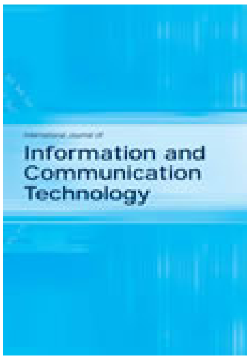


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Precise identification and traceability of fake e-commerce reviews integrating multimodal semantic understanding

Bing Duan

DOI: 10.1504/IJICT.2026.10077367

Article History:

Received:	09 December 2025
Last revised:	04 January 2026
Accepted:	06 January 2026
Published online:	17 April 2026

Precise identification and traceability of fake e-commerce reviews integrating multimodal semantic understanding

Bing Duan

Department of Finance and Business,
Qinghai Higher Vocational and Technical Institute,
Qinghai, 810799, China
Email: DB198012@163.com

Abstract: Aiming at the problem that the precise identification and traceability methods for false e-commerce reviews are difficult to comprehensively capture the complex semantics and potential features in the reviews, this paper first uses a text convolutional neural network and a pre-trained model to extract features from multi-dimensional text semantics respectively. By further integrating the features of the reviewers, the model's understanding of the overall semantics is further enhanced. The images posted by users in the comments are subjected to feature extraction using the residual network to obtain the corresponding visual semantic features. Subsequently, multimodal semantic feature fusion will be carried out to identify false comments. The experimental results on large-scale datasets show that the recognition and traceability accuracy rates of the proposed method have increased by at least 0.7% and 1.1% respectively, significantly outperforming the existing methods in the recognition and traceability of false comments.

Keywords: e-commerce fake reviews; identification and traceability; multimodal semantics; feature fusion; transformer model.

Reference to this paper should be made as follows: Duan, B. (2026) 'Precise identification and traceability of fake e-commerce reviews integrating multimodal semantic understanding', *Int. J. Information and Communication Technology*, Vol. 27, No. 35, pp.81–102.

Biographical notes: Bing Duan is a Lecturer at Qinghai Higher Vocational and Technical Institute. She obtained her Master's in Agriculture from Qinghai University in 2004, and later a Master's in Management from Qinghai Nationalities University in 2013. She has published eight papers. Her research interests include business administration and marketing.

1 Introduction

With the rapid growth of online users and merchants, user feedback systems have become an integral component of e-commerce platforms, serving to improve consumer experiences and foster healthy platform development. As this trend continues, the total volume of consumer reviews and feedback information has increased rapidly. However, these feedback data is highly valuable for merchants as they can serve as critical references for consumer decision-making and help merchants better understand customer needs, thereby more effectively meeting them (Li et al., 2024). Product reviews exert a substantial influence on consumers' purchasing decisions, effectively steering their consumption behaviours. A high volume of positive feedback can yield considerable commercial benefits for merchants. Consequently, some sellers resort to employing professional fake reviewers to secure undue profits. As online reviews grow in reach and impact, deceptive practices are proliferating, making fake reviews an escalating and widespread concern (Luo et al., 2023). By fabricating authentic-looking evaluations and disseminating misleading content, these actors erode consumer trust in review systems, often misleading unsuspecting buyers (Alsharif, 2022). The prevalence of such fraudulent information in today's digital landscape not only impedes the integrity of e-commerce ecosystems and distorts academic research, but also inflicts direct harm on consumer interests (Mughal et al., 2025). Thus, there is a pressing need for a robust and reliable framework capable of accurately detecting and tracing fake reviews, so as to mitigate their negative consequences and support sustainable platform development.

2 Related work

Early methods for identifying fake reviews in e-commerce primarily relied on analysing semantic features within review texts. These semantic features extracted from the textual content of fake reviews, include text length, vocabulary, and syntactic characteristics. The identification of fake reviews was treated as a binary classification problem. Zhang et al. (2020) proposed an e-commerce fake review detection method based on the TrustRank algorithm and decision trees. This approach more accurately captures the connectivity and information flow between text nodes, thereby enhancing detection accuracy. Vijayaragavan et al. (2020) employed a supervised learning support vector machine (SVM) model to distinguish between short and long Amazon reviews, achieving an accuracy rate of 88%. Alsubari et al. (2022) proposed a model integrating feature-based and graph-based techniques, employing SVM as the primary tool for fake review detection while converting adjacent node topology information into latent representation spaces. Baishya et al. (2021) utilised supervised machine learning to analyse sentiment features for fake review identification. Fan (2024) proposed a fake review detection method integrating domain expertise from the hotel industry with clustering algorithms, though it exhibits high computational complexity. Purifyregalia et al. (2025) introduced a hybrid model combining SVM and decision trees for classifying product reviews, significantly improving detection accuracy.

When data dimensions are high, early studies face the curse of dimensionality, where the number of variable configurations and features grows exponentially with the number of variables. Furthermore, since pre-designed discrete features cannot encode the entire review text information, researchers introduced deep neural network technology for multimodal feature fusion to detect fake reviews. In the dynamic analysis of fake reviews in e-commerce, deep learning technology constructs multi-layer neural network models to automatically extract textual features from reviews, user behaviour patterns, and temporal sequence correlations. This enables dynamic tracking of the propagation paths, velocity, and impact scope of fake reviews. Its core advantages lie in efficiently capturing complex features, adapting to dynamic environments, and reducing manual intervention costs. Qayyum et al. (2023) employed convolutional neural network (CNN) and long short-term memory (LSTM) to extract textual and visual semantic features, respectively.

By utilising a multi-layer perceptron module to predict source nodes after identifying candidate rumour source clusters, they substantially reduced the search space and improved efficiency. He et al. (2023) designed a dual-layer attention architecture to acquire multimodal review semantic data through specific attention mechanisms. They first employed n-gram syntax and CNNs to mine multidimensional meanings of review words. Abd-Alhalem et al. (2024) utilised recurrent neural network (RNN) and LSTMs to identify fake e-commerce reviews, leveraging natural language processing techniques and review content dictionaries to extract deep semantic features. Hu et al. (2025) proposed a novel method using graph convolutional networks (GCNs) with integrated multimodal semantic similarity to detect fake reviews. This approach combines textual and visual semantic information, significantly improving detection accuracy. Thuy et al. (2024) introduced a new model to efficiently learn distinct view representations of users and fuse learned features through a discriminative fusion mechanism to predict fake review labels. Hou et al. (2025) employed attention mechanisms to fuse multimodal semantic features and applied a Markov random field model for fake review detection. Experimental results demonstrate that integrating textual and visual semantic features enhances the accuracy of fake review identification.

Table 1 Analysis of the advantage and disadvantages of different research methods

Method category	*Key technologies*	*Advantages*	*Disadvantages*
Text analysis method	Machine learning (SVM), deep learning (CNN, RNN, BERT)	The model is relatively mature and effective in detecting false comments in plain text	The mode is single, making it difficult to deal with high-quality false texts. Ignoring key information such as images can easily lead to bypassing.
Multimodal fusion method	Feature concatenation, attention mechanism, cross-modal alignment	By comprehensively utilising multi-source information, the recognition accuracy and robustness are higher	Most of them stop at identification and lack the ability to trace the source. The feature fusion strategy may not be deep enough.
Tracing and sourcing methods	Graph neural networks, cluster analysis	Be capable of identifying groups that engage in collaborative cheating and cracking down on organised fraud	Over-reliance on metadata and failure to conduct in-depth traceability by leveraging the semantics of comment content.

