Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.040862

ARTICLE

DFE-GCN: Dual Feature Enhanced Graph Convolutional Network for Controversy Detection

Chengfei Hua[1,2,3], Wenzhong Yang[2,3,*], Liejun Wang[2,3], Fuyuan Wei[2,3], KeZiErBieKe HaiLaTi[2,3] and Yuanyuan Liao[2,3]

[1]College of Software, Xinjiang University, Urumqi, 830000, China

[2]Key Laboratory of Signal Detection and Processing in Xinjiang Uygur Autonomous Region, Xinjiang University, Urumqi, 830000, China

[3]Key Laboratory of Multilingual Information Technology in Xinjiang Uygur Autonomous Region, Xinjiang University, Urumqi, 830000, China

*Corresponding Author: Wenzhong Yang. Email: yangwenzhong@xju.edu.cn

Received: 01 April 2023 Accepted: 07 August 2023 Published: 31 October 2023

ABSTRACT

With the development of social media and the prevalence of mobile devices, an increasing number of people tend to use social media platforms to express their opinions and attitudes, leading to many online controversies. These online controversies can severely threaten social stability, making automatic detection of controversies particularly necessary. Most controversy detection methods currently focus on mining features from text semantics and propagation structures. However, these methods have two drawbacks: 1) limited ability to capture structural features and failure to learn deeper structural features, and 2) neglecting the influence of topic information and ineffective utilization of topic features. In light of these phenomena, this paper proposes a social media controversy detection method called Dual Feature Enhanced Graph Convolutional Network (DFE-GCN). This method explores structural information at different scales from global and local perspectives to capture deeper structural features, enhancing the expressive power of structural features. Furthermore, to strengthen the influence of topic information, this paper utilizes attention mechanisms to enhance topic features after each graph convolutional layer, effectively using topic information. We validated our method on two different public datasets, and the experimental results demonstrate that our method achieves state-of-the-art performance compared to baseline methods. On the Weibo and Reddit datasets, the accuracy is improved by 5.92% and 3.32%, respectively, and the F1 score is improved by 1.99% and 2.17%, demonstrating the positive impact of enhanced structural features and topic features on controversy detection.

KEYWORDS

Controversy detection; graph convolutional network; feature enhancement; social media

1 Introduction

In recent years, along with the rapid development of the Internet and other infrastructure, various social media applications have influenced every aspect of human life. While bringing convenience to

users, they have also given rise to many controversial phenomena. These controversial phenomena continue to fester on the Internet, triggering public opinions and undermining social stability and harmony. Controversy detection and regulation have therefore become vital social challenges. Controversial posts are those that cause significant differences of opinion among a significant number of people, and controversial posts on social media are texts that provoke strong reactions, both positive and negative, and spark heated debates among users [1]. Identifying controversial posts on social media has an important role as it allows people to receive controversial messages in advance, encourages fairness and impartiality, and discourages the spread of disinformation or hate speech [2]. Detecting social media controversy also provides academics with a window into public sentiment and opinion, allowing them to assess the reality and significance of events as well as the attitudes and opinions of users.

Social media often uses interactions and conversations between users to identify controversies. These interactions include social connections (retweets, followers, and comments on Reddit) [3,4], as well as Wikipedia citation links [5]. Recently, researchers have been interested in identifying controversial content on various social media platforms, including Twitter, Reddit, Weibo, and government websites. Garimella et al. [3] studied controversial topics on social media by constructing retweets, text, and follow graphs and measuring the level of controversy of the target topic. Hessel et al. [6] created a reply structure graph and used textual content and the structure of replies between users to identify controversial posts on the Reddit website. In addition, Zhong et al. [7] created a controversy detection Chinese dataset based on Weibo. They proposed a graph convolutional network-based method to detect intra-topic and inter-topic controversies on Weibo. Benslimane et al. [8] suggested a controversy detection approach based on user interaction graph structure and text features to identify post-level controversy on the Reddit platform. Although the above studies have yielded promising results for specific controversy detection tasks, some challenges remain. Current methods have limited ability to obtain information from such structures when certain data have a very deep tree structure, which is helpful for controversy detection [6]. Therefore, the first challenge is to obtain information about the deeper structure. In addition, topics are the root cause of controversial phenomena and often contain more feature information than other nodes. Effectively using controversial topic information is the second challenge to improving controversy detection capability.

Our task in this paper is to detect controversial posts on social media. Since topic-level controversy detection results are coarse, multiple controversial and non-controversial posts are often found under a single topic. To overcome two drawbacks of existing work, the DFE-GCN model utilizes a global structure encoding layer and a local structure encoding layer to obtain multi-scale structural features to enhance the representation of structural features. In addition, to fully use the rich topic information, we augment the topic features behind each graph convolution layer with an attention mechanism. Finally, fused post and comment update vectors are used to predict controversies. Extensive experiments show that our model outperforms existing methods and can exploit its features dynamically and effectively. The main contributions of this paper are as follows:

1. This paper proposes a DFE-GCN model that can accurately detect controversial posts on social media. The DFE-GCN is the first model that improves the performance of controversy detection by starting from both structure and topic levels.

2. In this paper, a structure enhancement module is utilized to acquire multi-scale structural information from the reply structure graph, enhancing the representation capability of structural features from both global and local perspectives. The structure enhancement module facilitates the exploration of deep-level structural features.

3. In this paper, we use a topic enhancement module that makes extensive use of topic feature information to achieve good controversy recognition performance. The topic enhancement module improves the impact of the root source of controversial events.

