GS2F: Multimodal Fake News Detection Utilizing Graph Structure and Guided Semantic Fusion

DONG ZHOU, School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China
QIANG OUYANG, School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China
NANKAI LIN*, School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China
YONGMEI ZHOU, School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China
AIMIN YANG, Guangdong University of Technology, Guangzhou, China

The prevalence of fake news online has become a significant societal concern. To combat this, multimodal detection techniques based on images and text have shown promise. Yet, these methods struggle to analyze complex relationships within and between modalities due to the diverse discriminative elements in the news content. In addition, research on multimodal and multi-class fake news detection remains insufficient. To address the above challenges, in this paper, we propose a novel detection model, GS^{2}F, leveraging *g*raph *s*tructure and *g*uided *s*emantic *f*usion. Specifically, we construct a multimodal graph structure to align two modalities and employ graph contrastive learning for refined fusion representations. Furthermore, a guided semantic fusion module is introduced to maximize the utilization of single-modal information and a dynamic contribution assignment layer is designed to weigh the importance of image, text, and multimodal features. Experimental results on Fakeddit demonstrate that our model outperforms existing methods, marking a step forward in the multimodal and multi-class fake news detection.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Computer vision representations**; • **Applied computing** → **Document management and text processing**.

Additional Key Words and Phrases: Fake news detection, Graph structure, Multi-view learning, Social media analysis.

*Nankai Lin is the corresponding author.

Authors' Contact Information: Dong Zhou, School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, Guangdong, China; e-mail: dongzhou@gdufs.edu.cn; Qiang Ouyang, School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, Guangdong, China; e-mail: ouyangq0011@163.com; Nankai Lin, School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China; e-mail: 229737513@qq.com; Yongmei Zhou, School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, Guangdong, China; e-mail: yongmeizhou@163.com; Aimin Yang, Guangdong University of Technology, Guangzhou, Guangdong, China; e-mail: amyang@gdut.edu.cn.

ACM 2375-4702/2024/12-ART
https://doi.org/10.1145/3708536

1 Introduction

The popularity of microblogging and short-video sharing has led to the proliferation of social media platforms such as Facebook [1], Twitter [2], Raddit [3] and Youtube [4]. The platforms facilitate rapid and extensive communication. However, they have also enabled the spread of fake news, causing significant damage to social order [42]. For instance, some public figures intentionally disseminate fake news for attention [4]. It was also a disruptive factor in the 2016 US presidential election [3]. Misinformation about diagnoses or vaccinations during pandemics can even incite social panic [18]. Therefore, early detection and prevention of fake news dissemination has become a recent priority and consequently research into fake news detection has gained considerable attention [34].

Early research in fake news detection focused predominantly on textual content analysis, using statistical text features for detection on social platforms [2, 25]. However, most online social content has recently evolved from plain text to multimodal forms incorporating text, images or videos. Recent years have seen preliminary progress in multimodal fake news detection research. Some studies address the underutilization of intra-modal information by extracting and processing relevant data to obtain detailed features. For instance, Fu et al. [13] concentrate on entity information in news text to enhance semantic comprehension of intramodal features. Singh et al. [35] manually design text and image features across four dimensions, content, organization, emotion as well as manipulation, and fuse them for improving fake news detection results. Wang et al. [41] extract a total of 16 features from text, images, and users to distinguish fake news. Other studies primarily focus on fusing multimodal information by combining representation features from different modalities. Wu et al. [46], for example, use image-extracted features with a multi-layered shared attention mechanism for coarse-grained fusion from frequency and spatial domains. Xiong et al.[47] propose a dual fusion mechanism that manipulates image features through text-image correlation to boost interaction between multimodalities. Wang et al. [44] propose the Event Adversarial Neural Network (EANN), which combined image and text features and leverages adversarial learning with event classification to extract common features across various events and improve detection accuracy. Chen et al. [6] examine the inherent ambiguity between image and text modalities through the lens of information theory to enhance the efficacy of fake news detection. Zheng et al. [51] introduce a social graph structure and employ a graph neural network to aggregate features, combined with a hierarchical fusion approach for fake news prediction.

Although significant progress has been made in multimodal fake news detection, several challenges remain:

- Firstly, much of the existing research [44, 46] focuses on coarse-grained fusion of text and image data, which may overlook nuanced details and fail to capture intricate intra-modal and inter-modal relationships that are important discriminative clues. Such oversights can be exploited by fake news creators who craft content with discrepancies that are not easily detected through general consumption habits prioritizing images and text messages. At the same time, extracting information from the entire image for fusion is overly general. Although coarse-grained features provide global semantic context, they are limited in capturing fine-grained implicit connections between modalities and struggle to fully express the logical dependencies between entity-level features in text and images. Therefore, the model must also focus on the interconnections between these entities to achieve a more comprehensive and accurate understanding, thereby improving its discriminative ability.
- Secondly, recent studies [6, 47] have highlighted the importance of integrating image-text content to identify cross-modal correlations. However, unlike other multimodal tasks, both fake and real news more or less exhibit cross-modal correlations. In such cases, cross-modal correlations may not necessarily play a

[1] https://www.facebook.com/
[2] https://twitter.com/
[3] http://reddit.com/
[4] https://www.youtube.com/

critical role. Although single-modal cues may reveal some deceptive practices, different model features show different effects in the decision-making procedures [37], and leveraging both single-modal and multimodal information effectively remains an area for further investigation.