For the research on the origin of false reviews in e-commerce, researchers mainly focus on the dynamic analysis of dissemination in deep learning. The GCNSI model proposed by Chai et al. (2021) innovatively transforms the topological information of adjacent nodes into potential representation Spaces, thereby converting the complex traceability problem into an operational node classification problem, providing an important idea for subsequent research. Shi et al. (2022), from the perspective of information dissemination dynamics, innovatively modelled the traceability problem as the reverse process of information diffusion and proposed a reversible validity perception graph diffusion model, effectively enhancing the traceability accuracy. Singh et al. (2024) innovatively integrated multiple tasks such as rumour source identification, rumour news detection, and popularity prediction in a joint learning framework. They predicted the source nodes after identifying candidate rumour source clusters through a multi-layer perceptron module, significantly reducing the search space and improving efficiency.

Research on the precise identification and traceability of e-commerce false reviews integrating multimodal semantic understanding has made certain progress. Different methods have their own advantages and disadvantages, as shown in Table 1.

From the above analysis, it can be seen that most current studies focus only on text semantic modalities or use relatively basic behavioural features, which are unable to fully and comprehensively mine the complex semantics and potential characteristics embedded in reviews. To address these issues, this paper proposes a precise identification and tracing method for e-commerce fake reviews based on multi-modal semantic understanding. This method has the following characteristics.

1 This paper introduces a novel multimodal framework that seamlessly integrates textual, visual, and reviewer-behavioural features. By leveraging TextCNN, pre-trained language models, and Residual Networks, our approach captures complex semantic patterns that unimodal methods often miss, leading to a more robust identification of fake reviews.

2 In light of the e-commerce fake review identification results, this paper uses a Transformer model for tracing e-commerce fake reviews, which can better capture long-range dependencies between source nodes and fake nodes, thus improving tracking performance.

3 Simulation experiments are conducted on real datasets. The results show that the proposed method improves the F1 of identification and tracing by at least 3.2% and 2.4%, respectively. Using a multi-modal feature fusion approach to supplement text semantics for identifying fake reviews is effective, as it can effectively improve overall detection accuracy for fake reviews.

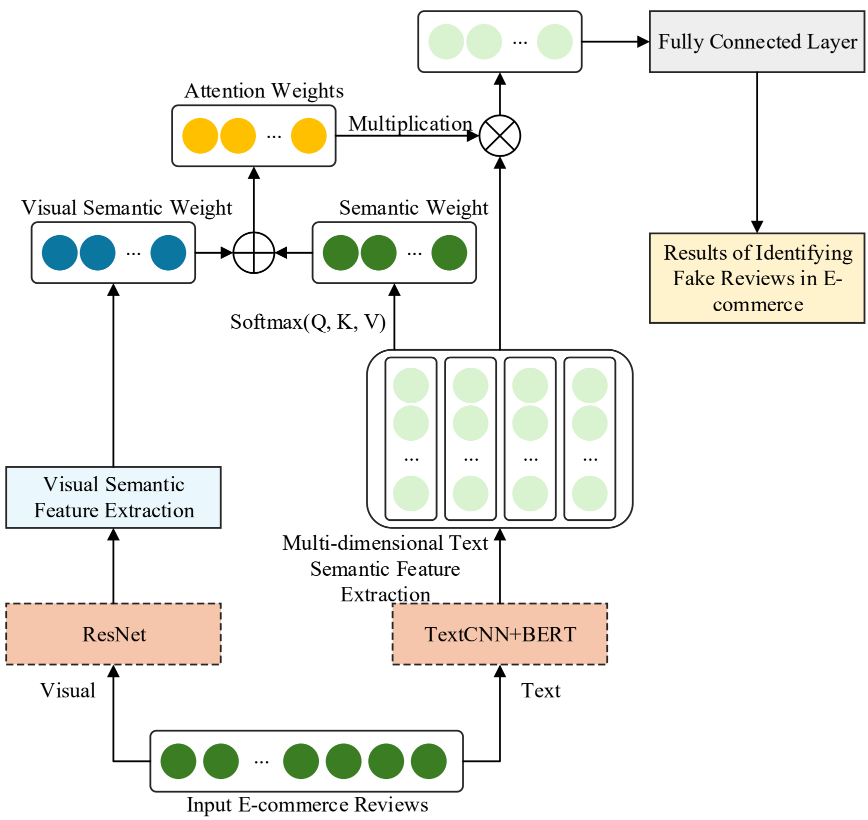
3 Fake review detection in e-commerce based on visual and text semantic fusion

3.1 Multi-modal semantic feature extraction for e-commerce fake reviews

User reviews on e-commerce websites are often considered direct sources of information for user experiences. Therefore, as an interaction feedback product from consumers after shopping, most user reviews on e-commerce platforms have certain reference value for consumers. According to existing research (Du and Jiang, 2025), current e-commerce user reviews mainly consist of two parts: review texts and review images, in which the review text contains three elements: linguistic style, semantic information, and sentiment orientation. Review images are pictures added by users when commenting, which can provide supplementary explanations for textual information. Existing research mostly focuses on unimodal semantics for identifying fake e-commerce reviews, while ignoring the multimodal semantics of e-commerce reviews, leading to unsatisfactory identification accuracy. To address the above issues, this paper proposes a multi-modal semantic fusion-based method for identifying fake e-commerce reviews called EFRMSF. The model structure of this method is shown in Figure 1.

The EFRMSF method employs text CNNs and BERT pre-trained models to extract features from text comment information, yielding corresponding feature vectors. The model incorporates reviewer characteristics by first processing the review text and the reviewer ID separately, then concatenating their output features. This process enriches the model's semantic understanding with reviewer context. Images posted by users in reviews undergo feature extraction via residual networks to obtain corresponding visual features. Finally, multimodal fusion of textual and visual features enables the identification of fraudulent e-commerce reviews.

Figure 1 Structure of the fake review detection model (see online version for colours)

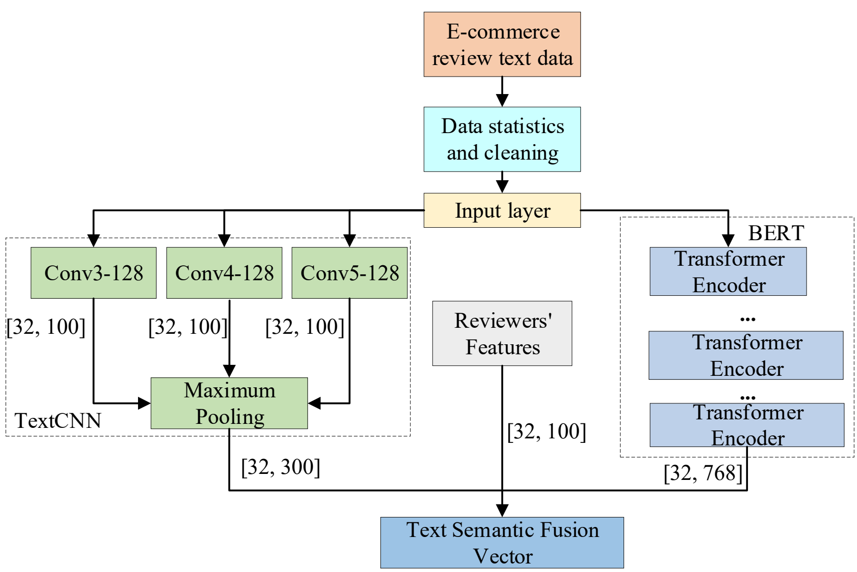


Text semantics are composed of review texts and reviewer IDs. The overall process of feature extraction and related parameter dimension settings are shown in Figure 1. First, preprocessing of the e-commerce review text data is carried out to accomplish tasks such as word segmentation, statistics, and cleaning. Then, bidirectional encoder representations from transformers (BERT) pre-training model and TextCNN are used for feature extraction from the textual semantics to obtain the review text feature vectors. The three feature vector matrices are fused at the embedding layer to obtain the final semantic fusion matrix, and the dimension of the fused matrix is 1,168.