The rest of this paper is structured as follows: Section 2 discusses related work and the context. Section 3 goes into the method in further depth. Section 4 discusses evaluation techniques, experimental results, and performance comparisons. The work is summarized in Section 5.

2 Related Work

The goal of social media controversy detection is to determine if a person's words are controversial by analyzing data from user interactions on social networking sites, such as comments, comment content, and response structure. Thus far, controversy detection approaches may be classified into three types: content-based, structure-based, and hybrid.

2.1 Content-Based

Text content features were widely used in early controversy detection tasks, where text content could be used to detect controversial topics and posts. According to Sznajder et al. [9], the context in which a concept is located has a solid indicative impact and can measure the level of controversy on a Wikipedia page. Instead of using Wikipedia metadata, the authors argue that the context of the concept may be a clear indication of controversy. The authors used pre-trained word embedding methods to represent article content and calculated the degree of controversy for these concepts using nearest neighbor and plain Bayes classifiers, respectively. The study by Dori-Hacohen et al. [10] focused on assessing whether content on web pages is controversial. They estimated the degree of controversy in Wikipedia by assessing the semantic differences between authors and measuring the proximity of controversy to the article topic to determine whether the subject was controversial and, thus, the controversial nature of the news article. By studying phrases that often appear in contexts with emotive words, Choi et al. [11] identified controversies in news articles. The goal of the study by Sriteja et al. [12] was to determine the controversial scores of articles using users' commentary views, combining them with information about the news article's content. Information on sentiment analysis, linguistic analysis, topic strength, and user interaction was also considered. Finally, a study by Jang et al. [13] concluded that controversy should be studied in a specific context, as it is not a generalized concept but relevant to the community. Although textual features are essential for controversy detection, relying on textual features alone is not sufficient to achieve the task.

2.2 Structure-Based

Texts on social media are often biased because their meaning may vary depending on many factors, such as the culture or language of the community, and caution should be exercised when dealing with such texts [14]. Therefore, it is essential to investigate controversies from a structural perspective. Guerra et al. [15] were the pioneers in utilizing topological features to extract controversies. They partitioned the nodes of two distinct communities into two non-overlapping sets. Then they identified polarizing topics by examining the internal degree and internal edges of the nodes, along with the boundary degree and the number of cross-group edges. Garimella et al. [3] constructed a conversation graph based on topics containing two types of edges: follow and retweet. After partitioning the conversation graph into two different communities, different methods were used to measure controversy, including a Random Walk Controversy (RWC) measure. Garimella et al. [2] used a similar approach in their study of controversy detection and demonstrated that by examining

users' relationships, their commitment to the community could be determined. They propose a new method based on biased random walks and use a new measure of controversy to quantify controversy. Mendoza et al. [16] emphasized the importance of communication between users and named entities. They generate condition graphs on partitioned named entities and use RWC controversy scores to quantify controversy. Although structural features are widely used and produce good results in controversy detection tasks, ignoring textual features is still a loss.

2.3 Hybrid Approaches

Recent research has focused on merging structural and content information to prevent the loss of important feature information in controversy detection tasks. In this context, many user interaction strategies on social media are crucial. Hessel et al. [6] demonstrated that textual features are as important as structural features for controversy detection tasks. They used text features generated by language models like Bidirectional Encoder Representation from Transformers (BERT) [17] combined with structural features from comment trees on Reddit's social platform to improve early post-level controversy detection performance. Zhong et al. [7] proposed a method based on a graph convolutional network (GCN) to achieve the same goal. It attempts to fuse the structural information of the comment tree with the textual information. It introduces concurrent multi-task classifiers in the model to distinguish topic-related and topic-irrelevant features for inter-topic detection. Benslimane et al. [8] proposed a controversy detection method based on discussion graph structure and textual features, using a graph neural network to encode the discussion graph. Two methods for detecting controversy are also provided. The first is representational learning of graphs using a hierarchical strategy. The second uses an attention mechanism to give different levels of importance to nodes and neighboring nodes. The drawback of the existing hybrid controversy detection methods is that they have limited access to structural features and ignore the influence of root information. The source of a controversial event is the topic, which is always informative and has a broad impact. Therefore, it is crucial to use the information on the topic effectively. To this end, we propose a dual feature enhancement graph convolutional network model that combines valuable textual and structural information and fully uses topical information. To compare the limitations of the various methods more clearly and intuitively, we summarize the different controversy detection methods in Table 1.

Table 1: Summary of controversy detection methods

Studies	Limitations
[9]	• Depending on nearest neighbors and naive Bayes classifiers, it may perform poorly in complex controversial situations.
[10]	• There is a greater reliance on the accuracy and effectiveness of the semantic difference metric.
[11]	• Important contextual information about other non-emotional words is ignored and may not be accurately identified for some controversial forms.
[12]	• It requires a substantial amount of user review data and additional information to support the analysis, which may present certain challenges in data acquisition and processing.
[13]	• Not widely applicable to different communities and cultural contexts and needs to be adapted and tailored to different contexts.

(Continued)

Table 1 (continued)

Studies	Limitations
[2,3,15,16]	• These approaches ignore the features of textual content and fail to capture the semantic information of controversial views and statements.
[6–8]	• Limited ability to acquire structural features and failure to learn deeper structural features. • The impact of topical messages is ignored and the topical features are not used effectively.

