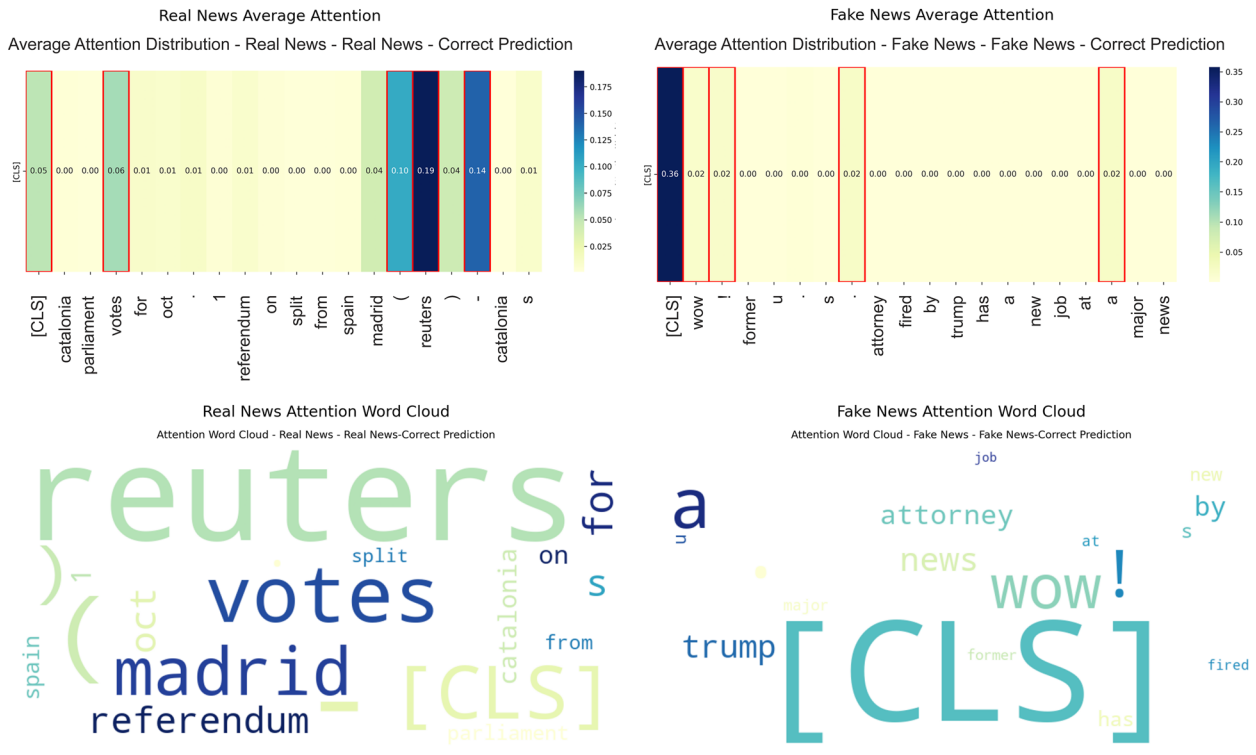


Figure 7
Average attention distribution–real news–real news–correct prediction

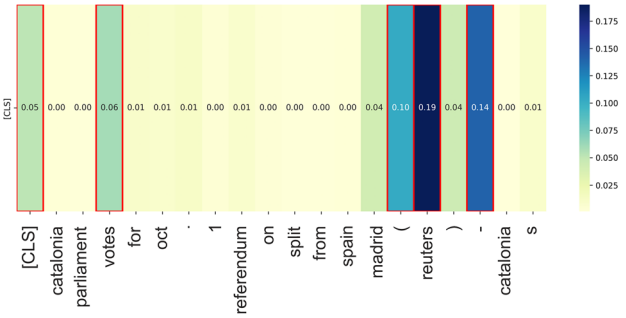


Table 10
Attention-based semantic feature analysis of real news samples

Semantic category	Typical high-weight lexical units	Average weight
Authoritative information source	reuters	0.19
Geographical location marker	madrid, catalonia, "("	0.10–0.14
Event description	votes, referendum	0.05–0.06
Syntactic marker	"(", ")"	0.04–0.10

Figure 8
Attention word cloud–real news–real news–correct prediction

Figure 9
Average attention distribution–fake news–fake news–correct prediction

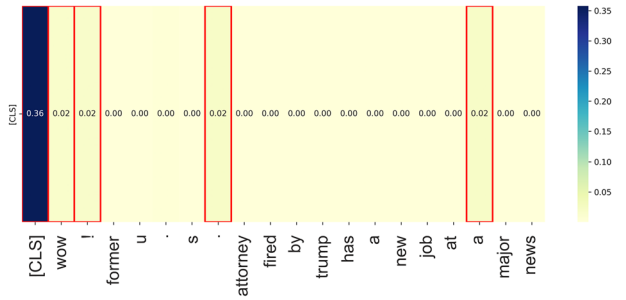


Figure 9, attention is mainly focused on a few lexical elements, especially [CLS] tags. Analyzing this unique pattern of attention distribution, we summarize the main focuses of attention in fake news as shown in Table 11.

Table 11
Attention-based semantic feature analysis of fake news samples

Feature	Typical tokens	Weight
[CLS] Extreme concentration	[CLS]	0.36
Emotional/Exaggerated markers	wow,"!"	0.02
Vague professional terms	attorney, major	0.02
Controversial figures	trump, fired	<0.01

This significant difference suggests that the [CLS] marker itself receives an unusually high attentional weight in the fake news sample, a phenomenon that may reflect the model's "information deficit" in the absence of reliable factual markers. At the same time, the model tends to focus on emotional and hyperbolic markers as potential signals of false content.

The word cloud visualization in Figure 10 visualizes the attention focus of fake news, where the extreme prominence of [CLS] markers and the secondary attention of word elements such as "wow" and "trump" further validate the results of our analysis.