1 Review text semantics feature extraction. The BERT pre-training model can fully describe features at the character level, word level, sentence level, and even relationships between sentences. It has achieved optimal results for tasks such as text classification, sentence pair classification, and sequence labelling, with significant advantages in extracting global text features. This paper combines TextCNN with the BERT pre-training model to capture complete semantic features of the review texts.

2 Fake reviewer feature extraction. By investigating the task of fake review identification, it is found that identifying fake reviews involves not only the fake reviews themselves but also fake reviewers and groups of fake reviewers. Since group relationships among fake reviewers are relatively complex and typically internal data on e-commerce websites, relationship data between fake reviewers cannot be obtained. During annotation of the crawled data, it was discovered that most accounts used by fake review publishers were small or temporary accounts with no profile pictures, and system default nicknames were common. In contrast, most normal users tend to use unique account ID information. Therefore, this paper studies from both the perspective of reviews themselves and reviewers, performing vectorisation on features of fake reviewers that are unique to the field of fake review detection, and integrating these into the final text semantic fusion vector, thereby enhancing the accuracy of identifying fake characteristics.

Figure 2 Multi-dimensional text semantic feature fusion (see online version for colours)



3 Text semantic feature fusion. As shown in Figure 2, the input sentences at the input layer are mapped through the embedding layer and converted into vector representations of fixed dimensions. The vector representation consists of three fused parts: the BERT pre-training model output vector, the TextCNN output vector, and the reviewer ID feature vector. The text semantics are extracted using the BERT model. Unlike traditional machine learning methods, the BERT pre-training model treats each input sentence as a character sequence $W = \{W_1, W_2, \ldots, W_N\} \in U_C$,

where U_C is the set of all occurring characters, and W is the total sum of the character sequences for each sentence. The BERT model is shown in equation (1), where T_i^{bert} is the i^{th} character's BERT vector in the input, and $Bert(s)$ is the vector expression output of sequence s in BERT.

$$T_i^{bert} = \left(Bert(s)\right)_i \tag{1}$$

TextCNN is good at capturing local feature information, while BERT has a better representation of global features. Therefore, this paper combines the advantages of both for feature vector fusion. In TextCNN, convolution is a special type of linear operation, as shown in equation (2), where x is the input to the convolutional layer, w is the i^{th} character's corresponding TextCNN vector in the kernel function input, and $S(t)$ is its output feature mapping.

$$S(t) = x \times \omega \tag{2}$$

After obtaining the feature vectors from BERT, the TextCNN embedding layer, and the reviewer ID, a Concat operation needs to be performed to fuse the features into the same dimension. Concat is one of the most straightforward methods for feature fusion. Its core logic involves concatenating features end-to-end along a specified dimension without altering the size of other dimensions, ultimately merging multiple feature tensors into a single dimension. As shown in equation (3), where E_i is the concatenated vector of the i^{th} character's BERT vector, TextCNN vector, and reviewer ID after fusion, and E^w is the Word2Vec vector corresponding to the reviewer ID.

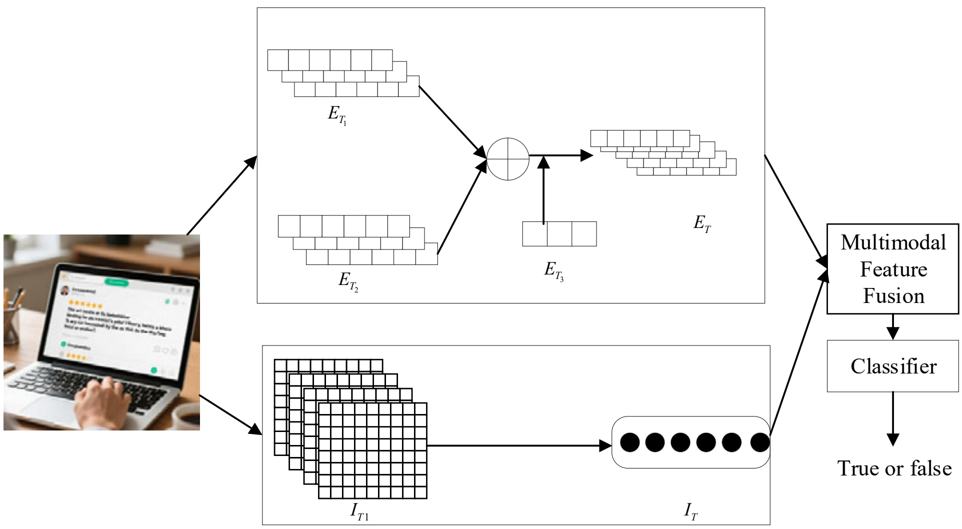
$$E_i = Concat\left(T_i^{Bert}, S(t), E^w\right) \tag{3}$$

Complete and useful reviews are mostly composed of text and images, and identifying visual reviews can improve the accuracy of detecting fake e-commerce reviews. As one of the most effective techniques in deep learning networks, CNN has achieved success in image classification. ResNet expands the depth of neural networks to 152 layers based on CNN (Wang and Wang, 2024), and uses tricks in the residual learning (residual learning) architecture to address gradient explosion and vanishing issues that are common in deep networks, enabling network depth to play its intended role for better performance (Xu et al., 2023). ResNet, with its unique residual connections and deep feature extraction capabilities, is particularly well-suited for processing images of fake e-commerce reviews. Its advantages lie in addressing core issues such as gradient vanishing and performance degradation, while enabling deep networks to capture complex anomalous patterns within fake review images. Traditional deep neural networks often suffer from gradient vanishing or exploding during training, making deep networks difficult to train. ResNet addresses this by introducing residual blocks, which add the input directly to the output, forming a residual structure between input and transformation. This design allows gradients to propagate directly through nonlinear transformation layers, solving the gradient vanishing problem and ensuring effective training of deep networks. Based on ResNet's powerful image classification capability demonstrated in image processing, this paper adopts ResNet to vectorise review images and obtain a visual semantic sequence representation $I_r = \{I_{t1}, I_{t2}, I_{t3}, \ldots, I_{tn}\}$.

3.2 Visual and text semantic multimodal fusion

Multimodal fusion is based on the features of text and visual semantics, fusing them into the same sequence as input variables for prediction. According to research on human reading habits and the subjective visual evidence provided by visuals to complement the text, readers focus attention on images while reading review texts. To simulate this mechanism computationally, this paper concatenates the textual and image features of reviews into a hybrid sequence. This combined matrix is then fed through an Attention layer to assign weights, followed by a fully connected layer and a Softmax layer to generate the final determination, as illustrated in Figure 3.

Figure 3 Multimodal semantic feature fusion (see online version for colours)



The multimodal fusion module adopts an early fusion strategy based on the model, concatenating visual features with textual features into a hybrid sequence S_T = $Concat\{E_{t1}, E_{t2}, \ldots, E_{tn}, \ldots, I_{t1}, I_{t2}, \ldots, I_{tn}\}$ on the basis of text feature fusion. An Attention layer is added before the fully connected layer to assign weights and obtain the most important feature vector representation, as shown in equation (4), where ω_z and b_z are weights and biases.

$$Z_g = \omega_z S_T + b_z \tag{3}$$

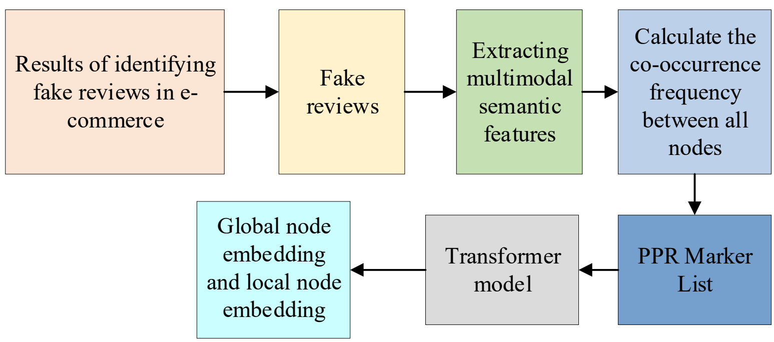
After passing through the Attention layer, the sequence Z_g is input into a fully connected neural network for training. It then goes through a Softmax layer for normalisation to output the final e-commerce fake review detection result.