3 Method

This section describes post-level controversy detection methods, a classification task. Given the post text (Post) and the social media reply network structure graph G = (V, E), the controversy detection model $f_{controversy}(.)$ maps the input (text, G) to the set of tags Y = {controversial, non-controversial}, i.e., controversial posts and non-controversial posts. $f_{controversy}(\text{text, G}) = Y_{\{controversial\}}$. The main idea is to represent the exchange of user responses on social media as a graph of user responses and to iterate over the vector representation of the target nodes using GCN.

The overall structure diagram of the DFE-GCN model proposed in this paper is shown in Fig. 1, which consists of six modules: graph construction, global structure information encoding layer, semantic information encoding layer, feature fusion layer, local neighborhood information fusion and attention root node feature enhancement layer, and controversy classification layer. Especially for the feature fusion layer, direct concatenation of these features will lead to the failure of certain features due to their widely varying distribution. Therefore, we make use of the attention mechanism to better integrate features. In the following, we describe each module in depth.

3.1 Graph Construction

The purpose of using subject hashtags on social media is to help users choose areas of interest. Users can join discussions on particular topics by clicking on the hashtags that interest them. Topical discussions often include posting, forwarding, answering, and the like. As a result, the most typical information diffusion path in various social media is from themes to posts and then to comments. We first construct a topic-post-comment graph to simulate the message transmission chain between topics, posts, and comments. G = (V, E) for the target post, where G is an undirected graph, V is the collection of nodes, and E is the set of edges. The following describes the node and edge information.

Nodes: There are three kinds of nodes in G: topic nodes, post nodes, and comment nodes. The topic node can be thought of as a hub node for integrating and exchanging information. The root node also connects all post nodes for this theme. Then, each comment node is linked to the post or comment node to which it responds.

The topic-post-comment graph we construct has three different types of edges, and each node in the graph can represent a topic, a post, or a comment. The types of the three edges are described as follows:

Topic-Post edge: We consider the topic node a hub node for integrating and exchanging information. Then, we connect all the post nodes under the topic to a topic node to capture helpful information between related posts under the same topic.

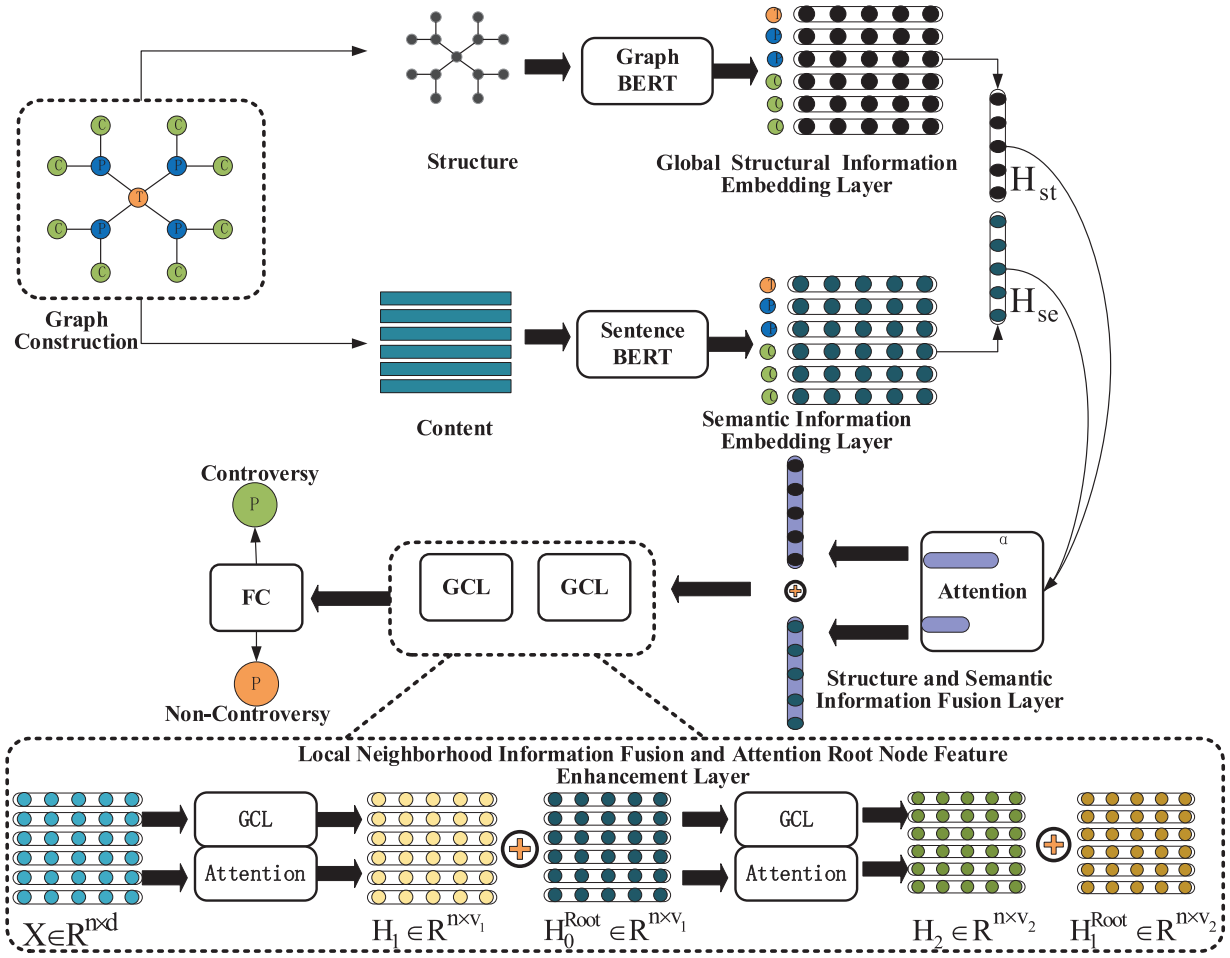

Figure 1: Structural diagram of the DFE-GCN model. The global structural information encoding layer and the local neighborhood information encoding layer extract multi-scale structural features of the response graph to enhance the representation of structural features, allowing access to deep structural features. The topic feature enhancement layer enhances the topic features using the attention mechanism to improve the impact of the root cause of controversial events

Post-Comment edge: We link each comment node to the post node it replies to maintain the relationship between posts and comments. It is possible to model the relationship between posts and comments.