4.4.3. Multi-pronged synergistic mechanisms

The 16 attention heads of the BERT Large model exhibit a clear specialized division of labor, forming a distributed cognitive structure. By analyzing Figures 11 and 12 in detail, we found that.

Header 6 and Header 2: specifically capture authoritative information sources, and the attention weights of "reuters" are as high as 0.44 and 0.41, respectively, indicating that the model has taken the reliability of the information sources as an important indicator for judging authenticity.

Head 1: Focusing on geolocation information in real news scenarios, the attentional weights of "madrid", "spain," and "reuters" are 0.32, 0.13, and 0.17, respectively, to form contextual associations with geographic entities;

Head 12: Exhibits multifocusing characteristics, focusing on both the topic content word "votes" (0.12) and the information source "reuters" (0.29), and constructing topic-source associations for validation;

Head 8: Attentional weight of 0.13 for the classification marker "[CLS]" in the fake news scenario, indicating that this head is directly involved in the final classification decision;

Heads 1, 5, and 15: The word "attorney" shows the same attentional weight (0.07) in fake news, showing a particular sensitivity to legal terms.

Notably, the maximum attention weight triggered by real news (0.44) is significantly higher than that of fake news (0.13), implying that the model has a more deterministic attentional pattern for real content. This multi-head synergistic mechanism enables the model to assess the authenticity of the text from multiple semantic and pragmatic dimensions at the same time, realizing a complex judgment process similar to that of

Figure 10
Attention word cloud–fake news–fake news–correct prediction

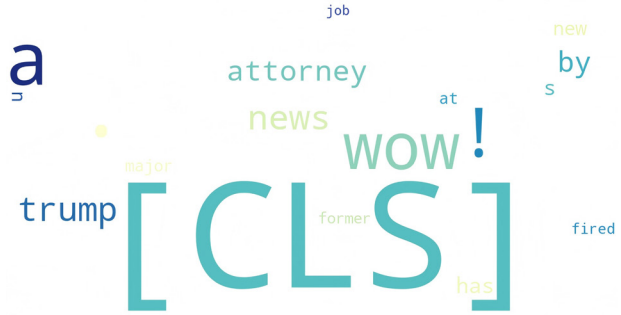
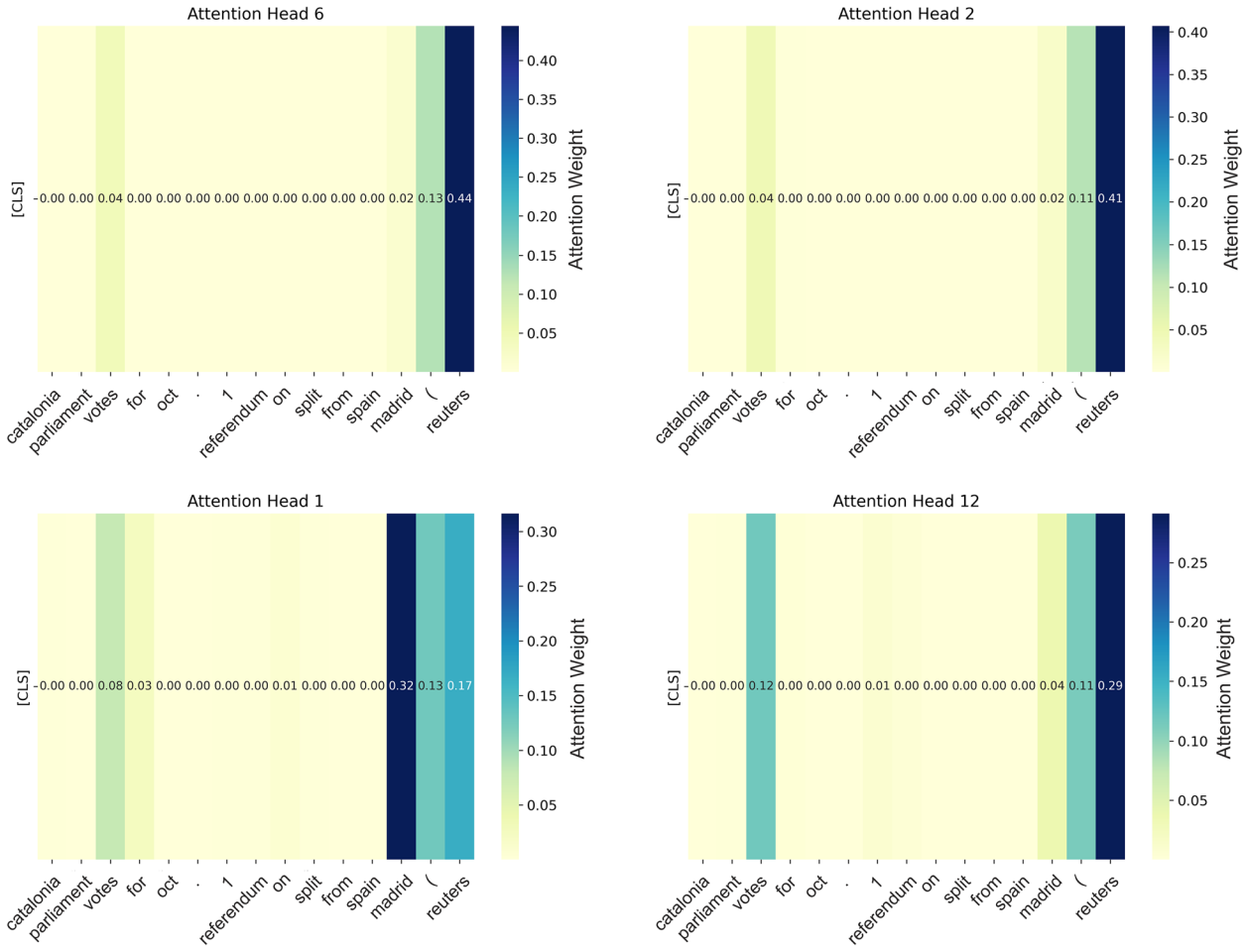


Figure 11
BERT Large attention distribution -real news (predicted: real)



human experts. The specialized division of labor among different heads enhances the model's robustness, enabling it to maintain high accuracy despite insufficient factual cues or strong emotional noise.