4 Tracing the spread of e-commerce fake reviews based on transformer model

4.1 Framework for tracing the spread of e-commerce fake reviews

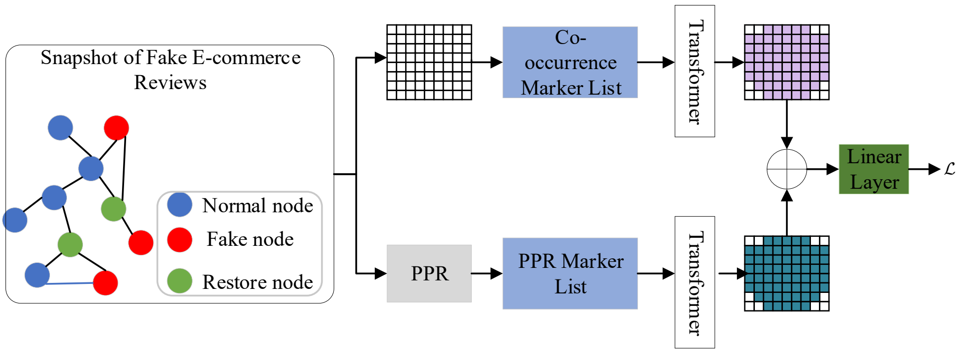
Based on the results of identifying fake e-commerce reviews, this paper analyses the propagation paths of fake reviews to achieve preliminary tracing of their origins and provide strong technical support for e-commerce platform governance. Existing methods for tracing the spread of fake e-commerce reviews have failed to address long-distance interaction issues, which severely limits their ability to accurately identify source nodes. In addition, previous approaches typically perform tracing algorithms on the entire propagation graph (Banerjee and Chua, 2021). Although this method provides comprehensive information, it lacks the capability to distinguish or selectively filter data, making it difficult for models to effectively utilise available information. This limitation reduces model performance. The relationship between identifying and tracing fake reviews in e-commerce is illustrated in Figure 4.

Figure 4 The relationship between identifying and tracing fake reviews in e-commerce (see online version for colours)



To address the above challenges, this paper proposes the GTransformer framework for tracing the spread of fake e-commerce reviews. First, the feature generation module extracts multi-modal semantic features from each node based on identified fake reviews and then combines them into comprehensive features. Next, the co-occurrence tokenisation module calculates co-occurrence frequencies among all nodes, normalises these values, and generates a list of tokens for each node according to their co-occurrence frequency with other nodes. This module prioritises nodes with higher co-occurrence frequencies to provide long-range dependencies for traceability. Subsequently, the PageRank (PPR) tokenisation module utilises PPR to generate an additional list of tokens for each node, emphasising highly relevant nodes within local neighbourhoods. This module captures localised fake review information to support tracing tasks. Finally, these two lists of tokens are input in mini-batch form into two standard Transformer models to produce global and local node embeddings. These embeddings are combined and used for the traceability task. The overall architecture is shown in Figure 5.

Figure 5 GTransformer framework for tracing the spread of fake reviews (see online version for colours)



4.2 Feature generation module

In source tracing tasks, the individual state of a node and the distribution of its neighbours' states play a crucial role in inferring the transmission source. To capture this information, this paper constructs two types of features: node state features to represent the individual state of a node, and node statistical features to encode the distribution of each state among its neighbours.

1 State features: Let the node state column vector be denoted as y. By applying one-hot encoding to y as part of the node features, we can enhance the model's learning capability and interpretability. Therefore, the state feature X_i^1 of node v_i can be represented as follows.

$$X_i^1 = \begin{cases} [1, 0], \ y_i = 0 \\ [0, 1], \ y_i = 1 \end{cases} \tag{4}$$

2 Statistical features: In addition to node state features, this paper also considers the distribution of neighbour states. Specifically, we focus on the following aspects: the degree of a node, i.e., the number of neighbours surrounding it; the number of infected nodes among its neighbours; and the number of uninfected nodes among its neighbours. Therefore, the statistical feature X_i^2 of node v_i can be constructed as follows, where $\oplus$ denotes the concatenation operation.

$$X_i^3 = |N(v_i)| \tag{5}$$

$$X_i^4 = \sum\nolimits_{v_j \in N(v_i)} y_i \tag{6}$$

$$X_i^5 = X_i^3 - X_i^4 \tag{7}$$

$$X_i^2 = X_i^3 \oplus X_i^4 \oplus X_i^5 \tag{8}$$

Finally, this paper combines state feature X_i^1 and statistical feature X_i^2 using a concatenation operation to obtain the final composite node feature.

$$X_i = X_i^1 \oplus X_i^2 \tag{9}$$

This feature fusion method not only preserves the individual information of each node but also fully leverages the distribution information of infected and uninfected states among its neighbours, providing richer and more effective input features for subsequent tracing tasks.

4.3 Co-occurrence tokenisation module

In traceability tasks, network propagation often follows certain paths, and the source node may be far from fake review nodes; therefore, accurately capturing long-range dependencies between nodes is crucial for effectively locating the source node (Li et al., 2023). However, existing studies have not adequately considered this aspect, resulting in significant limitations in identifying sources of fake review propagation. To address this, this paper proposes a co-occurrence tokenisation mechanism. This mechanism analyses historical propagation data to calculate co-occurrence frequencies between each node and all other nodes, normalises these frequencies, and generates a list of tokens for each node. The token list reflects the relative importance of other nodes in the context of propagation, enabling the model to better understand long-range dependencies and structural patterns within the network. Specifically, for a dataset containing p propagation paths, a fake review matrix I is constructed, where element $I_{ij} = 1$ if node j appears in the i^{th} propagation path, otherwise $I_{ij} = 0$. Subsequently, this paper calculates the co-occurrence matrix C, as shown in equation (11), where $C_{i,j}$ represents the number of times nodes v_i and v_j issue fake reviews together. Next, this paper performs row normalisation on C, as shown in equation (12).

$$C = I^T I \tag{10}$$

$$C_{i,j} \leftarrow \frac{c_{i,j}}{c_{i,i}} \tag{11}$$

For each node v_i, the top k_1 nodes from the i^{th} row of C are selected, forming its co-occurrence token list T_c^i.

$$T_c^i = TopK\left(C_{i,:}, k_1\right) \tag{12}$$

where $C_{i,:}$ is the i^{th} row of C. To preserve the relative importance of nodes in T_c^i, ranking scores from C are used as positional encodings. For each node v_j belonging to T_c^i, its feature input to the transformer model is the concatenation of the node feature X_j and its ranking score $C_{i,j}$.

$$X_c^{ij} = X_j \oplus C_{i,j} \tag{13}$$

where $\oplus$ denotes the concatenation operation. This approach enables the Transformer to simultaneously capture a node's intrinsic attributes and its contextual importance in the propagation structure, thus more effectively extracting global long-range dependencies.