Comment-Comment edge: The connection method of comment-comment edges is similar to that of post-comment edges, A edge $<V_i, V_j>$ will be created when V_j responds to the post P or any comment posted by V_i.

3.2 Global Structure Information Encoding Layer

Structural information is also crucial when identifying controversial events on social media. Nodes with controversial posts typically have more comment nodes and diverse neighborhood structures.

Intuitively, nodes with similar structures also tend to perform the same functions. Therefore, we must take a global perspective to find structural similarities between crucial nodes.

In this paper, we learn structural embeddings using Graph-Bert [18], a pre-trained model that can be applied to graph data and can help us find the structural similarity of key nodes from a global perspective. It does not consider information about edges in the graph by sampling the original graph into multiple subgraphs and learning representations of the subgraphs only using an attention mechanism. As it is not constrained by the graph structure, it can be well-pre-trained and migrated for learning. Formally, Graph-Bert learns a structural embedding matrix $H_{st} \in R^{N*F_{st}}$, which contains global structural information about the graph. Each node $h_{st} \in R^{F_{st}}$ is represented by an F_{st} dimensional vector.

3.3 Semantic Information Encoding Layer

The analysis of controversial speech cannot be separated from the study of the text's content. In deep learning-based controversy detection tasks, the first step of most algorithms represents the content of the text as a vector. The vector is then classified using a classification algorithm, with some using the vector as the input sequence to a recurrent neural network based on the chronological order of the posts. Text vectors are an integral part of a neural network's understanding of text content, and their quality will directly determine the performance of the model.

In this paper, we use the Sentence-BERT [19] model, a dedicated model for generating sentence or paragraph vectors, and output fixed-dimensional sentence vectors using a twin neural network architecture. Secondly, the creators of the Sentence-BERT model created the Sentence-Transformer library, which contains several sentence vector generation models based on various pre-trained models and trained on different corpora. Thus, based on the above models, a text semantic representation vector $h_{se} \in R^{F_{se}}$ is extracted for each node, which contains the hidden syntactic and semantic aspects of the text content. h_{se} can be constructed as a semantic embedding matrix $H_{se} \in R^{N*F_{se}}$ for each node in the network.

3.4 Feature Fusion Layer

As the distribution of values between different features can vary greatly, directly stitching the features from each part and feeding them into the graph neural network may lead to training instability and the loss of some important features. To solve these problems, we perform an attentional dynamic fusion of the acquired text features and structural features. We employ the following attention mechanism, as shown in Eqs. (1)–(3):

$$\mathcal{F}(h_s) = v^T \tan h(W_{\mathcal{F}} h_s + b_{\mathcal{F}}), s \in st, se \tag{1}$$

$$\alpha_s = \frac{\exp(\mathcal{F}(h_s))}{\sum_{s \in st,se} \exp p(\mathcal{F}(h_s))} \tag{2}$$

$$h_s^{att} = \alpha_s h_s, s \in st, se \tag{3}$$

where $W_{\mathcal{F}}$ is the weight parameter, $b_{\mathcal{F}}$ is the matrix of paranoid parameters, and F(.) Outputs the attention raw score, after which the adjusted semantic representation h_{se}^{att} and the structural representation h_{st}^{att} are concatenated to obtain the fusion representation h_{fu}. As shown in Eq. (4):

$$h_{fu} = \text{concat}\left(h_{st}^{att}, h_{se}^{att}\right) \tag{4}$$

3.5 Local Neighborhood Information Fusion and Attention Root Node Feature Enhancement Layer

GCN [20] is a multilayer neural network that processes graphical data and generates embedding vectors of nodes based on their neighbors. The GCN can capture information from nodes' direct and indirect neighbors through cascading layered convolutions. We input the above-generated text and structure fused features h_{fu} into the GCN, the hidden feature matrix H_1, which is the output of the first layer of the GCN, is defined as Eq. (5):

$$H_1 = \sigma\left(\hat{A}h_{fu}\left(W_0\right)\right) \tag{5}$$

where $\hat{A}$ is the normalized adjacency matrix, defined as Eq. (6):

$$\hat{A} = \widetilde{D}^{\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}} \tag{6}$$

where $\widetilde{A}$ is calculated from the adjacency matrix and the diagonal matrix I_N, defined as Eq. (7):

$$\widetilde{A} = A + I_N \tag{7}$$

$\widetilde{D}$ is the degree matrix corresponding to the adjacency matrix, defined as Eq. (8):

$$\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij} \tag{8}$$

$\sigma(\cdot)$ is the nonlinear ReLU activation function.