- Lastly, current studies predominantly address binary classification in multimodal fake news detection [10, 26, 40], despite the existence of various types of fake news [24]. Finer-grained classifications could yield deeper insights into its origins, propagation mechanisms, and the development of more robust prevention strategies. While multi-class detection has been explored in text-based contexts [8, 14], similar research on multimodal fake news is scarce. Applying multi-class models from other domains to this challenge has not yielded optimal results. Thus enhancing multi-class detection capabilities for multimodal fake news represents a critical research direction.

To address the above challenges, in this paper, we propose a novel multimodal fake news detection model, GS^{2}F, which leverages ***g***raph ***s***tructures and ***g***uided ***s***emantic ***f***usion to achieve superior performance in multi-class fake news detection. Our model employs two principal design strategies:

(1) We enhance multimodal content fusion by extracting both global and local features from text and images. For text, we construct a graph using syntactic dependency trees derived from semantic dependency information. For images, we generate a graph based on spatial location information for each entity region. We utilize the output from the [CLS] token of the pre-trained model and the features extracted from the full image to represent the global features of text and images, respectively. Global features provide comprehensive insights for an overall understanding of the news content. Image and text representations share semantic information but exist in different feature spaces. To align feature distances, contrastive learning has been used to adjust modality features in the absence of misinformation [43]. Subsequently, we consider the information of inconsistencies across modalities by weighting edges connecting modal entities according to their similarity degrees and employ dual graph contrastive learning objectives to uncover intra- and inter-modal semantic relationships. Specifically, we utilize label-supervised graph contrastive learning guided by news labels to refine graph representational features, thereby improving class discrimination for enriched fusion features. For instance, within the same type of news, cross-modal fusion reveals certain commonalities among the news items. This exploration becomes particularly important when dealing with a greater variety of types. It aids the model in learning shared contextual features from multimodal data that are relevant for assessing the veracity of news [15]. Simultaneously, self-supervised graph contrastive learning is implemented to discover an optimal graph structure and derive robust multimodal graph representations. Through graph networks, we can effectively model the relationships between nodes and augment the graph structure, a process that is both feasible and easy to implement. This enables the derivation of multiple, yet similar, graph structures, which facilitates the identification of a broader range of features. In summary, using multimodal graph-structured fused features provides a clearer expression of semantic information across modalities, enhancing the model's ability to distinguish between classes in complex classification tasks. This approach offers better generalization, transferability, and robustness when learning clues from news content.

(2) Our previous model primarily focused on cross-modal information fusion. To improve the discriminative capacity of single-modal features, we introduce a guided semantic fusion module that utilizes supplementary information to enrich single-modal features. Given that news content often depends on domain-specific knowledge not readily extracted from domain-free datasets, GS^{2}F incorporates external knowledge as background information for complementing the news content analysis. Furthermore, effective extraction is facilitated by applying Discrete Cosine Transform (DCT) [17], especially in scenarios involving tampered image content [46]. We further employ a standard attention mechanism to independently model the intrinsic semantics of each modality. Then, we assign appropriate contribution weights to single-modal

and multimodal features so that single-modal features of text and images can augment overlooked aspects during multimodal feature fusion.

Through these strategies, GS^{2}F effectively identifies fake news with enhanced ability by capturing intricate patterns across text and image modalities. It detects news content fusion from three perspectives: text features, image features, and multimodal features. Additionally, it integrates domain-specific knowledge into the analysis framework.

In summary, our work makes the following significant contributions:

(1) We propose a fine-grained multimodal graph structure to enhance the detection of fake news across multiple modalities. It can explicitly learn the logical dependencies between text words and image regions. Graph contrastive learning is also introduced to implicitly capture interactions among modalities.
(2) We design a new guided semantic fusion module that enriches the representation of image and text features. Through our novel dynamic contribution assignment approach, we effectively merge content from three dimensions, optimizing the use of semantic information across modalities.
(3) We conduct extensive experiments on the representative multi-class dataset Fakeddit [24], and the experimental results demonstrate its superior performance, setting new benchmarks in accuracy for this task.

2 Relate work

2.1 Unimodal Fake News Detection

Fake news detection methods were categorized into single-modal and multimodal detection based on the utilized mode. Recent years have witnessed extensive research on both these methods, leveraging deep learning techniques. In single-modal analysis, the detection of fake news primarily relied on single-modal information to capture potential features for categorization. Traditional approaches centered on extracting statistical and semantic information from post content. For instance, Rashkin et al. [30] focused on textual features and identified specific words commonly present in fake news. Ma et al. [22] employed recurrent neural networks (RNN) to learn temporal representations from contextual information in related posts. Chen et al. [5] added an attention mechanism to the recurrent neural network (RNN), directing focus to various temporal linguistic features. Dong et al. [11] proposed a two-path deep semi-supervised learning framework, with one path undergoing supervised learning on limited labeled data, and another path engaging in unsupervised learning on unlabeled data, enhancing false news detection. Emotional features in the text were also considered in fake news detection, revealing evident emotional biases [1, 9]. Beyond sentiment analysis, researchers endeavored to enhance detection accuracy by integrating textual information with other rumor cues, including comment information[31] and conversation structure [48].

2.2 Multimodal Fake News Detection

Prior research on fake news detection has predominantly concentrated on text-based information. With the increasing prevalence of multimedia, attention is shifting towards multimodal fake news detection to better reflect current trends. This area has been explored by various scholars who have recognized the untapped potential of multimodal content for improving the accuracy of fake news detection models. For instance, Zhang et al. [50] proposed a novel bi-emotional feature descriptor that assessed the sentiment discrepancy between publishers and commenters to distinguish between fake and real news. Qi et al. [27] highlighted limitations in previous studies, noting that they overlooked non-textual information embedded in images that cannot be captured by pre-trained extractors. To address this gap, they manually extracted such information to complement textual analysis. However, as many real-world images lack accompanying text, relying solely on this type of data may not be feasible for widespread application in feature extraction.