4.4 Personalised pageRank tokenisation module

In addition to capturing long-range dependencies, local information is equally crucial for source tracing tasks. Personalised PPR can effectively quantify the relevance of each node in a local neighbourhood through random walks. To complement the global information provided by co-occurrence tokenisation, this paper employs PPR to generate an additional token list for each node v_i. This token list captures its local structural context by prioritising nodes that are highly related to v_i within their neighbourhoods. The combination of co-occurrence and PPR tokenisations enables the model to leverage both global and local information simultaneously, achieving more accurate source identification.

For each node v_i, its PPR vector r is computed based on equation (15). Subsequently, the top k_2 nodes ranked in r are sampled to construct the PPR token list T_p^i.

$$r_i = \alpha \,\mathrm{Pr} + (1-\alpha) q \tag{14}$$

where P is the transition matrix; q is a one-hot random vector; r is the stationary distribution of the random walk, commonly referred to as the PPR vector; and α is the damping constant.

Similar to co-occurrence tokenisation, this paper uses the ranking scores in r as positional encodings. Specifically, for each node v_j, its input features to the transformer model are a concatenation of the node feature X_j and its PPR ranking score $r_{i,j}$, as shown below:

$$X_p^{ij} = X_j \oplus r_{i,j} \tag{15}$$

4.5 Transformer-based propagation source tracing

After obtaining the co-occurrence and PPR token lists for each node, this paper implements a mini-batch training scheme for the graph Transformer by loading mini-batches of token lists and utilising self-attention mechanisms within each batch. This design not only improves computational efficiency but also ensures scalability of the model on large-scale graphs, especially suitable for fake review source tracing tasks that require simultaneous analysis of global and local propagation patterns.

For each node v_i, this paper separately stacks the features of nodes in their co-occurrence token list T_c^i and PPR token list T_p^i to construct two feature matrices $X_c^i \in R^{k_1 \times (d+1)}$ and $X_p^i \in R^{k_2 \times (d+1)}$, where k_1 and k_2 denote the number of tokens taken from T_c^i and T_p^i respectively, d denotes the dimension of two vectors. These two matrices serve as inputs for two independent standard Transformer models, thereby extracting global and local embeddings for each node. Taking T_c^i as an example, its initial input is $Z_{i,c}^{(0)} = X_c^i \in R^{k_1 \times (d+1)}$. For the l^{th} layer of the Transformer, the calculation process is defined as follows:

$$Z_{i,c}^{(l)} = FFN\left(Softmax\left(\frac{Q_{i,c} K_{i,c}^T}{\sqrt{h}} \right) V_{i,c} \right) \tag{16}$$

$$Q_{i,c} = Z_{i,c}^{(l-1)} W_Q^{(l)} \tag{17}$$

$$K_{i,c} = Z_{i,c}^{(l-1)} W_K^{(l)} \tag{18}$$

$$V_{i,c} = Z_{i,c}^{(l-1)} W_V^{(l)} \tag{19}$$

where $W_Q^{(l)}$, $W_K^{(l)}$, and $W_V^{(l)}$ are the trainable weights in the linear projection layers of the l^{th} layer, and FFN denotes a feed-forward neural network. Residual connections and layer normalisation are applied in each attention block and FFN block. To obtain the representation vector of node v_i, this paper applies a summation operation to $Z_{i,c}^{l}$ to derive the global embedding h_i^c for the node. For the PPR token list, a similar process is followed, using another Transformer model to extract the local embedding h_i^p of the node. Finally, the global and local embeddings are fused in the following way to obtain the final representation of node v_i.

$$h_i = h_i^c \oplus h_i^p \tag{20}$$

By employing different k_1 and k_2 values, the model can flexibly adjust its receptive field to capture both global and local fake review information simultaneously. This adaptability is especially critical for source tracking tasks because it captures long-range dependencies through co-occurrence relationships while also capturing local neighbourhood structures via PPR models. The combination of these two complementary mechanisms enables the model to comprehensively understand global propagation patterns and local propagation characteristics, providing a solid foundation for accurately identifying the sources of fake reviews in e-commerce. Finally, this paper inputs the node embedding matrix into a linear layer to generate the predicted source vector, denoted as $\tilde{s}$. The loss function $\mathcal{L}$ used for inferring the source node is defined as binary cross-entropy loss, with the equation as follows. This loss function aims to optimise model parameters by minimising the gap between predictions and true labels, thus improving the accuracy of source tracing for fake e-commerce reviews.

$$\mathcal{L} = \sum_i^n \left(-s_i \log \tilde{s}_i - (1 - s_i)\log(1 - \tilde{s}_i)\right) \tag{21}$$

5 Implementation and results

5.1 Experimental setup

The data used in the experiments comes from Jingdong (JD.com), containing all user comments and product information on JD.com from 2020 to 2022. The total number of products involved in these reviews is 448,000, while the number of users who participated in commenting reaches 1.578 million. Among them, 68% are one-time commenters, accounting for two-thirds of all platform users. This paper preprocesses the aforementioned dataset by removing samples with empty comment content, missing user IDs, or abnormal ratings. Duplicate samples are deduplicated based on user ID, product

ID, and comment content to prevent model overfitting. Missing values are subsequently handled to standardise the data. Unstructured review text is converted into a structured and standardised format to facilitate feature extraction.

After preprocessing, 44,550 effective text comment data points were obtained. Next, the training dataset needed to be prepared and annotated. Data annotation requires significant manpower. From a cost perspective, this study randomly sampled 1,666 text comments from each store to form the training set, totalling 13,588 text comments requiring annotation. Five graduate students specialising in e-commerce were organised for this study. After familiarising themselves with the annotation rules, they annotated the text reviews. During annotation, they comprehensively considered factors mentioned earlier (such as dates with abnormal review counts for each store and duplicate reviews). They annotated the data based on multiple characteristics of fake reviews, including high repetition, abnormal timing, inclusion of other platform names, abnormal sentiment, and content mismatch with the product. to identify fake reviews (labelled as 1) and genuine reviews (labelled as 0). The table below details the dataset specifications. This dataset was divided into training, testing, and validation sets in a 6:2:2 ratio.

Table 2 Details of the dataset

Content	*Fake comments*	*Real comments*	*Total number*
Number of comments	4,324	9,264	13,588
The number of users	4,019	8,973	12,992
Number of IPs	1,542	6,539	8,081
Number of texts	3,029	6,081	9,110
Number of pictures	1,295	3,183	4,478

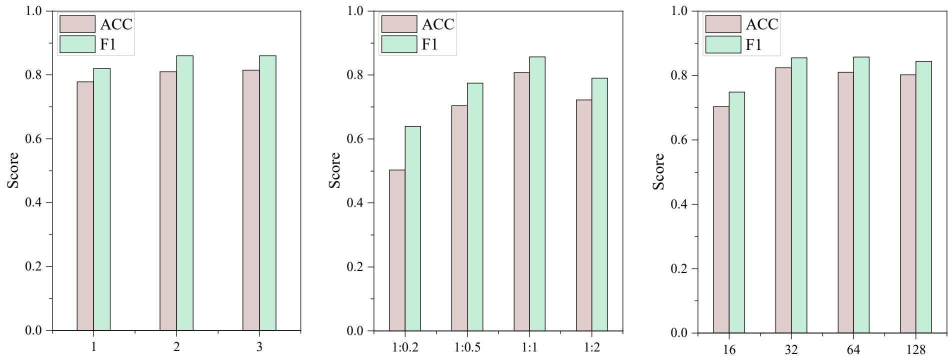
The hardware used for the experiment is an Intel Xeon E5-2603 v4 processor, 64 GB memory, Nvidia Tesla k80 graphics card, and the operating system is Ubuntu 16.04.10. The software configuration environment of the experiment is Python3.6.8, and the compilation software is Jupyter Notebook. In the TextCNN model, the convolution kernel size in the convolutional pooling is 256, the batch training size is 800, and the total number of iterations is 20. To prevent rapid convergence and overfitting during prediction, dropout is set to 0.2, batch size is 64, and the total number of iterations is 10.