Our findings provide an interpretable perspective on Transformer-based news veracity assessment, revealing how attentional mechanisms can distinguish between true and false content by developing specialized attention patterns that are highly consistent with journalistic verification principles.

Figure 13 presents a comprehensive comparison of the attention distribution between real news and fake news, highlighting the key difference between the two models: real news is characterized by a "multi-point synergistic" mechanism, whereas fake news is characterized by a "single-point over-concentration." This comparison provides visual evidence for the decision-making mechanism of the model.

These visualization results offer more profound insight into the model's decision-making process. Specifically, BERT Large assigns higher attention weights to credibility-indicative elements such as named entities (e.g., "Reuters") and geospatial markers (e.g., "Madrid"), suggesting that the model prioritizes factual and contextual cues. In contrast, fake news samples tend to trigger more centralized attention patterns—especially over the [CLS] token or emotionally charged words—implying the model's reliance on superficial cues when clear factual markers are absent. This behavior underscores the model's ability to differentiate between semantically rich versus shallow content structures, which aligns with its high performance in classification accuracy.

4.4.4. Quantitative assessment of attention-based interpretability

To complement the visual attention map presented earlier, we conducted a quantitative analysis using four metrics widely adopted in explainable NLP research (AOPC, Sufficiency, Comprehensiveness, Attention Entropy & Sparsity).

AOPC is a key metric for evaluating the quality of feature importance explanations, assessing the effectiveness of explanations by calculating the area of performance degradation as important features are progressively removed. As shown in Figure 14, we conducted a statistical analysis of the AOPC score distributions for real and fake news, and further employed the Sufficiency and Comprehensiveness metrics to assess the fidelity of attention explanations.

The experimental results show that the AOPC scores of real news are significantly higher than those of fake news ($p < 0.0001$), with medians of 0.25 and 0.065, respectively. This difference indicates that the model forms a more concentrated and meaningful attention pattern when processing real news, while the attention distribution of fake news is relatively dispersed, with weaker ability to locate key information. A high AOPC score indicates that the model can accurately identify the text fragments that are most important for classification decisions.

Sufficiency metric: Measures whether using only features identified as important is sufficient to maintain the model's predictive performance. Experimental results show that fake news performs significantly better than real news on the sufficiency metric (0.91 vs.

Figure 12
BERT attention distribution–fake news (predicted: fake)

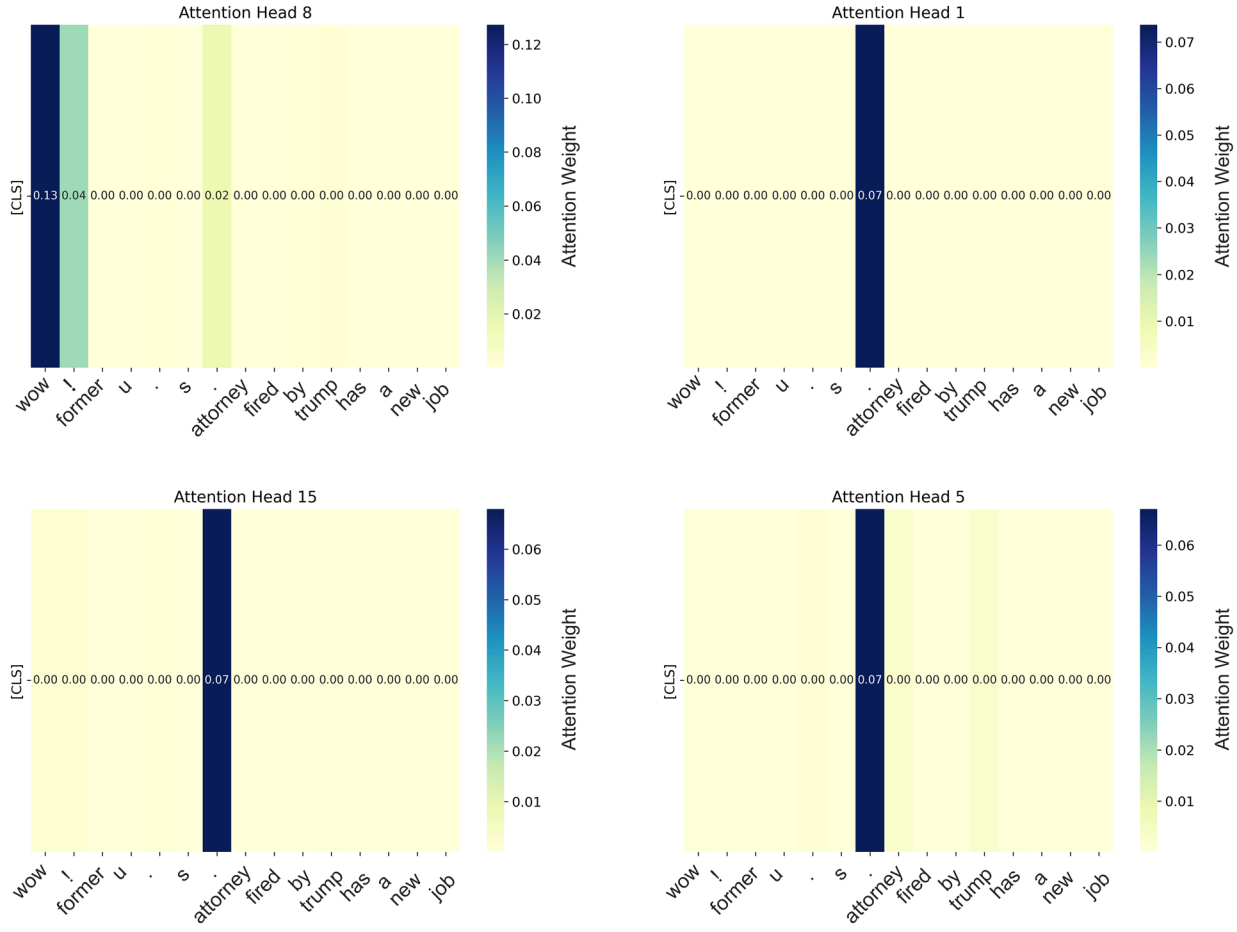
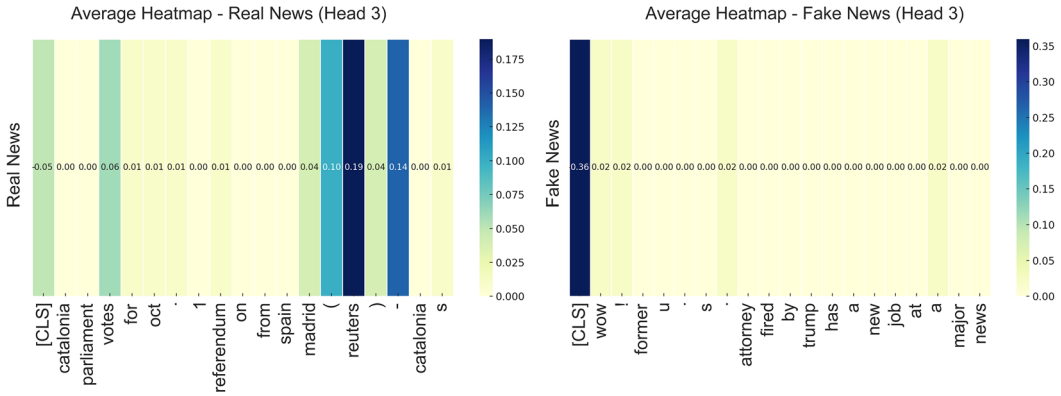