Multimodal fake news detection necessitates the consideration of interactions among modalities and the feature learning pertinent to news content. Singhal et al. [36] utilized pre-trained language models and the ImageNet model for feature extraction, subsequently concatenating the features and inputting them into a stacked fully connected layer for classification. However, the direct concatenation cannot effectively learn multi-modal mutual interaction. The researchers examined the correlations between different modes and improved the fusion model by introducing more advanced aggregation methods. For instance, Wu et al. [46] utilized a superimposed cross-attention module with mixed feature vectors for classification, employing coarse-grained modular vectors that yielded satisfactory results in fake news identification. Qian et al. [28] implemented two context converters with multiple attention heads and layer configurations to capture multimodal interactions, demonstrating notable efficacy in multimodal fake news detection tasks. Chen et al. [6] considered the inherent ambiguity of different modes by aggregating single-modal features and cross-modal correlations, enhancing fake news detection accuracy. Zhang et al. [49] improved model performance by segmenting images into patches, integrating them with textual data, and applying contrastive learning at the final stage. Wu et al. [45] explored fine-grained semantics in the text modality, while Wang et al. [10] attempted fine-grained extraction of both text and images. Müller-Budack et al. [23] also considered more entity details in their analyses. Graph neural networks were employed to aggregate information from different nodes. Dhawan et al. [10] utilized graph structure to establish connections among all entities within the intra-model, employing a graph attention network for fusion interaction to detect fake news. However, complete connections within the intra-model may introduce redundancy, potentially impacting the inherent semantic relationships. Fine-grained features can reveal more details in news content. Considering the correlation between these detailed clues can aid in detecting news. However, if we overly rely on pre-trained object detection models to identify target entities, it may result in missing objects whose classes are not predefined by the model. Although global coarse-grained features offer a coarser representation, they still provide global semantic information and can complement object-level features to some extent. In our model, we have also considered this aspect, leveraging coarse-grained features to provide comprehensive global semantic information while facilitating semantic interactions between entities through fine-grained object-level features.

Recently, Nakamura et al. [24] introduced a fine-grained fake news detection task and presented the Fakeddit dataset, which distinguishes not only between true and fake news but also subdivides the latter into five detailed categories based on the reasons behind misinformation. The complexity of these detailed labels presents challenges for accurate classification. Our preliminary experiments indicate that multi-classification methods from other domains are suboptimal for this nuanced task in fake news detection. Therefore, it is crucial to develop sophisticated and integrated fusion techniques to improve multimodal fake news detection efficacy. This presents a significant research opportunity.

3 Methodology

In this section, we describe the proposed GS^{2}F that leverages ***g***raph ***s***tructures and ***g***uided ***s***emantic ***f***usion. Our model processes multimodal news samples containing image information I and text information T, each tagged with a label Y. We categorize label Y into binary, ternary, and hexary classes for our experiments. Figure 1 depicts the framework of our proposed model. We begin by extracting both global and local features from individual modalities in the dataset. A multimodal graph is then constructed to encapsulate the logical dependencies between these features, facilitating an initial aggregation of multi-granularity information. Graph contrastive learning is employed to learn key intra- and inter-modal features. The guided semantic fusion module is subsequently harnessed to refine single-modal feature representations. Later, a dynamic contribution assignment layer is introduced to integrate single-modal and multimodal features by assigning weights according to their relevance. These features are used for the final fake news classification.

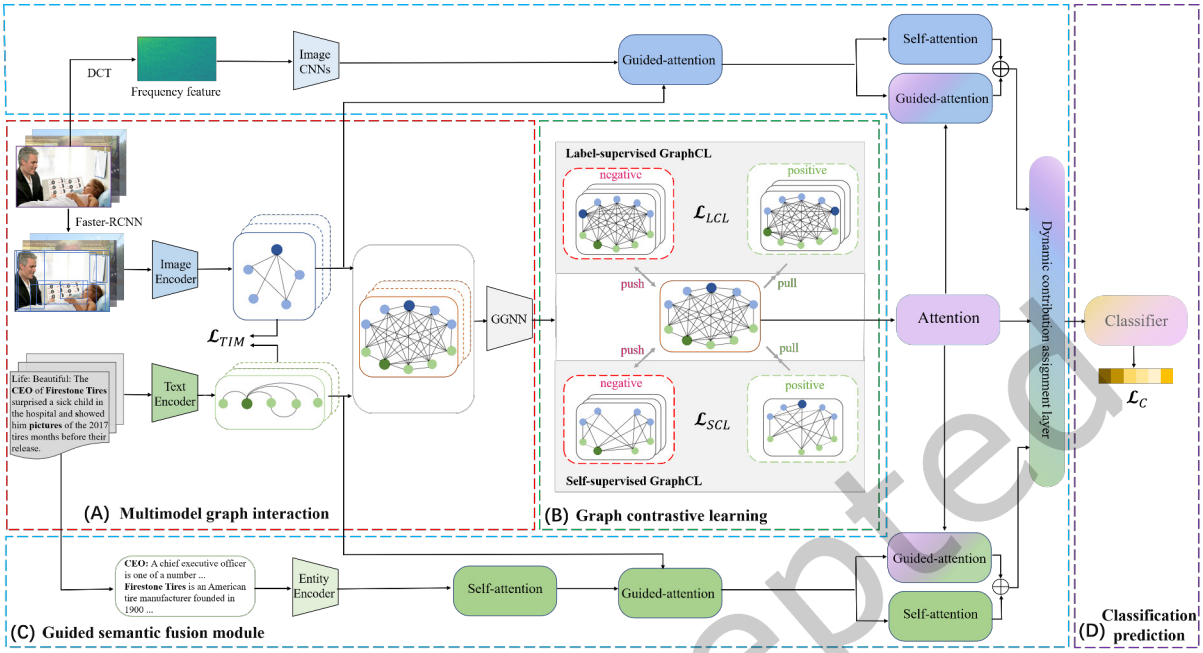


Fig. 1. The architecture of the proposed framework GS^{2}F.