In controversy generation, the topic corresponding to the root node contains more important feature information relative to other nodes, and better use of root information and the relationship from the root node to other nodes is essential for controversy detection. Therefore, it is necessary to perform root node feature augmentation for the nodes in the reply graph. The process of its root node feature augmentation is as follows: For node i, calculate the root node's influence weight on it, as shown in Eq. (9):

$$\text{weight}_i = \text{softmax}\left(\omega\left(Uh_i, Uh_0\right)\right) \tag{9}$$

where $\omega(.)$ denotes the activation function, U is a trainable parameter matrix, and h_0 denotes the node root feature vector. The root node feature enhancement vector of node i is created based on this weight value. The calculation method is shown in Eq. (10):

$$h_i^r = \sigma\left(\text{weight}_i Uh_i\right) \tag{10}$$

Therefore, the root node enhancement matrix in the first layer is shown as Eq. (11):

$$H_k^{\text{Root}} = \left[\left(h_0^r\right)^T, \left(h_1^r\right)^T, \ldots, \left(h_{n-1}^r\right)^T\right]^T \tag{11}$$

The hidden matrix $\widetilde{H_1}$, after introducing the root node feature enhancement mechanism, is defined as Eq. (12):

$$\widetilde{H_1} = \text{concat}\left(H_1, H_0^{\text{Root}}\right) \tag{12}$$

Afterward, the hidden feature matrix H_2 of the second GCN layer is calculated using Eq. (5), and the hidden matrix $\widetilde{H_2}$ for introducing the root node attention mechanism is calculated according to Eqs. (9)–(12).

3.6 Controversy Classification Module

In this paper, a fully connected layer as well as the non-linear activation function softmax are applied to obtain the controversial probability of post nodes in the reply structure graph. It can be defined as Eq. (13):

$$P_i = \sigma\left(W_0 h_{post\_i} + b_0\right) \tag{13}$$

where W_0 and b_0 are both learnable parameters. Finally, in this paper, the loss is computed by comparing the controversy probability values with the true labels using the binary cross-entropy loss function, as shown in Eq. (14):

$$L_c = -\frac{1}{N}\sum_i \left(\left(1 - y_i^c\right)\log\left(1 - P_i\right) + y_i^c \log\left(P_i\right)\right) \tag{14}$$

where y_i^c is the true label value, and p_i is the predicted controversial probability value by the model. The algorithmic flow chart of the model is shown in Table 2.

Table 2: Algorithmic flow of the DFE-GCN model

Training process of dual feature enhanced graph convolutional network model
Given: The structural graph G = (V, E), the structural representation matrix H_{st}, and the semantic representation matrix H_{se}.
Objective: Model parameters Θ.
for each epoch **do**
Randomly sample a mini batch
for each node i **do**
$h_{st}^{att}, h_{se}^{att} \leftarrow$ Attention(h_{st}, h_{se})
$h_{fu} \leftarrow$ Concatenate(h_{st}^{att}, h_{se}^{att})
end for
for each node i **do**
for each GCN layer j **do**
$h_i \leftarrow GCN_j\ (h_{fu}, A)$
$h_i^r \leftarrow$ Attention(h_{fu})
$\widetilde{h}_i \leftarrow$ Concatenate(h_i, h_i^r)
$h_{fu} \leftarrow \widetilde{h}_i$
end for
$p_i \leftarrow$ Output($\widetilde{h}_{post\_i}$)
end for
Θ ← BackProp(L, y_i, p_i, Θ) ΔAdam step
end for

4 Experiments

The following will provide detailed explanations of the motivation and implementation details of the experimental settings in the controversy detection task. The experiments in this paper are divided into the following three parts:

1. Comparative experiments: We compared our method with current mainstream methods for controversy detection (content-based, structure-based, and hybrid) to investigate whether our method has superior performance in detecting controversial social media posts.

2. Ablation experiment: This experiment removes different modules from the proposed method, involving a global structure encoding module and an attentional topic feature enhancement module, to investigate the effect of different modules in the model on the experimental results.

3. Network layer analysis: This experiment performs a layer sensitivity analysis for the number of layers of the model to explore the layer settings suitable for the model to achieve the best performance.

4.1 Dataset

We conducted experiments on two real datasets in two different languages. Table 3 displays partial statistical information on the two datasets.

Table 3: Dataset statistics

Number	Weibo	Reddit
Topics (Hashtags/Subreddits)	49	6
Controversial posts	1992	7515
Non-controversial posts	3684	7518
All posts	5676	15033
Comments of controversial posts	35632	578879

Weibo Dataset [7]: The dataset is a Chinese dataset for social media post controversy detection, which contains 49 topics from different domains, such as "Candidates respond to being rejected three times by Peking University", "Proposal to abolish motor vehicle driving license", etc. There are a total of 1992 controversial posts and 3684 non-controversial posts. The posts in each topic fall under that topic, and each post contains at least 2 comments and up to 15 comments. In addition, the dataset's author has retained the structure of the replies to the dataset.

Reddit Dataset [6]: The Reddit dataset was collected from 2007 to February 2014, and the data covers six specific online channels, namely AskMen (am), AskWomen (aw), Fitness (fn), LifeProTips (lt), PersonalFinance (pf), and Relationships (rs). On Reddit, each user can post comments under threads of posts about a specific topic (subreddit). Each topic contains a set of posts, with metadata and a comment structure associated with each post. Noting that posts with low comments may not be constructed into a significant graph, only posts with at least 30 comments are retained.

4.2 Experimental Environment and Hyperparameters

The experimental environment is shown in Table 4 and the experimental hyperparameters are shown in Table 5.