5.2 Parameter sensitivity analysis

To evaluate the proposed method for fake review detection, we conduct sensitivity analyses to assess the impact of data imbalance and model configurations. Specifically, we compare the effects of varying the number of attention heads (k), the downsampling ratio, and the node embedding dimension. The results are presented in Figure 6.

As the number of heads k in multi-head attention increases, the model detection performance improves; however, when k = 2 and k = 3, there is little difference, indicating that two-headed attention is sufficient to capture different relational information. From Figure 5(b), it can be seen that imbalance between positive and negative samples causes overfitting in the model, and a one-to-one sampling strategy of positive and negative samples helps improve model performance. Figure 5(c) shows the model's performance under different node embedding dimensions. When the embedding dimension is 32, 64, or 128, the model performs similarly; when the embedding dimension is 16, the performance is worst. The accuracy comparison of different text feature vectors on the OURS model is shown in Table 3.

Figure 6 Sensitivity analysis results for different parameters, (a) number of attention heads *k* (b) data sampling ratio (c) node embedding dimension (see online version for colours)



(a) (b) (c)

Table 3 Accuracy comparison across different document vector dimensions

Vector dimension	*50*	*100*	*150*	*200*	*250*	*300*
Accuracy/%	75.6	82.4	94.8	86.3	82.4	80.7

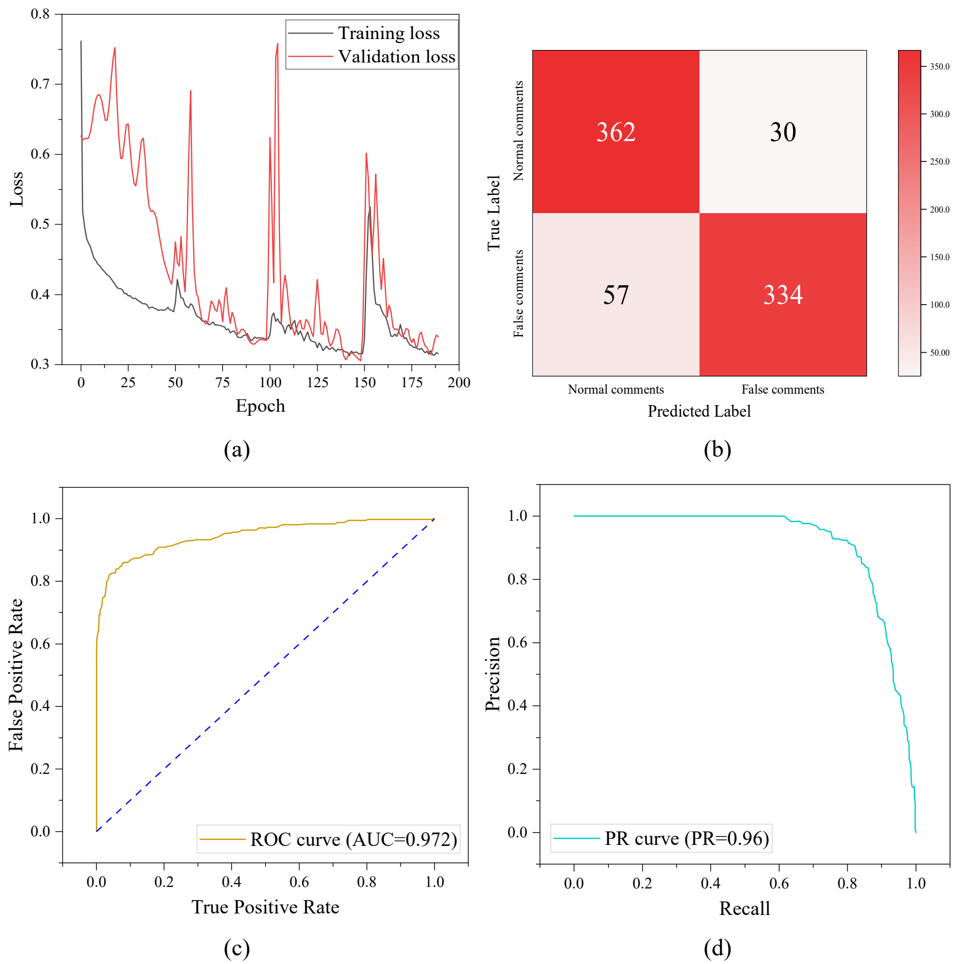
As indicated by the results in the table, the optimal document vector dimension for the OURS model is 150. Recognition accuracy falls below 90% for other vector dimensions. Therefore, the document vector dimension in the OURS model is set to 150.

5.3 Fake review identification results in e-commerce

Figure 6 presents four distinct evaluation metrics for the proposed method OURS. From left to right and top to bottom, these are the loss curve, confusion matrix, ROC curve, and PR curve. In the training and validation loss curves, the blue curve represents training loss, while the orange curve represents validation loss. As shown in Figure 7(a), although both training and validation losses exhibit fluctuations, there is no scenario where training loss continuously decreases while validation loss sharply increases, accompanied by a widening gap between the two. This indicates no obvious signs of overfitting. Such fluctuations may be attributed to factors such as data noise, model complexity, or randomness during the training process.

The confusion matrix is used to evaluate the performance of recognition models. The confusion matrix for OURS is shown in Figure 7(b). In this binary classification problem, the 362 in the upper left corner represents the number of genuinely normal comments correctly predicted as normal; the 30 in the upper right corner represents the number of genuinely normal comments incorrectly predicted as fake; the 57 in the lower left corner represents the number of true fake comments incorrectly predicted as normal comments; the 334 in the lower right corner indicates the number of true fake comments correctly predicted as fake comments. The orange curve in Figure 7(c) is the ROC curve, labelled with AUC = 0.972 below it, signifying an area under the curve of 0.972. The closer the AUC value is to 1, the better the model's performance. The PR curve illustrates the relationship between precision and recall at different thresholds. The blue curve in Figure 7(d) represents the PR curve, with an area under the curve of 0.96. A higher AP value indicates the model effectively balances precision and recall across various thresholds. Overall, these charts demonstrate the model's training process and performance from multiple perspectives, showing strong capability in the identification task.

Figure 7 Visualisation of different evaluation metrics for the proposed method, (a) loss value (b) Confusion matrix (c) ROC curve (d) PR curve (see online version for colours)



5.4 *Analysis of recognition and tracing performance*

To further verify the accuracy of recognition and tracing capabilities of the proposed method, this paper selects commonly used evaluation metrics accuracy (ACC), F1, and AUC to analyse the recognition and tracing performance of different models. Comparative methods include SUSVM (Purifyregalia et al., 2025), FRD-LSTM (Qayyum et al., 2023), TCLMD (He et al., 2023), IMCPE (Hu et al., 2025), DFRMDL (Hou et al., 2025), DBERT (Hyder et al., 2024), DLTRANS (Alsaad and Joshi, 2024), MFPN (Jing et al., 2023) and QMFND (Qu et al., 2024). The comparison of recognition and tracing performance metrics for these methods is shown in Table 4. The recognition and attribution accuracy rates of OURS reached 94.2% and 93.5%, respectively, representing improvements of at least 0.7% and 1.1% over the baseline method. Compared to the F1 scores, the harmonic mean of precision and recall, OURS achieved F1 scores of 96.3% for recognition and 97.6% for attribution, surpassing the other five methods by at least 0.5% and 2%, respectively.