3.1 Multimodel graph interaction

3.1.1 Feature extraction. To obtain sentence-level and word-level text representations, we employ BERT [16] on news text T. Specifically, the output vectors of each token as well as $[CLS]$ are employed to capture feature representations. The final output from the language model is formulated as follows:

$$X^T = [x_0, x_1, x_2, \ldots, x_L] \tag{1}$$

where X^T denotes the encoded representation of the original input text. L is the length of the text. The output corresponding to the $[CLS]$ token is utilized as the sentence-level semantic feature of text modality. To ensure uniformity in dimensionality across multimodal features and to effectively encapsulate sequential context, we employ a BiGRU network to process the textual modality features.

$$H^T = \left[h_0^T, h_1^T, h_2^T, \ldots, h_L^T\right] = BiGRU(X^T) \tag{2}$$

where $h_i^T \in \mathbb{R}^{d_h}$ represents the feature corresponding to the i-th word in the text and d_h is the dimension of the hidden layer. H^T denotes text modal features that contain sentence-level and word-level features of the text.

To extract fine-grained features from local regions in the image, we utilize the Faster R-CNN model [32] to detect specific objects in image I. We select the top r entities based on their confidence scores to construct entity-level features for these regions. These features are then fed into a vision pre-trained Swin Transformer model [21] to extract local image modal characteristics. Recognizing that reliance on object detection models may result in missed objects, we complement the region-specific features with global image descriptors to provide comprehensive semantic representation. The final image representation is thus formally defined as follows:

$$H^I = \left[h_0^I, h_1^I, h_2^I, \ldots, h_r^I\right] \tag{3}$$

where $h_j^I \in \mathbb{R}^{d_h}$ is the feature representation of the j-th region object, d_h is the dimension of the hidden layer. r is the number of the region object. H^I denotes a feature representation containing both global and local information in the image.

3.1.2 Graph construction. To exploit intra- and inter-modal semantic relationships effectively, we utilize a graph-based approach to connect features across modalities. Different from Dhawan et al. [10] utilized all nodes connected within the intra-model. We consider semantic correlations among tags in a text sentence and spatial correlations among region objects in an image. Moreover, semantic correspondences between words in the sentence and region objects in the image can be present. Such relationships contribute to improved news detection performance. Therefore, we construct three graphs: a text-modality graph, an image-modality graph, and a multimodal graph. In these structures, edges represent intra- and inter-modal connections, whereas nodes correspond to features from the text and images.

Text-Modality graph: We construct a text-modality graph $G^T \in \mathbb{R}^{(L+1)\times(L+1)}$ to capture the semantic relationships inherent in textual data. It consists of semantic feature nodes at word-level and sentence-level. This graph is derived from the syntactic dependency tree, which leverages syntactic structures to enhance textual comprehension. If a relationship exists between two words in the dependency tree, the corresponding words are connected, facilitating logical reasoning in the text and fully connecting the information from text semantic and word levels. The weight of the edges in the text graph is set to 1.

Image-Modality graph: We construct an image-modality graph $G^I \in \mathbb{R}^{(r+1)\times(r+1)}$ for each image-modality instance, utilizing the Intersection over Union (IoU) score [33] to highlight spatial correlations among visual objects while preserving information on the region locations. Some region feature nodes may lack connecting edges while the global image node is connected to all region nodes.

Multimodal graph: To capture the relationship between textual and visual features in multimodal news, we construct a multimodal graph $G^M \in \mathbb{R}^{(r+L+2)\times(r+L+2)}$. This graph serves to integrate text and image information by modeling intermodal relationships. It is crucial to address both consistency and inconsistency present in the content of text-image pairs. Accordingly, an edge is established between every text node and its corresponding image node, with similarity scores employed to quantify their relatedness, thereby capturing associations in each modality. The similarity score ranges from 0 to 1. Subsequently, we add 1 to the original value on the cross-modal edges to better focus on cross-modal node aggregation while avoiding the model overly emphasizing relationships between unimodal features. At the same time, this adjustment ensures that even with low similarity, each edge in the graph still exerts a certain influence, preventing the omission of some information. These edges are instrumental in enabling the model to learn from both intra- and inter-modal dependencies. The architecture of this multimodal graph is delineated below:

$$G_{ij}^M = \begin{cases} G_{ij}^T, \text{ if } G_{ij}^T > 0 \\ G_{ij}^I, \text{ if } G_{ij}^I > 0 \\ Sim(h_i, h_j) + 1, \text{ if } i < L+1, j \geq L+1 \\ \qquad \text{or } i \geq L+1, j < L+1 \\ 0, \text{ otherwise,} \end{cases} \tag{4}$$

where $i < L+1$ and $j \geq L+1$, or $i \geq L+1$ and $j < L+1$, the nodes represent the association values between text and image. For $i < L+1$ and $j < L+1$, the graph represents values from the text graph G^T. For $i > L+1$ and $j > L+1$, it represents values from the image graph G^I.