Table 4: Hardware and software in the experiment

Hardware/environment	Description
GPU	TITAN RTX
OS	Ubuntu18.04
Framework	Pytorch
RAM	32G
CPU	Intel(R) Xeon(R) Gold 5120 CPU @ 2.20 GHz

Table 5: Experimental hyperparameters

Models layer	2
Epoch	200
Hidden dimension	200
Learning rate	0.0001
Optimizer	Adam
Batch size	8
Dropout rate	0.15

4.3 Evaluation Metrics

This article treats controversy detection as a binary classification problem and evaluates performance using Precision (P), Recall (R), Accuracy (ACC), and F1.

The Precision metric compares the proportion of accurately predicted controversy samples to all predicted controversy samples. The calculation process is shown in Eq. (15):

$$P = \frac{TP}{TP + FP} \tag{15}$$

The Recall metric compares the proportion of correctly predicted controversy samples to all controversy samples. The calculation process is shown in Eq. (16):

$$R = \frac{TP}{TP + FN} \tag{16}$$

The Accuracy metric computes the percentage of correct prediction samples among all samples. The calculation process is shown in Eq. (17):

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

F1 was employed as an overall measure of model performance metrics because Precision and Recall are frequently not positively associated and are often contradictory. The calculation process is shown in Eq. (18):

$$F1 = 2 * \frac{P \cdot R}{P + R} \tag{18}$$

To demonstrate the validity of the method proposed in this paper and fully evaluate its performance, experiments were conducted to answer the following questions:

RQ1: Can the controversy detection method proposed in this paper improve the performance of controversy detection?

RQ2: Can the method capture deeper structural features?

RQ3: What is the effect of root feature attention enhancement on performance improvement?

4.4 Baseline

In this paper, the proposed controversy detection method, which combines reply graph structure and text features with topic feature enhancement, is compared with multiple baseline algorithms to verify its performance. We used three representative methods, including content-based methods, structure-based methods, and hybrid methods, and their descriptions are as follows:

1. Structure-based controversy detection methods:

GCN [20]: We employed a two-layer basic GCN model with non-linear activation and dropout between them.

Graph Attention Network (GAT): GAT [21] is a novel neural network architecture that employs multi-head attention techniques. We employed a GAT model with two layers and used non-linear activation and dropout between them.

2. Content-based controversy detection methods:

BERT: For the controversy identification work, we fine-tuned multilingual BERT, which has been pre-trained in 104 languages and has learned text vectors with strong universal representation.

Text Convolutional Neural Network (TextCNN): TextCNN [22] is a famous text categorization model. It employs one-dimensional convolution to obtain n-gram feature representation in the sentence, and the embedding layer features pre-trained word vectors.

3. Fusion methods:

Topic-Post-Comment Graph Convolutional Network (TPC-GCN): For post-level controversy detection, TPC-GCN [7] combines data from the graph's structure with the content of topics, posts, and comments. BERT is used in this technique to extract text features, while graph convolutional networks are used to learn structural features.

4.5 Results and Analysis

To answer RQ1, we compared the performance of all methods on both datasets. Tables 6 and 7 show the detection performance of the proposed model compared to other state-of-the-art methods. The proposed method in this paper achieves better performance with higher precision, recall, F1, and accuracy than the other methods. DFE-GCN and TPC-GCN use a hybrid structure incorporating structural and semantic features. Therefore, the detection performance is generally higher than those using only semantic or structural features. Compared to TPC-GCN, DFE-GCN provides a higher level of performance by enhancing structural and thematic features. On the Weibo dataset, the F1 value improved by 1.99 percentage points, and the accuracy improved by 5.92 percentage points. On the Reddit dataset, the F1 value increased by 2.17 percentage points, and the accuracy rate increased by 3.32 percentage points. This is a result of capturing the deep structural features of the reply structure graph as well as leveraging topic information. However, content-based methods, such as TextCNN

and BERT, score low on both datasets. We believe that the simple concatenation of the textual content of topics, posts, and comments during preprocessing of the data makes the semantics of the text extremely complex, which is not conducive to the understanding of the model and causes low detection performance. Among the content-based approaches, the BERT model outperforms TextCNN on both datasets because BERT can provide the better semantic understanding and contextual relevance, while TextCNN is relatively limited in understanding the semantic and contextual aspects of the text. We also find that the structure-based approach outperforms the semantic feature-based approach, illustrating that the structural information of the response graph is important for our controversy detection task.

Table 6: Controversy detection capability of different methods on *Weibo*

Method	Precision	Recall	F1	Accuracy
GCN	0.6515	0.6627	0.6570	0.7210
GAT	0.6761	0.6612	0.6685	0.7677
TextCNN	0.6180	0.6193	0.6186	0.6783
BERT	0.6317	0.6472	0.6393	0.7135
TCP-GCN	0.6717	0.6922	0.6811	0.8158
Ours	0.7006	0.7015	0.7010	0.8750

Table 7: Controversy detection capability of different methods on *Reddit*

Method	Precision	Recall	F1	Accuracy
GCN	0.6193	0.6347	0.6261	0.6637
GAT	0.6283	0.6222	0.6287	0.7056
TextCNN	0.5653	0.5577	0.5615	0.5833
BERT	0.5830	0.5897	0.5863	0.6284
TCP-GCN	0.6487	0.6544	0.6514	0.7327
Ours	0.6743	0.6720	0.6731	0.7659

4.6 Ablation Experiment

To explore the impact of different parts of the model on the experimental results, we also conducted ablation experiments. The experiments were conducted by removing different modules as follows:

1. w/o Att: The attention mechanism in the topic feature enhancement module was removed, and only simple feature concatenation was used for topic node feature enhancement to treat each node in the response graph equally.