Figure 13
Attention distribution comparison between real and fake news samples (Head 3)



0.36), indicating that fake news detection often relies on a few key linguistic features or patterns.

Comprehensiveness metric: Assesses the extent to which model performance declines after removing important features. Real news performs better on this metric (0.63 vs. 0.045), indicating that real news processes information more comprehensively and requires consideration of a broader range of text features.

To further understand the characteristics of attention distribution, we also calculated the attention entropy and sparsity metrics. As shown in Figure 15, the attention entropy distributions of real and fake news, as well as the relationship between attention sparsity and AOPC scores, are illustrated.

It can be observed that there are significant differences in the attention entropy distributions of real news and fake news. The attention

Figure 14
Attention-based interpretability metrics analysis

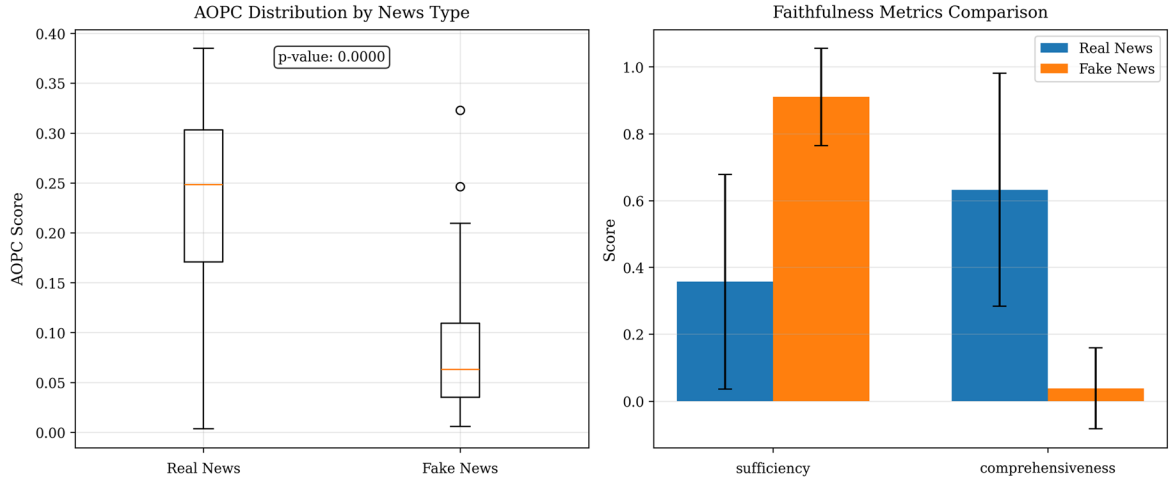
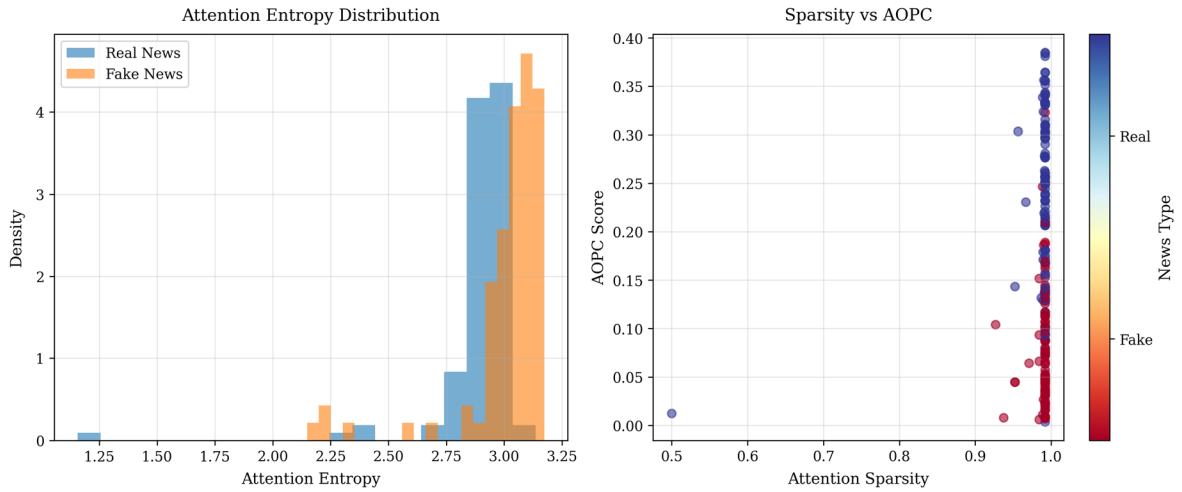