Table 4 Recognition and tracing performance of different methods

Method	*Recognition*			*Traceability*		
	ACC	*F1*	*AUC*	*ACC*	*F1*	*AUC*
SUSVM	0.803	0.828	0.824	0.812	0.849	0.854
FRD-LSTM	0.828	0.845	0.857	0.838	0.861	0.897
TCLMD	0.871	0.897	0.881	0.856	0.905	0.929
IMCPE	0.905	0.924	0.928	0.891	0.916	0.948
DFRMDL	0.929	0.931	0.953	0.919	0.952	0.976
DBERT	0.895	0.927	0.904	0.906	0.892	0.917
DLTRANS	0.920	0.932	0.949	0.915	0.908	0.952
MFPN	0.935	0.958	0.964	0.924	0.920	0.958
QMFND	0.914	0.928	0.941	0.919	0.956	0.951
OURS	0.942	0.963	0.972	0.935	0.976	0.984

When comparing the area under the ROC curve AUC, our method achieves the highest AUC values for both identification and attribution, approaching 1. This demonstrates the high accuracy of our proposed approach in both tasks. SUSVM obtains classification results for fake e-commerce reviews using SVM, but its classification accuracy is low, leading to a high number of fake positives and fake negatives, which in turn reduces traceability accuracy. FRD-LSTM utilises CNN and LSTM to extract semantic features from text and visual data respectively. After identifying candidate rumour source clusters via a multi-layer perceptron module, it predicts source nodes. However, this approach heavily relies on node content features while failing to fully leverage network topology characteristics, resulting in suboptimal identification and attribution accuracy. TCLMD employs a dual-layer attention structure for comment identification and propagation tracing. However, attention mechanisms are sensitive to noise and outliers in input data. If comment datasets contain numerous spelling errors, non-standard terminology, or maliciously fabricated fake reviews, model performance may significantly deteriorate. IMCPE employs GCN for identifying and tracing fake e-commerce reviews. While it leverages graph structures to capture complex relationships between reviews, user behaviour and product reviews on e-commerce platforms are dynamic. The static graph structure of GCN struggles to incorporate new data in real-time, requiring periodic retraining that may introduce historical data forgetting. DFRMDL employs attention mechanisms to analyse multimodal semantics for fake review detection and attribution. However, when fake reviews originate from multiple independent sources, the multimodal attention structure may struggle to distinguish features from different sources, potentially reducing attribution accuracy. OURS not only extracts multimodal semantics from e-commerce reviews but also leverages the powerful expressive capabilities of Transformer models to achieve efficient fusion and computation of both global and local information, thereby enhancing the accuracy of identification and attribution. DBERT utilises the BERT model to identify fake reviews. However, fake reviewers can generate adversarial examples by replacing keywords with synonyms or adjusting sentence structures, potentially leading BERT to misjudge due to over-reliance on local word meanings. DLTRANS employs CNNs and Transformers for fake review detection. However, e-commerce reviews often include images and ratings, multimodal information that aids in detecting fakes, such as when a review praises a product but features blurry images or when ratings contradict the text. CNNs and Transformers process only text, ignoring cross-modal correlations. MFPN often assumes multimodal information is complementary. Yet, in real scenarios, certain modalities may be low-quality, introducing noise during fusion. For example, fake reviewers might use irrelevant images or copy pictures from other products. If the model fails to recognise such low-quality modalities, it may erroneously rely on their features for judgment. QMFND's lightweight design may sacrifice some performance in identifying fake reviews.

After 100 independent runs, the mean and standard deviation of the ACC for tracing fake e-commerce reviews were obtained. Furthermore, the renowned Wilcoxon signed-rank test was employed to analyse the results. Here, '+', '−', and '≈' denote that the comparison algorithm is significantly better than OURS, significantly worse than OURS, or statistically similar to OURS, respectively. The ACC and F1 scores alongside Wilcoxon rank-sum test results are presented in Table 5. It is evident that OURS exhibits lower maximum, minimum, average, and standard deviation values for the ACC metric compared to the comparison algorithms, indicating that OURS significantly outperforms the other models in terms of diversity and convergence. The test results demonstrate that OURS achieves outstanding accuracy in tracing fake reviews on e-commerce platforms.

Table 5 Outcome of ACC indicators and Wilcoxon rank-sum

Model	*Mean value*	*Standard deviation*	*Max*	*Min*	*Significance test*
SUSVM	0.794	0.805	0.803	0.758	–
FRD-LSTM	0.806	0.824	0.849	0.783	–
TCLMD	0.867	0.851	0.885	0.807	–
IMCPE	0.883	0.893	0.909	0.839	–
DFRMDL	0.905	0.907	0.917	0.858	–
DBERT	0.884	0.892	0.903	0.869	–
DLTRANS	0.908	0.912	0.924	0.846	–
MFPN	0.926	0.935	0.916	0.851	–
QMFND	0.901	0.911	0.942	0.874	–
OURS	0.937	0.942	0.965	0.916	

The training time, recognition time, and attribution time for different models are shown in Table 6. Our model achieved a training time of only 1,205 minutes, significantly lower than other models.

This efficiency stems from the proposed model's full utilisation of LSTM's sequence processing capabilities and Transformer's high-performance parallel processing capabilities. Comparing fake review detection and attribution times, OURS achieves 7.8 s and 14.6 s respectively, both lower than other models. This fully validates the proposed model's high efficiency.

Table 6 Model training time comparison

Model	*Training time/min*	*Identification time/s*	*Traceability time/s*
SUSVM	2,801	22.1	36.8
FRD-LSTM	3,169	25.9	28.7
TCLMD	2,098	18.3	22.5
IMCPE	2,394	17.2	26.4
DFRMDL	1,569	15.4	43.8
DBERT	1,237	20.9	35.6
DLTRANS	1,854	12.6	29.4
MFPN	1,283	16.8	18.7
QMFND	1,369	11.5	20.7
OURS	1,205	7.8	14.6

5.5 Ablation experiment

To further verify the significance of each component in the proposed model, this paper conducts an ablation experimental study on each component in the recognition model. Mark the text feature extraction module removed from the model as OURS/T. Remove the image feature extraction module and denote it as OURS/F. Change the attention feature fusion module to splicing fusion and denote it as OURS/S. The ablation experiment results of different components are shown in Table 7.

Table 7 Results of the ablation experiment

Model	*ACC*	*F1*	*AUC*
OURS/T	0.825	0.806	0.857
OURS/F	0.858	0.822	0.906
OURS/S	0.879	0.851	0.939
OURS	0.942	0.963	0.972

The recognition accuracy of OURS has increased by 11.7% and 15.7% respectively compared with F1 for OURS/T, by 8.4% and 14.1% respectively compared with OURS/F, and by 6.3% and 11.2% respectively compared with OURS/S. When comparing the recognition accuracy, the AUC of OURS has increased by 13.42%, 7.29%, and 3.51% respectively compared to OURS/T, OURS/F, and OURS/S. The above results indicate that the accuracy of identifying fake e-commerce reviews based on a single modality is inferior to that of multi-modal methods. In addition, achieving multimodal feature fusion through the attention mechanism can significantly improve the recognition accuracy of the model. Based on the above analysis, the e-commerce fake review identification model that integrates all components has achieved good identification results.