2. w/o Att_Topic: The topic feature attention enhancement module in the model was removed, that is, the topic features were not enhanced, and the fused features were directly passed through two layers of GCN.

3. w/o Structure Encoder: The global structural encoding module in the model was removed, and only the semantic features of the text were used as the input of the model.

To answer RQ2 and RQ3, we conducted ablation experiments to explore the contributions of the different modules. The results in Tables 8 and 9 show that the model with root node feature augmentation improves significantly in all metrics compared to the model without root node augmentation. Building upon this foundation, we enhance the features of the root node by utilizing attention mechanisms, resulting in further performance improvement. This is because different comments or posts have different levels of importance for the topic. Treating every node in the comment graph equally leads to a certain loss of accuracy. Enhancing the root node's features using an attention mechanism allows for assigning different weights to other nodes based on the varying degrees of influence from the root node. This helps alleviate the problem to some extent. When we removed the global structure encoding module from the model, the F1 score dropped by 1.81 percentage points, and the accuracy dropped by 2.18 percentage points on the Weibo dataset. On the Reddit dataset, the F1 score decreased by 1.13 percentage points, and the accuracy dropped by 2.01 percentage points, indicating that adding the global structure encoder to our approach can extract deeper structural features and positively impact model recognition.

Table 8: The impact of different parts of the model on *Weibo dataset*

Condition	Precision	Recall	F1	Accuracy
Ours	0.7006	0.7015	0.7010	0.8750
w/o Attention	0.6892	0.6917	0.6902	0.8535
w/o Attention Topic	0.6693	0.6957	0.6822	0.8477
w/o Structure Encoder	0.6795	0.6864	0.6829	0.8532

Table 9: The impact of different parts of the model on *Reddit dataset*

Condition	Precision	Recall	F1	Accuracy
Ours	0.6743	0.6720	0.6731	0.7659
w/o Attention	0.6766	0.6641	0.6703	0.7541
w/o Attention Topic	0.6633	0.6574	0.6605	0.7478
w/o Structure Encoder	0.6649	0.6720	0.6618	0.7458

4.7 Layer Analysis

The above ablation study demonstrates the superior performance of our proposed dispute detection method in terms of structural feature enhancement and topic feature enhancement. To further evaluate this approach, we conducted experiments on GCN layer count analysis. We tried 1, 2, 3, 4, and 5 GCN layers, as shown in Figs. 2 and 3, and we chose to set the number of GCN layers to two, which provided the maximum performance of our method. However, as the number of GCN layers increases, we find that the extraction performance of the model gradually decreases. We believe this is due to the model's inability to understand the nodes' features due to the excessive number of layers. In the GCN, each update aggregates data from surrounding nodes to adjust the node features. As the number of GCN layers increases, the 'aggregation radius' of each node also increases. This means that the node covers more and more neighboring nodes, resulting in a lack of diversity in the local network structure of the node, which is not conducive to the learning of the node's features. Therefore, increasing the number of GCN layers will reduce the extraction performance of the model.

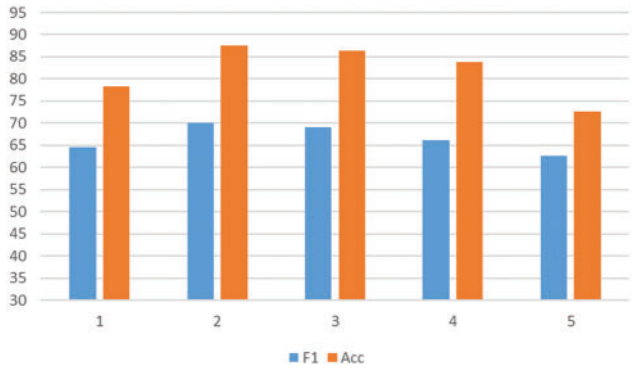

Figure 2: Analysis of the number of layers on the *Weibo dataset*

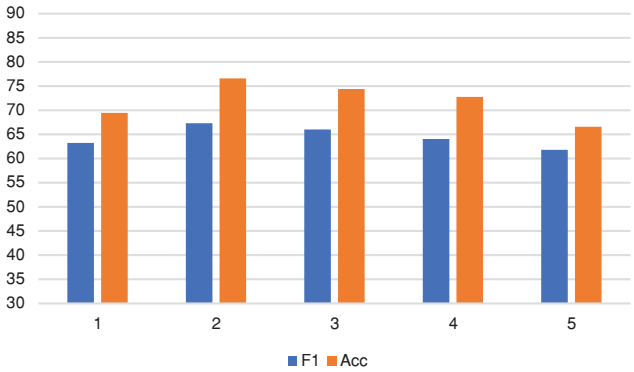

Figure 3: Analysis of the number of layers on the *Reddit dataset*

5 Conclusion

Detecting controversies is a significant challenge in social networks and online information analysis. This study proposes a dual-feature enhanced graph convolutional network for controversy detection. It utilizes a global structure coding module and a local structure coding module to obtain multi-scale structural features for enhanced representation of structural features. In addition, to fully use the rich topic information, we augment the topic features behind each graph convolutional layer with an attention mechanism. Finally, the update vectors of posts and comments are combined to predict controversies. Experimental results on two real datasets show that the proposed DFE-GCN approach achieves good results in the controversy detection task, outperforming other baseline models. The accuracy and F1 score were improved by 5.92% and 1.99%, respectively, on the Weibo dataset and by 3.32% and 2.17%, respectively, on the Reddit dataset. These changes demonstrate that incorporating enhanced structural and topic features has a positive impact on controversy detection. Afterward, we conducted ablation experiments, and the experimental results further confirmed the effectiveness and rationality of each module in DFE-GCN. In future work, we will further explore the potential features of controversies and try to take more modal information into account to achieve multimodal controversy detection.