3.1.3 Modal alignment. Bridging the heterogeneity gap across various modalities necessitates mapping the feature spaces of each modality into a common semantic space. Recent studies [43] have investigated the use

of contrastive learning to align modalities by pulling in paired image-text data and pushing away unpaired instances. In our approach, we select a subset of real news in a mini-batch, resulting in the following text-to-image contrastive loss:

$$\mathcal{L}_{T->I} = \mathrm{E}_{\mathrm{p}}\left[-\log \frac{\exp\left(S\left(H^{T}, H^{I+}\right)/\tau\right)}{\sum_{k=1}^{K}\exp\left(S\left(H^{T}, H_{k}^{I-}\right)/\tau\right)}\right] \tag{5}$$

where H^{I+} represents the image features that match the text features H^T, and H^{I-} represents the non-matching image features. Similarly, the image-to-text contrastive loss is $\mathcal{L}_{I->T}$, so the final loss for modality alignment is as follows:

$$\mathcal{L}_{TIM} = \frac{\mathcal{L}_{T->I} + \mathcal{L}_{I->T}}{2} \tag{6}$$

3.1.4 Graph aggregation. The Gating Graph Neural Network (*GGNN*) [20] is then employed to aggregate multimodal graphs to learn inter- and intra-modal dependencies. During information propagation, the *GGNN* selectively filters node information, enhancing the identification of salient features. For a multimodal graph G^M with text node features H^T and image node features H^I, the process is formalized as follows:

$$m = pool(GGNN([H^{T}; H^{I}], G^{M})) \tag{7}$$

where, $W_n \in \mathbb{R}^{d_h \times d_h}$, $b_n \in \mathbb{R}^{d_h}$ are trainable parameters and $m \in \mathbb{R}^{d_h}$ is a multimodal fusion feature.

3.2 Graph contrastive learning

To enhance the model's ability to extract key features from news data and refine multimodal fusion representations, we amplify disparities in the fusion representations across different classes [15]. We introduce two fine-grained graph contrastive learning objectives that leverage multimodal graphs to fuse text and image features. Specifically, we utilize label-supervised graph contrastive learning guided by news labels to refine graph representational features, thereby improving class discrimination for enriched fusion features. Simultaneously, self-supervised graph contrastive learning is implemented to discover an optimal graph structure, resulting in robust representational features.

3.2.1 Label-supervised graph contrastive learning. To facilitate the learning of shared features in multimodal news content and to exploit class relationships for refining initial feature representations, we introduce a label-supervised graph contrastive learning objective. This objective enhances model training by reinforcing intra-class similarities and inter-class distinctions. Each mini-batch contains N samples, and we use the multimodal initial fusion feature M as the anchor. In the feature space, the anchor and its positive samples should be brought closer, while the anchor and negative samples should be pushed away. Positives are defined as samples in the same class, while negatives encompass samples from different classes. The positive and negative sets for each class in the mini-batch are denoted as $P(m)$ and $N(m)$. Therefore, the loss function of label-supervised contrastive learning can be defined as:

$$\mathcal{L}_{LCL} = -\frac{1}{|P(m)|}\left[\sum_{m_p \in P(m)} \log \frac{\exp\left(S\left(m, m_p\right)/\tau\right)}{\sum_{m_n \in N(m)}\exp\left(S\left(m, m_n\right)/\tau\right)}\right] \tag{8}$$

where $S(\cdot) = m_i(m_j)^T$ is used to calculate the similarity between two vectors, m_p is a positive sample with the same label and m_n is a negative sample with different labels. τ is a temperature coefficient.

3.2.2 Self-supervised graph contrastive learning. To improve the robustness of our model and its capacity to learn invariant features from the data, we introduce a self-supervised graph contrastive learning objective. We construct positive samples by manipulating nodes and edges in multimodal graphs. By randomly adding or removing edges and deleting nodes, we generate a set of similar graph structures $\bar{G}$. These varied structures facilitate the model's ability to implicitly discern complex intra- and inter-modal relationships, thereby enabling the identification of a broader range of features. Simultaneously, the model's improved capacity to make accurate judgments when encountering new and unseen fake news instances increases detection accuracy and robustness. Later, we employ a Gated Graph Neural Network (*GGNN*) with shared weights for feature aggregation, which integrates text and image features into a unified graph representation denoted as z. For each sample in a mini-batch, the enhanced graph $\bar{G}$ multimodal feature is treated as the positive sample z_i and all other features are considered negative samples. Consequently, the self-supervised graph contrastive loss is defined as:

$$l_s(m_i, z_i) = \frac{\exp\left(S\left(m_i, z_i\right)/\tau\right)}{\exp(S(m_i, z_i)/\tau) + \sum_{n=1, n\neq i}^{N}(\exp(S(m_i, m_n)/\tau) + \exp((S((m_i, z_n)/\tau)} \tag{9}$$

$$\mathcal{L}_{SCL} = \frac{1}{2N}\sum_{i=1}^{N}(l_s(m_i, z_i) + l_s(z_i, m_i)) \tag{10}$$

where m_i represents the multimodal fusion feature of the i-th sample in the mini-batch.