6 Conclusions

With the rapid development of e-commerce, fake reviews have become a severe challenge in misleading consumer decisions. Existing research primarily relies on single-text modalities or simple behavioural features, which are difficult to adapt to increasingly complex strategies for fake reviews, such as carefully fabricated texts accompanied by misleading images. To address these issues, this paper proposes an accurate method for detecting and tracing e-commerce fake reviews that integrates multimodal semantic understanding. First, based on the text and images posted by users combined with reviewer characteristics, a multimodal learning approach comprehensively utilises textual semantics features and image visual semantics features to assess user review authenticity. Based on this foundation, a Transformer model is used to trace e-commerce fake reviews. Co-occurrence tokenisation lists and PPR lists are processed separately by two standard Transformer models in mini-batch fashion, generating global and local node embeddings. Finally, a linear layer fuses the global and local embeddings to determine whether each node is a propagation source. Experimental results show that the identification accuracy and tracing accuracy of the proposed method are improved by at least 1.3% and 0.8%, respectively, effectively enhancing the accuracy of fake review detection and tracing.

In the future work, this study will further expand and deepen our research on tracing fake reviews in e-commerce, striving to achieve new breakthroughs at both theoretical and practical levels. Moving forward, we plan to explore how to effectively incorporate higher-level structural features into tracing tasks.

Declarations

All authors declare that they have no conflicts of interest.

References

Abd-Alhalem, S.M., Ali, H.A., Soliman, N.F., Algarni, A.D. and Marie, H.S. (2024) 'Advancing e-commerce authenticity: a novel fusion approach based on deep learning and aspect features for detecting false reviews', *IEEE Access*, Vol. 12, pp.116055–116070.

Alsaad, M.M.B. and Joshi, H. (2024) 'Combination between deep learning and transformer models to detect fake yelp electronic product reviews', *Journal of Computational Analysis and Applications*, Vol. 33, No. 7, pp.13–25.

Alsharif, N. (2022) 'Fake opinion detection in an e-commerce business based on a long-short memory algorithm', *Soft Computing*, Vol. 26, No. 16, pp.7847–7854.

Alsubari, S.N., Deshmukh, S.N., Alqarni, A.A., Alsharif, N., Aldhyani, T., Alsaade, F.W. and Khalaf, O.I. (2022) 'Data analytics for the identification of fake reviews using supervised learning', *Computers, Materials & Continua*, Vol. 70, No. 2, pp.3189–3204.

Baishya, D., Deka, J.J., Dey, G. and Singh, P.K. (2021) 'SAFER: sentiment analysis-based fake review detection in e-commerce using deep learning', *SN Computer Science*, Vol. 2, No. 6, pp.59–67.

Banerjee, S. and Chua, A.Y. (2021) 'Calling out fake online reviews through robust epistemic belief', *Information & Management*, Vol. 58, No. 3, pp.44–56.

Chai, Y., Wang, Y. and Zhu, L. (2021) 'Information sources estimation in time-varying networks', *IEEE Transactions on Information Forensics and Security*, Vol. 16, pp.2621–2636.

Du, P. and Jiang, Y. (2025) 'Enhancing cross-border e-commerce English text classification using graph neural networks and transfer learning', *International Journal of Information and Communication Technology*, Vol. 26, No. 39, pp.22–36.

Fan, D. (2024) 'Research on the identification method of fake reviews in e-commerce', *Academic Journal of Business & Management*, Vol. 6, No. 4, pp.276–281.

He, Y., Li, Z., Li, Z., Zhou, S., Yu, T. and Zhang, J. (2023) 'Towards cross-lingual multi-modal misinformation detection for e-commerce management', *IEEE Transactions on Network and Service Management*, Vol. 20, No. 2, pp.1040–1050.

Hou, J., Tan, Z., Zhang, S., Hu, Q. and Wang, P. (2025) 'Detecting fake review intentions in the review context: a multimodal deep learning approach', *Electronic Commerce Research and Applications*, Vol. 70, pp.91–105.

Hu, J., Yang, M., Tang, B. and Hu, J. (2025) 'Integrating message content and propagation path for enhanced false information detection using bidirectional graph convolutional neural networks', *Applied Sciences*, Vol. 15, No. 7, pp.34–57.

Hyder, S.B., Tariq, N., Moqurrab, S.A., Ashraf, M., Yoo, J. and Srivastava, G. (2024) 'Bert-based deceptive review detection in social media: introducing deceptivebert', *IEEE Transactions on Computational Social Systems*, Vol. 11, No. 6, pp.7234–7243.

Jing, J., Wu, H., Sun, J., Fang, X. and Zhang, H. (2023) 'Multimodal fake news detection via progressive fusion networks', Information Processing & Management, Vol. 60, No. 1, pp.10–23.

Li, Q., Park, J. and Kim, J. (2024) 'Impact of information consistency in online reviews on consumer behavior in the e-commerce industry: a text mining approach', *Data Technologies and Applications*, Vol. 58, No. 1, pp.132–149.

Li, W., Guo, C., Liu, Y., Zhou, X., Jin, Q. and Xin, M. (2023) 'Rumor source localization in social networks based on infection potential energy', *Information Sciences*, Vol. 634, pp.172–188.

Luo, J., Luo, J., Nan, G. and Li, D. (2023) 'Fake review detection system for online e-commerce platforms: a supervised general mixed probability approach', *Decision Support Systems*, Vol. 175, pp.45–61.

Mughal, N., Mujtaba, G., Mughal, M.H., Manaf, A. and Kamangar, Z. (2025) 'Fake reviews detection on e-commerce websites using novel user behavioral features: an experimental study', *ACM Transactions on Asian and Low-Resource Language Information Processing*, Vol. 24, No. 9, pp.1–44.

Purifyregalia, K., Umam, K., Wibowo, N.C.H. and Handayani, M.R. (2025) 'Detecting fake reviews in e-commerce: a case study on Shopee using support vector machine and random forest', *Journal of Applied Informatics and Computing*, Vol. 9, No. 3, pp.955–965.

Qayyum, H., Ali, F., Nawaz, M. and Nazir, T. (2023) 'FRD-LSTM: a novel technique for fake reviews detection using DCWR with the Bi-LSTM method', *Multimedia Tools and Applications*, Vol. 82, No. 20, pp.31505–31519.

Qu, Z., Meng, Y., Muhammad, G. and Tiwari, P. (2024) 'QMFND: A quantum multimodal fusion-based fake news detection model for social media', *Information Fusion*, Vol. 104, pp.102–115.

Shi, C., Zhang, Q. and Chu, T. (2022) 'Locating the source of diffusion in networks under mixed observation condition', *Physics Letters A*, Vol. 434, pp.74–86.

Singh, T., Kumari, M. and Gupta, D.S. (2024) 'Rumor identification and diffusion impact analysis in real-time text stream using deep learning', *Journal of Supercomputing*, Vol. 80, No. 6, pp.15–26.

Thuy, D.T.T., Thuy, L.T.M., Bach, N.C., Duc, T.T., Bach, H.G. and Cuong, D.D. (2024) 'Designing a deep learning-based application for detecting fake online reviews', *Engineering Applications of Artificial Intelligence*, Vol. 134, pp.108–123.

Vijayaragavan, P., Ponnusamy, R. and Aramudhan, M. (2020) 'An optimal support vector machine based classification model for sentimental analysis of online product reviews', *Future Generation Computer Systems*, Vol. 111, pp.234–240.

Wang, D. and Wang, L. (2024) 'Deep learning semantic understanding and classification of student online public opinion for new media', *International Journal of Information and Communication Technology*, Vol. 25, No. 10, pp.62–76.

Xu, W., Fu, Y-L. and Zhu, D. (2023) 'ResNet and its application to medical image processing: Research progress and challenges', *Computer Methods and Programs in Biomedicine*, Vol. 240, pp.66–73.

Zhang, F., Hao, X., Chao, J. and Yuan, S. (2020) 'Label propagation-based approach for detecting review spammer groups on e-commerce websites', *Knowledge-Based Systems*, Vol. 193, pp.105–113.