Figure 15
Attention entropy and sparsity analysis



entropy of real news is primarily concentrated in the 2.75–3.0 range, while the distribution of fake news is relatively more dispersed, with a peak in the 3.0–3.25 range. Higher attention entropy indicates that the model distributes attention more evenly when processing fake news, making it difficult to form a clear focus.

From the relationship between attention sparsity and AOPC scores, it can be observed that real news (blue dots) is primarily distributed in the high sparsity, high AOPC region, indicating that the model can effectively focus on key information. Fake news (red dots) exhibits low sparsity and low AOPC characteristics, with a more dispersed attention distribution.

Based on the above quantitative analysis results, we can draw the following conclusions:

Differentiated attention patterns: Real news and fake news exhibit significant differences in attention distribution, providing quantitative support for the model's interpretability.

Asymmetry in explanation quality: Real news detection achieves higher AOPC scores and comprehensiveness, indicating that the model can form more reliable and comprehensive explanations when processing real news.

Differences in feature-dependent strategies: Fake news detection tends to rely on a few key features (high sufficiency), while real news detection requires more comprehensive information integration (high comprehensiveness).

Attention focusing capability: The positive correlation between high sparsity and high AOPC confirms the effectiveness of the model's attention mechanism, particularly in the task of detecting real news.

These quantitative evaluation results not only validate the effectiveness of our model's attention mechanism but also provide quantitative evidence for understanding the fundamental differences between real news and fake news in terms of linguistic features.

4.4.5. Limitations and prospects

Although attentional visualization provides important insights for model decision-making, there are still some limitations:

Attention ≠ causation: high weight does not necessarily represent a deterministic feature and needs to be cross-validated with a combination of gradients, probe networks, and other methods;

Granularity limitation: the current analysis focuses on the lexical meta-level and does not yet fully capture inter-sentence and paragraph-level semantic relationships;

Domain adaptation: The above attention model is mainly based on political news samples, and migration to domains such as finance and health still requires systematic evaluation.

Future research can be deepened in three aspects: (1) fusing multi-dimensional interpretation means such as gradient accumulation and integrated probes; (2) designing interpretability-driven distillation and fine-tuning strategies; and (3) constructing a hybrid detection framework that combines expert knowledge and attention analysis to enhance model transparency and practicality.

In summary, explain ability analysis conducted through attention visualization indicates that BERT-Large prioritizes specific keywords related to source credibility. Attentional visualization provides transparent evidence of BERT Large's decision-making process in fake news detection, laying the methodological foundation for the development of a more trustworthy and auditable AI system.

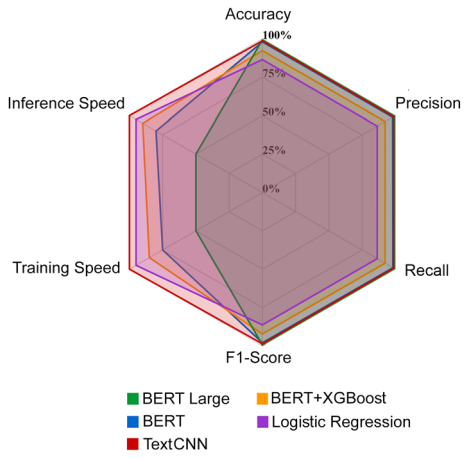
4.5. Comprehensive performance comparison of multiple models

To visualize the performance of each model on different performance metrics, Figure 16 presents a comprehensive comparison of the five models on six key metrics in the form of a radar chart.

It is visible in the radar chart:

1) The BERT Large model (green line) demonstrates its superior classification ability by leading across the board in all four classification performance metrics: accuracy, precision, recall and F1 score, forming the outermost polygon. However, in terms of training speed and inference speed, the performance is relatively weak due to the large number of parameters.
2) The TextCNN model (red line) is second only to BERT Large in classification performance, with an accuracy of 98.7%. However, its most significant advantage lies in the training and inference speed, especially since its inference speed is outstanding, attributed to its concise convolutional network structure and parallel computing capability.
3) The BERT/RoBERTa model (blue line) performs well overall, with balanced indicators and an accuracy of 98.0%. Its performance is slightly lower than that of BERT Large, but higher than other models, reflecting the strong ability of the basic pre-trained model.
4) The BERT(CLS) + XGBoost model (yellow line) performs next best, with an accuracy of about 92.3%. While this feature extraction and traditional machine learning approach has some advantages in inference speed, the classification performance is significantly lower than that of the end-to-end deep learning approach.
5) The logistic regression model (purple line) exhibits the weakest performance across all classification metrics, with an accuracy of approximately 86.5%. Despite its fast training and inference speed, the simple word frequency features are challenging to capture the complex semantic and contextual information in the text, leading to its limited performance on the fake news detection task.

Figure 16
Model performance comparison



4.5.1. Error analysis and case studies

To gain a deeper understanding of the strengths and weaknesses of each model, we conducted an error analysis of BERT Large and logistic regression, the two models with the most significant performance differences, by randomly selecting 10 samples from the test set where the two judgments were inconsistent. Table 12 shows some of the results of the case studies.

Based on the error analysis, we classified model errors into four categories: hyperbolic description misclassification, sentiment bias misclassification, factual inconsistency misclassification, and linguistic complexity misclassification. Figure 17 shows the distribution of each model on these error types.