3.3 Guided semantic fusion module

To address the limitations of relying solely on multimodal features, which may overlook the discriminative capabilities inherent to single-modal data, we propose a single-modal perspective-oriented guided semantic fusion module. This module is structured into three distinct components: an information extraction component for augmenting single-modal data, a guided semantic interaction to facilitate meaningful integration of modalities, and a contribution allocation layer that assigns feature-based weights. The subsequent sections provide detailed descriptions of each component.

3.3.1 External information extraction. To enhance the semantic richness of single-modal data and bolster the capability to distinguish single-modal features, we apply external entity knowledge to consider the context description information of the text as a means to explain and complement the entities. For each news item, we utilize the *TAGME* [12] toolkit for entity recognition and extract background descriptions ($B = [b_1, b_2, b_3, \ldots, b_e]$) from Wikipedia. This information is then fed into BERT to acquire a feature representation $B^T \in \mathbb{R}^{e\times d_b}$ for the descriptive text of entities.

$$B^T = BERT(B) \tag{11}$$

In addition, given that images in fake news are often subject to re-compression or manipulation in the frequency domain [46], we utilize the Discrete Cosine Transform (DCT) to extract pertinent frequency-domain features. We begin by resizing image I to a 224×224 resolution and dividing it into non-overlapping 28×28 blocks. The number of acquired blocks is denoted as $t = 64$. Subsequently, we apply DCT to each block to obtain frequency-domain representations. These representations are then fed into a Convolutional Neural Network (CNN) for feature extraction. Specifically, we adopt the Inception V3 [38] architecture to derive a feature vector $C^I \in \mathbb{R}^{t\times d_c}$, which encapsulates critical information from the frequency domain for further analysis.

To achieve dimension consistency, a linear projection is employed to map the inputs to the fully connected layer:

$$\bar{B^T} = B^T W_1 + b_1 \tag{12}$$

$$\bar{C^I} = C^I W_2 + b_2 \tag{13}$$

where $W^1 \in \mathbb{R}^{d_b \times d_h}$,$W^2 \in \mathbb{R}^{d_c \times d_h}$,$b^1 \in \mathbb{R}^{d_h}$,$b^2 \in \mathbb{R}^{d_h}$ are trainable parameters.

3.3.2 Guided semantic interaction. In addition to multimodal data, single-modal data can also enhance prediction performance [52]. To extract essential information from each modality, we propose a guided semantic interaction and establish a single-modal branch for preserving and integrating features of individual modalities. Specifically, we employ a multi-head attention mechanism [39] to improve the representational capacity of each modality. In this mechanism, Q denotes the query presenting the source information, which is informed by the knowledge in keys(K) and values(V). The operation proceeds as follows:

$$Attn(Q, K, V) = Softmax(\frac{QK^\top}{\sqrt{d}})V \tag{14}$$

$$head_n = Attn(QW_h^Q, KW_h^K, VW_h^V) \tag{15}$$

$$MH(Q, K, V) = (head_1; \ldots; head_H)W^O \tag{16}$$

$$MHL(Q, K, V) = LayerNorm(MH(Q, K, V) + Q) \tag{17}$$

where h denotes the number of attention heads and the projection matrix $W_h^Q \in \mathbb{R}^{d_h \times d_h}$, $W_h^K \in \mathbb{R}^{d_h \times d_h}$, $W_h^V \in \mathbb{R}^{d_h \times d_h}$, $W^O \in \mathbb{R}^{d_h \times d_h}$ are trainable parameters.

Taking the textual information as an example, we initially employ self-attention to capture internal long-range dependencies and structural features in the text description. We aim to facilitate semantic interactions in text by directing the flow of descriptive knowledge into single-modal representations, thereby improving feature recognition capabilities. The enhancement process for the text feature T^m is detailed as follows:

$$B^m = MHL(\bar{B^T}, \bar{B^T}, \bar{B^T}) \tag{18}$$

$$T^m = MHL(H^T, B^m, B^m) \tag{19}$$

The feature representation of the text perspective will be conducted from two angles: one involves obtaining an in-depth representation of the text mode through self-attention, referred to as H_{slf}^T. The other involves promoting guided semantic interactions between text and multimodal features to derive feature H_{crs}^T. By calculating the correlation between text and multimodal features, the model identifies interrelated components and sensitively distinguishes subtle differences. The model flexibly selects and focuses on information relevant to specific contexts, integrates useful visual signals, and enhances key information in textual features, thereby improving overall recognition performance. This approach is taken to preserve the integrity and diversity of semantic information in the modality, thereby enhancing recognition ability. To enhance the global perception capability of the multimodal feature, we also add the attention interaction to multimodal feature m, as follows:

$$H_{slf}^m = MHL(m, m, m) \tag{20}$$

$$H_{slf}^T = MHL(T^m, T^m, T^m) \tag{21}$$

$$H_{crs}^T = MHL(T^m, H_{slf}^m, H_{slf}^m) \tag{22}$$

Finally, the representations H_{slf}^T and H_{crs}^T are subjected to average pooling followed by concatenation. The textual representation is characterized as follows:

$$H_{fus}^T = \left[pool(H_{slf}^T) \oplus pool(H_{crs}^T)\right] \tag{23}$$

Based on Eqs. (20)-(23), we derive the image perspective feature, denoted as H_{fus}^I, which integrates external information and exhibits strong discriminative capability.