Acknowledgement: Thank you to our tutors and researchers for their guidance and assistance throughout this research process. Additionally, we would like to express our gratitude to the anonymous reviewers for providing valuable comments.

Funding Statement: This research was funded by the Natural Science Foundation of China Grant No. 202204120017, the Autonomous Region Science and Technology Program Grant No. 2022B01008-2, and the Autonomous Region Science and Technology Program Grant No. 2020A02001-1.

Author Contributions: Study conception and design: C. Hua, F. Wei; data collection: C. Hua, Y. Liao, K. HaiLaTi; analysis and interpretation of results: C. Hua, W. Yang, L. Wang; draft manuscript preparation: C. Hua. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data used in this study are public datasets, which can be obtained from the following links: Weibo dataset: https://pan.baidu.com/s/1McUURxWhQc54zk9yblT0Mw?pwd=93ht or https://drive.google.com/file/d/1SvB3qRkT-N745W44ZW44gfeAevTmsyTz/view?usp=sharing.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Dori-Hacohen, "Controversy analysis and detection," Ph.D. Dissertation, University of Massachusetts Amherst, USA, 2017.

[2] K. Garimella, G. D. F. Morales, A. Gionis and M. Mathioudakis, "Reducing controversy by connecting opposing views," in *Proc. of the Tenth ACM Int. Conf. on Web Search and Data Mining*, New York, NY, USA, pp. 81–90, 2017.

[3] K. Garimella, G. D. F. Morales, A. Gionis and M. Mathioudakis, "Quantifying controversy on social media," *ACM Transactions on Social Computing*, vol. 1, no. 1, pp. 1–27, 2018.

[4] A. J. Morales, J. Borondo, J. C. Losada and R. M. Benito, "Measuring political polarization: Twitter shows the two sides of Venezuela," *Chaos*, vol. 25, no. 3, pp. 33114, 2015.

[5] M. Jang and J. Allan, "Improving automated controversy detection on the web," in *Proc. of the 39th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, New York, NY, USA, pp. 865–868, 2016.

[6] J. Hessel and L. Lee, "Something's brewing! Early prediction of controversy-causing posts from discussion features," in *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, USA, pp. 1648–1659, 2019.

[7] L. Zhong, J. Cao, Q. Sheng, J. Guo and Z. Wang, "Integrating semantic and structural information with graph convolutional network for controversy detection," 2020. [Online]. Available: https://aclanthology.org/2020.acl-main

[8] S. Benslimane, J. Azé, S. Bringay, M. Servajean and C. Mollevi, "A text and GNN based controversy detection method on social media," *World Wide Web*, vol. 26, no. 2, pp. 799–825, 2023.

[9] B. Sznajder, A. Gera, Y. Bilu, D. Sheinwald, E. Rabinovich *et al.,* "Controversy in context," arXiv preprint arXiv:1908.07491, 2019.

[10] S. Dori-Hacohen, D. Jensen and J. Allan, "Controversy detection in wikipedia using collective classification," in *Proc. of the 39th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, New York, NY, USA, pp. 797–800, 2016.

[11] Y. Choi, Y. Jung and S. H. Myaeng, "Identifying controversial issues and their sub-topics in news articles," in *Proc. of the 2010 Pacific Asia Conf. on Intelligence and Security Informatics*, Berlin, Heidelberg, Germany, pp. 140–153, 2010.

[12] A. Sriteja, P. Pandey and V. Pudi, "Controversy detection using reactions on social media," in *2017 IEEE Int. Conf. on Data Mining Workshops*, New Orleans, LA, USA, pp. 884–889, 2017.

[13] M. Jang, J. Foley, S. Dori-Hacohen and J. Allan "Probabilistic approaches to controversy detection," in *Proc. of the 25th ACM Int. on Conf. on Information and Knowledge Management*, New York, NY, USA, pp. 2069–2072, 2016.
[14] S. Benslimane, J. Azé, S. Bringay, M. Servajean and C. Mollevi, "Controversy detection: A text and graph neural network based approach," in *Web Information Systems Engineering*, Melbourne, VIC, Australia, pp. 26–29, 2021.
[15] P. Guerra, W. Meira Jr, C. Cardie and R. Kleinberg, "A measure of polarization on social media networks based on community boundaries," in *Proc. of the Int. AAAI Conf. on Web and Social Media*, Washington, USA, pp. 215–224, 2013.
[16] M. Mendoza, D. Parra and Á. Soto, "GENE: Graph generation conditioned on named entities for polarity and controversy detection in social media," *Information Processing & Management*, vol. 57, no. 6, pp. 102366, 2020.
[17] J. Devlin, M. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, USA, pp. 4171–4186, 2019.
[18] J. Zhang, H. Zhang, C. Xia and L. Sun, "Graph-Bert: Only attention is needed for learning graph representations," arXiv preprint arXiv:2001.05140, 2020.
[19] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-Networks," in *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing*, Hong Kong, China, pp. 3982–3992, 2019.
[20] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.
[21] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio *et al.,* "Graph attention networks," arXiv preprint arXiv:1710.10903, 2017.
[22] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1746–1751, 2014.