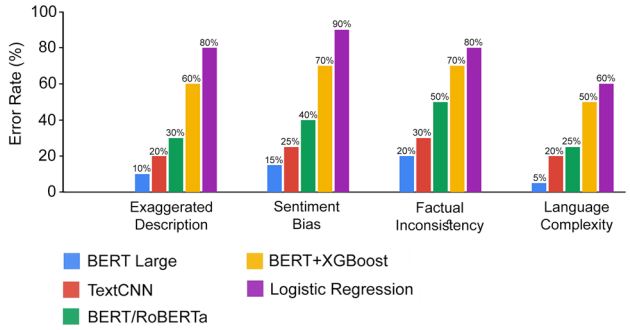
As shown in Figure 17:

1) The BERT Large model exhibits a lower misclassification rate across all error types, particularly in cases of high linguistic complexity, with a misclassification rate of only 5%, which is significantly lower than that of other models. This indicates that the Large pre-trained model has stronger language comprehension ability and can parse complex linguistic expressions correctly.
2) The TextCNN model performs better in factual inconsistency and linguistic complexity misclassification but has a relatively high misclassification rate of 25% when dealing with sentiment-biased cases. This may be because CNN focuses more on local features and has a limited understanding of sentiment transitions within a given context.
3) Traditional models (e.g., logistic regression) perform poorly across all error types, particularly in sentiment bias misclassification, where

Table 12
Model error case analysis examples

News text excerpt	True label	BERT Large prediction	Logistic regression prediction	Analysis
"Study claims global warming will cause sea levels to rise 10 m..."	Fake (0)	Fake (0)	Real (1)	Logistic regression focuses on keywords "study", "global warming" while ignoring exaggerated values
"Antarctic ice cap suddenly melts significantly within 24 hours..."	Fake (0)	Fake (0)	Real (1)	Logistic regression cannot understand the unreasonable relationship between "24 hours" and "significant melting"
"NASA's latest detection shows liquid water exists on Mars..."	Real (1)	Real (1)	Fake (0)	Logistic regression may consider the combination of "Mars" and "liquid water" as indicators of fake news

Figure 17
Error distribution analysis across models



the error rate reaches as high as 90%, indicating that it is challenging to comprehend the nuances and transitions in sentiment expression.

4.5.2. Computing resources and performance balance analysis

Considering the limitation of computational resources in practical application scenarios, we have comparatively analyzed the resource consumption and performance of each model, as shown in Table 13.

To visualize the trade-off between the model's accuracy and resource consumption, we constructed the efficiency analysis graph shown in Figure 18. The horizontal axis of the graph represents the resource consumption index, which combines the number of parameters, training time, and memory occupation, and the vertical axis represents the accuracy rate.

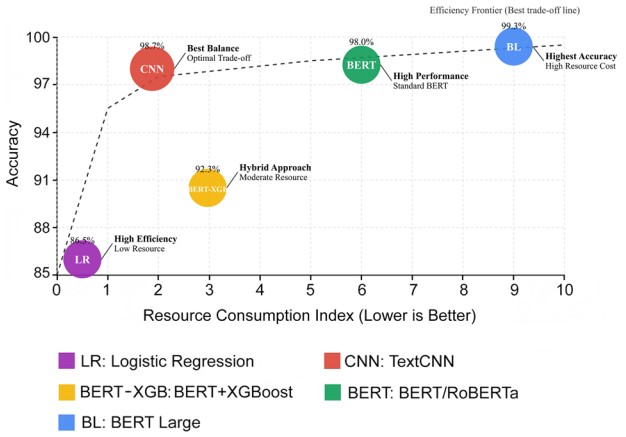
As shown in Figure 18:

1) The TextCNN model is positioned near the efficiency frontier, representing the optimal balance between accuracy and resource consumption. With 98.7% accuracy and very low resource consumption, the model offers significant advantages in real-world deployments.
2) Although the BERT Large model achieves the highest accuracy rate, it also consumes the largest amount of resources. It is best suited for scenarios that require extremely high accuracy rates and substantial resources.
3) Although the logistic regression model has the lowest resource consumption, its accuracy rate lags significantly behind and is not suitable for application scenarios with high accuracy requirements.

Table 13
Model computational resources and performance comparison

Model	Parameters	Training time (total)	Inference speed (/1000 items)	Accuracy	Memory usage
BERT Large	340 M	4824 s	118 s	99.3%	5.2 GB
TextCNN	14.76 M	130 s	1 s	98.7%	0.3 GB
BERT/ RoBERTa	110 M	1005 s	87 s	98.0%	1.8 GB
BERT(-CLS) + XGBoost	-	~500 s	2 s	92.3%	1.5 GB
Logistic Regression	<10 K	~10 s	<1 s	86.5%	0.1 GB

Figure 18
Model efficiency analysis: accuracy vs. resource consumption



4) The BERT/RoBERTa and BERT + XGBoost models are below the efficiency frontier, indicating that they have room for improvement in resource utilization efficiency.

In addition, we evaluated computational costs and found that TextCNN, with its smaller architecture, achieved significantly faster inference speeds compared to BERT Large (0.06 ms per input vs. 10.32 ms, a 172-fold improvement in speed). BERT Large achieved the highest accuracy (99.33%), while TextCNN demonstrated highly competitive performance (98.60%), with significantly reduced computational requirements, making it more suitable for resource-constrained environments or real-time applications.

5. Discussion

This study systematically evaluates and compares deep learning models for fake news detection, with a particular focus on the balance between accuracy, efficiency, and interpretability. The BERT Large model emerged as the most accurate, achieving an impressive 99.33% accuracy due to its extensive parameterization and powerful semantic representation capabilities. These results significantly surpass traditional models, such as logistic regression (86.46%), and hybrid approaches, like BERT + XGBoost (92.31%). The TextCNN model, despite its simpler architecture and fewer parameters, demonstrated a competitive accuracy of 98.77%, highlighting its utility in resource-constrained environments and validating findings from previous research that emphasized the practicality of lightweight CNN-based models [15, 24].