3.3.3 Dynamic contribution assignment layer. In preparation for classification, we acknowledge that single-modal and multimodal features do not contribute equally to the detection task. Specifically, certain multimodal news items rely heavily on the integration of multimodal fusion features, whereas other instances may benefit more from enhanced single-modal image and text features. To address this disparity in feature contribution, we introduce a dynamic contribution assignment layer to model the importance of different features. This method trains the system to highlight key components pertinent to news discrimination, thereby yielding weighted and diverse representations. The learnable parameter U is dynamically tuned during training to effectively handle diverse multimodal news content. The vector $Z = [H^T_{fus}; H^m_{slf}; H^I_{fus}]$ represents the composite representation of the image, text, and multimodal fusion features. We assign contribution weights to each feature and these weights are later used to proportionally scale the respective features prior to their integration into a unified representation.

$$H_z = ZW_z + b_z \tag{24}$$

$$A_z = Softmax((UW_u)(H_zW_k)^{\mathrm{T}}) \tag{25}$$

$$\bar{H_z} = A_z(H_zW_v) \tag{26}$$

where A_z denotes the weight of each feature, meanwhile $\bar{H_z}$ is the weighted superposition of there features, $W_z \in \mathbb{R}^{d_h \times d_h}$, $W_u \in \mathbb{R}^{d_h \times d_h}$, $W_k \in \mathbb{R}^{d_h \times d_h}$, $W_v \in \mathbb{R}^{d_h \times d_h}$, $b_z \in \mathbb{R}^{d_h}$ are trainable parameters. We decompose $\bar{H_z}$ into three distinct weighted features, which are then concatenated to form the final fusion feature f.

3.4 Classification prediction

The fused feature vector f is fed into a multilayer perceptron (MLP) layer to reduce its dimensionality. A softmax function is then applied to derive the predicted probabilities $\hat{y}$ for the detection task. The cross-entropy loss function is utilized to compute the prediction loss. The prediction process is formulated as follows:

$$\hat{y} = Softmax(MLP(f)) \tag{27}$$

$$\mathcal{L}_C = \sum_{k-1}^{n} y_k log(\hat{y}_k) \tag{28}$$

where y represents the true label and n represents the number of classification classes. In summary, the total loss of our model is defined as follows:

$$\mathcal{L} = \mathcal{L}_C + \alpha(\mathcal{L}_{LCL} + \mathcal{L}_{SCL}) + \beta\mathcal{L}_{TIM} \tag{29}$$

where α and β serve as hyperparameters that modulate the influence of the contrastive learning loss and the modal alignment loss, respectively.

4 Experiment

4.1 Dataset

The Fakeddit dataset [24] represents the most extensive multimodal collection derived from the social networking platform Reddit. This dataset encompasses over one million samples drawn from 22 diverse subreddits, ranging from political news to typical user posts. It offers classifications across 2-class, 3-class, and 6-class schemes. We evaluate our model on all schemes. 2-class labels indicate the authenticity of the news, 3-class labels expand the fake label to fake news with true text and with false text, and 6-class labels are detailed in Fig. 2. To maintain data relevance, entries lacking either image or text were excluded. Given the voluminous nature of Fakeddit, a subset constituting 10% of the multimodal content was randomly selected for experimental purposes, ensuring the retention of original class proportions. The training-to-test set ratio was preserved at 10:1 as per dataset standards. The chosen classification granularity mirrors real-world prevalence patterns of such news types.

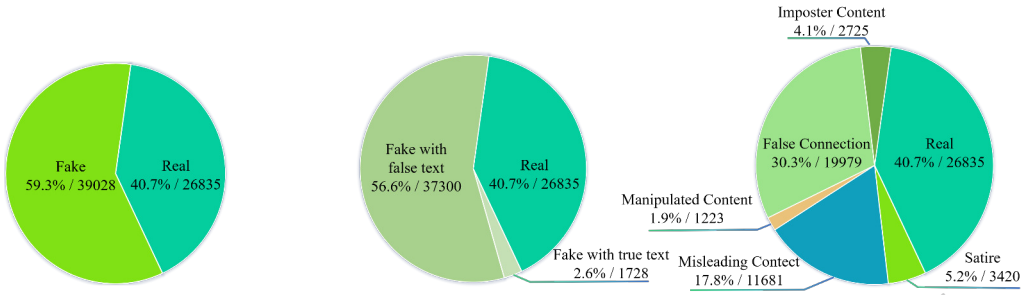


Fig. 2. Distribution of classes in the Fakeddit dataset

4.2 Baseline models

To evaluate the performance of our model, we compare it with state-of-the-art fake news detection models as well as multi-class detection models from various domains.

- **Text (single-modal)**: This model exclusively leverages textual data and employs BERT for feature extraction and classification.
- **Visual (single-modal)**: This model utilizes solely image modality for detection purposes. An image is input into a Swin Transformer to obtain its hidden representation, following which classification predictions are made through a fully connected layer coupled with a Softmax layer.
- **MCAN** [46]: This model exploits multimodal features by incorporating frequency domain information of images. It employs a stacked attention mechanism for effective interactive fusion, thereby enhancing classification performance.
- **BTIC** [49]: This model segments the image into patches and extracts multimodal features from both textual information and these patches. It also selects representative samples for contrastive learning during training.
- **TRIMOOM** [47]: This model emphasizes the importance of textual information and leverages text-image correlations to enhance image features. It employs text features to guide the feature fusion process, subsequently utilizing the integrated features for fake news detection.
- **LIIMR** [37]: This model extracts fine-grained features from images and text to establish inter-modal relationships, facilitating the extraction of salient information from dominant modalities to enhance prediction accuracy.
- **MTTV** [40]: This model improves the utilization of visual images by integrating both global and local entity features. It utilizes a multi-layer Transformer architecture to facilitate extensive interaction with multimodal data.
- **CLIP** [29]: The model jointly trains image and text modalities through contrastive learning, generating respective image and text representations, and optimizes by maximizing the similarity between correct image-text pairs. It effectively retrieves and classifies across modalities, demonstrating excellent performance across multiple tasks.
- **BLIP** [19]: The model employs a self-supervised pre-training approach to learn multimodal associations between images and text, achieving cross-modal information fusion via a bidirectional Transformer structure. It demonstrates significant performance improvements in multimodal tasks. Given the relevance to multimodal fake news detection, we selected the BLIP model's parameters, trained on the visual question answering (VQA) task, to optimize its application in this area.