Contextualizing these findings within the existing literature, our results align with prior studies emphasizing the superior semantic comprehension of transformer-based models (e.g., BERT and RoBERTa) compared to shallow machine learning approaches [16, 18]. The robust performance of BERT Large corroborates the established effectiveness of fine-tuning large-scale pre-trained models for domain-specific tasks, particularly in fake news detection [17, 23]. The TextCNN results also confirm the value of capturing local textual features efficiently, reinforcing conclusions drawn from previous literature on convolutional architectures [24].

Despite significant advances demonstrated here, several limitations must be acknowledged. The dataset's potential bias, primarily focused on political news, may limit its generalizability across other domains, such as health or finance. Additionally, despite employing advanced threshold optimization techniques, imbalances in news categories remain a

challenge. Future studies may explore cross-domain adaptation strategies and develop hybrid architectures that incorporate multimodal data to enhance robustness against diverse and multimodal fake news content [11]. For example, these findings could inform the development of improved filtering algorithms for social media platforms or help create fact-checking tools.

The implications of this study are substantial for both academia and industry. Practically, platforms handling vast amounts of user-generated content can significantly benefit from deploying BERT Large when computational resources permit. Conversely, environments with constrained computational capabilities, such as mobile or edge computing scenarios, can effectively utilize TextCNN to achieve near-state-of-the-art performance with minimal resource requirements. Furthermore, the interpretability analysis through attention visualization provides crucial insights into model decision-making, increasing transparency and trustworthiness in critical application scenarios, thus addressing ethical concerns prevalent in AI deployments.

The findings of this study have direct implications for real-world systems, particularly in the design of automated content moderation tools, misinformation flagging systems, and AI-assisted fact-checking platforms. For example, platforms like Twitter/X or Facebook could integrate models like BERT Large for high-accuracy offline verification pipelines. At the same time, TextCNN could be deployed in real-time detection modules on mobile or edge devices due to its computational efficiency. Furthermore, attention-based interpretability could be embedded into transparency dashboards to help moderators and end-users understand why content is flagged, enhancing trust and accountability in AI-driven decision systems.

Recent developments in generative AI present compelling opportunities for enhancing the security and privacy of mobile crowdsensing (SPPMCS) systems through the generation of synthetic data, thereby mitigating risks related to data leakage and malicious attacks. Yang et al. [27] investigate this integration by identifying core challenges, proposing targeted solutions, and introducing a generative AI-based framework for data protection, supported by simulation-based validation. Complementing this work, Yang et al. [28] explore the application of attention mechanisms within mobile crowdsensing (MCS), demonstrating their effectiveness in optimizing task allocation, privacy safeguards, and data transmission. Their attention-driven framework significantly improves network performance in large-scale MCS environments and outlines key directions for future research. Collectively, these studies highlight the transformative potential of advanced AI methods—such as generative and attention-based models—for building secure, interpretable, and high-performance systems, offering relevant parallels to the optimization and interpretability strategies employed in our fake news detection framework.

This research engages deeply with ongoing concerns about algorithmic opacity and public accountability. Building on the framework proposed by Torabi Asr and Taboada [29], who emphasize the importance of both automated detection and public education, we argue for the development of interactive, AI-driven media literacy tools that simulate fake news detection using models such as BERT. These tools can enhance user trust and foster critical thinking by revealing the linguistic cues and decision processes that AI systems rely on to flag deceptive content. Embedding such tools in school curricula, public libraries, or even within social media platforms could cultivate cognitive resilience and empower individuals in the face of increasingly sophisticated misinformation campaigns.

Reflecting on the initial research objectives presented in the Introduction, this study successfully addressed key technical challenges identified, including model generalization, computational efficiency, and interpretability. Our comparative analysis framework and systematic optimization strategies provided practical solutions and benchmarks that are useful for further research.

While this study focuses primarily on fake news in the political domain, the proposed models—particularly BERT Large and TextCNN—are inherently domain-agnostic due to their underlying language modeling capabilities. This allows for their extension to other domains such as health, finance, and science misinformation. To support future real-world deployment, we plan to conduct cross-domain evaluations using datasets from diverse fields. This line of investigation will help assess model robustness, adaptivity, and generalizability across heterogeneous content environments.

In conclusion, this research substantiates the superior capability of transformer-based pre-trained models in detecting fake news, particularly highlighting the trade-off between model complexity and operational efficiency. TextCNN provides a viable alternative in efficiency-sensitive contexts. Future research directions include enhancing multimodal integration, exploring adversarial robustness, and refining interpretability techniques, thereby moving towards the development of comprehensive and practical fake news detection solutions that are adaptable across various operational contexts.

6. Conclusion

This study demonstrates the substantial effectiveness of deep learning models, particularly BERT Large and TextCNN, in accurately detecting fake news, achieving remarkable accuracy rates of 99.33% and 98.77%, respectively. By systematically comparing various pretrained and lightweight models, the research fills a significant gap in evaluating the trade-offs between accuracy, efficiency, and interpretability while also introducing a comprehensive framework for hyperparameter optimization and interpretability analysis. Nonetheless, the research acknowledges limitations related to potential dataset biases, primarily centered on political news, which may affect generalizability. These limitations, however, present valuable avenues for future investigations, including multimodal content integration and cross-domain adaptability. Ultimately, this research provides a robust methodological foundation for developing more accurate and transparent fake news detection systems, thereby significantly enhancing public trust and informed decision-making in information-intensive environments, such as social media platforms and news dissemination services.

Recommendations

The finding revealed that the lack of training for both teachers and students was the main factor that prevented them from using educational technology tools in teaching and learning Ecology. Therefore, training on educational technology for both teachers and students is recommended. Since educational technology tools have arisen excitement and curiosity amongst students, they recommended other module tutors to use educational technology tools as well. Educational technology tools integrated in the module will be further replicated by student's teacher during teaching practice or as a full fledge teacher. Therefore, tutors were recommended to use variety of educational technology tools in learning, teaching and an assessment.

Acknowledgement

The authors would like to extend their gratitude to Khon Kaen University for their support through Announcement No. 2580/2563 on "The Criteria for the Acceptance of an Inbound Visiting Scholar from a Foreign Institution or Organization to Khon Kaen University". This support has been invaluable to the success of our research project.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data are available on request from the corresponding author upon reasonable request.

Author Contribution Statement

Qiuping Li: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Fen Fu:** Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Yinjuan Li:** Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Bhunnisa Wisassinthu:** Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Wirapong Chansanam:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Tossapon Boongoen:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ..., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

[2] Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

[3] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

[4] Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, *4*(5), 472–480. https://doi.org/10.1038/s41562-020-0833-x

[5] Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, *9*(1), 104–117. https://doi.org/10.1017/xps.2020.37

[6] Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology*, *10*(3), 21. https://doi.org/10.1145/3305260

[7] Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, *53*(5), 109. https://doi.org/10.1145/3395046

[8] Oshikawa, R., Qian, J., & Wang, W. Y. (2020). A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 6086–6093.

[9] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

[10] Pierri, F., & Ceri, S. (2019). False news on social media: A data-driven survey. *ACM SIGMOD Record*, *48*(2), 18–27. https://doi.org/10.1145/3377330.3377334

[11] Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). MVAE: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, 2915–2921. https://doi.org/10.1145/3308558.3313552

[12] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[13] Altunbey Özbay, F., & Alatas, B. (2020). Fake news detection within online social media using supervised artificial intelligence algorithms. *Physica A: Statistical Mechanics and its Applications*, *540*, 123174. https://doi.org/10.1016/j.physa.2019.123174

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

[15] Ouassil, M.-A., Cherradi, B., Hamida, S., Errami, M., El Gannour, O., & Raihani, A. (2022). A fake news detection system based on combination of word embedded techniques and hybrid deep learning model. *International Journal of Advanced Computer Science and Applications*, *13*(10), 525–534. https://doi.org/10.14569/ijacsa.2022.0131061

[16] Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, *80*(8), 11765–11788. https://doi.org/10.1007/s11042-020-10183-2

[17] Gagiano, R., Kim, M. M.-H., Zhang, X., & Biggs, J. (2021). Robustness analysis of Grover for machine-generated news detection. In *Proceedings of the 19th Annual Workshop of the Australasian Language Technology*, 119–127.

[18] Essa, E., Omar, K., & Alqahtani, A. (2023). Fake news detection based on a hybrid BERT and LightGBM models. *Complex & Intelligent Systems*, *9*(6), 6581–6592. https://doi.org/10.1007/s40747-023-01098-0

[19] Chen, C., Zhang, P., Zhang, H., Dai, J., Yi, Y., Zhang, H., & Zhang, Y. (2020). Deep learning on computational-resource-limited platforms: A survey. *Mobile Information Systems*, *2020*(1), 8454327. https://doi.org/10.1155/2020/8454327

[20] Mazzeo, V., Rapisarda, A., & Giuffrida, G. (2021). Detection of fake news on COVID-19 on web search engines. *Frontiers in Physics*, *9*, 685730. https://doi.org/10.3389/fphy.2021.685730

[21] Przybyla, P. (2020). Capturing the style of fake news. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(01), 490–497. https://doi.org/10.1609/aaai.v34i01.5386

[22] Keya, A. J., Shajeeb, H. H., Rahman, M. S., & Mridha, M. F. (2023). FakeStack: Hierarchical Tri-BERT-CNN-LSTM stacked model for effective fake news detection. *PLoS ONE*, *18*(12), e0294701. https://doi.org/10.1371/journal.pone.0294701

[23] Reza, S., Ferreira, M. C., Machado, J. J., & Tavares, J. M. R. (2022). A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent

neural networks. *Expert Systems with Applications*, *202*, 117275. https://doi.org/10.1016/j.eswa.2022.117275

[24] Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1746–1751. https://doi.org/10.3115/v1/d14-1181

[25] Jane, M. (2025). Evaluating the performance of BERT, XGBoost, and hybrid models for fake news detection. *ResearchGate*. https://www.researchgate.net/publication/389499023

[26] Pradeepa, S., Jomy, E., Vimal, S., Hassan, M. M., Dhiman, G., Karim, A., & Kang, D. (2024). HGATT_LR: Transforming review text classification with hypergraphs attention layer and logistic regression. *Scientific Reports*, *14*(1), 19614. https://doi.org/10.1038/s41598-024-70565-6

[27] Yang, Y., Zhang, B., Guo, D., Du, H., Xiong, Z., Niyato, D., & Han, Z. (2024). Generative AI for secure and privacy-preserving mobile crowdsensing. *IEEE Wireless Communications*, *31*(6), 29–38. https://doi.org/10.1109/MWC.004.2400017

[28] Yang, Y., Du, H., Xiong, Z., Niyato, D., Jamalipour, A., & Han, Z. (2025). Enhancing wireless networks with attention mechanisms: Insights from mobile crowdsensing. *IEEE Wireless Communications*, *32*(5), 152–161. https://doi.org/10.1109/MWC.003.2400231

[29] Torabi Asr, F., & Taboada, M. (2019). Big data and quality data for fake news and misinformation detection. *Big Data & Society*, *6*(1), 2053951719843310. https://doi.org/10.1177/2053951719843310

How to Cite: Li, Q., Fu, F., Li, Y., Wisassinthu, B., Chansanam, W., & Boongoen, T. (2025). Fake News Detection with Deep Learning: Insights from Multi-dimensional Model Analysis. *Journal of Computational and Cognitive Engineering*, *4*(4), 603–619. https://doi.org/10.47852/bonviewJCCE52026051