- **CBAN** [7]: This model is a multi-classification model equipped with a bipolar attention mechanism to effectively accommodate both consistent and inconsistent information in images and text. It demonstrates robust performance across various multi-class datasets, including applications in crisis judgment and emotion analysis.
- **ITIN** [53]: This multi-classification model is utilized for multimodal emotion analysis, leveraging a cross-modal alignment module to capture the associations between regions and words. Multimodal features are integrated via an adaptive cross-modal gating mechanism that incorporates contextual visual information, thereby enhancing performance.

4.3 Experiment settings

Experiments are conducted using the PyTorch framework [5]. Image features are extracted with a pre-trained Swin Transformer model on ImageNet [21], and text features are obtained using the BERT-base model. The object detection model identifies 20 image regions (r = 20). We set the hyperparameters α to 0.15. β is the weight of the control mode alignment, which needs only to be set to 0.1. Both edge modification and point deletion ratios are fixed at 0.1. To prevent overfitting, a normalization layer is applied after the fully connected layers in both text and image modalities. The dropout rate is set to 0.5. The model is optimized using the AdamW optimizer with a learning rate of $1e-4$, a batch size of 64, and 100 epochs. For evaluation purposes, we adopt macro-level metrics due to imbalanced class distribution in the dataset.

[5]https://github.com/pytorch/pytorch

Table 1. Main results

	Method	Accuracy	Precision	Recall	F1-score
2-class	Text	0.848	0.840	0.846	0.843
	Visual	0.778	0.773	0.781	0.774
	BTIC	0.867	0.860	0.865	0.862
	MCAN	0.863	0.859	0.856	0.857
	TRIMOOM	0.873	0.867	0.874	0.870
	BLIP	0.862	0.862	0.860	0.861
	CLIP	0.886	0.883	0.881	0.882
	LIIRM	0.881	0.879	0.874	0.876
	MTTV	0.884	0.881	0.878	0.879
	GS^{2}F	**0.896**	**0.889**	**0.893**	**0.891**
3-class	Text	0.833	0.823	0.830	0.826
	Visual	0.765	0.757	0.718	0.736
	BTIC	0.853	0.858	0.832	0.845
	MCAN	0.852	0.849	0.830	0.837
	TRIMOOM	0.870	0.848	0.866	0.856
	BLIP	0.843	0.869	0.843	0.854
	CLIP	0.880	0.886	0.825	0.856
	MTTV	0.878	0.859	**0.884**	0.871
	CBAN	0.863	0.859	0.854	0.856
	ITIN	0.866	0.882	0.857	0.868
	GS^{2}F	**0.885**	**0.894**	0.876	**0.885**
6-class	Text	0.760	0.660	0.581	0.605
	Visual	0.729	0.627	0.511	0.534
	BTIC	0.813	0.732	0.629	0.646
	MCAN	0.806	0.607	0.633	0.619
	TRIMOOM	0.828	0.775	0.662	0.689
	BLIP	0.816	0.739	0.606	0.634
	CLIP	0.865	0.702	0.646	0.667
	MTTV	0.855	0.739	0.718	0.728
	CBAN	0.831	0.721	0.668	0.691
	ITIN	0.850	0.779	0.654	0.693
	GS^{2}F	**0.870**	**0.801**	**0.724**	**0.751**

4.4 Main results

In this section, we assess the performance of our GS^{2}F model by benchmarking it against a range of detection methods. The results are presented in Table 1. Our evaluation metrics included accuracy, precision, recall, and F1 score. The findings reveal that:

(1) GS^{2}F outperforms other models in both binary and multiclass scenarios with accuracies of 89.6%, 88.5%, and 87.0% for 2-, 3-, and 6-class classifications respectively, demonstrating its superiority across different tasks as evidenced by top-ranking accuracy, recall, and F1 scores.

(2) The multimodal approaches significantly surpass single-modal methods with at least a 1.5%, 1.9%, and 4.6% increase in accuracy for the respective classifications, highlighting the benefits of exploiting multimodal data.
(3) Various models such as MCAN utilize coarse-grained semantic information from multimodal data while TRIMOOM focuses on text features to guide image processing and MTTV emphasizes image information utilization rate. Our GS^{2}F exceeds these methods by integrating multi-granularity information through graph structures, applying graph contrastive learning to extract deep semantic connections. It can effectively enhance detection via three interactive perspectives.
(4) Models originally intended for multi-class tasks generally perform better in 6-class detection compared to those developed for 2-class classification. GS^{2}F not only excels at 2-class classification task but also demonstrates superior performance in 6-class challenges due to its comprehensive integration of multi-modal information. It also substantially improves single-modal data usage efficiency thus augmenting discriminative capacity.

4.5 Ablation studies

We conduct experiments to evaluate the impact of various model components on overall performance. The following modifications were applied:

- The removal of the text description and the frequency domain feature information of the image (w/o external text and image).
- The exclusion of label-contrastive and self-contrastive tasks, which omits graph contrastive learning objectives to assess their effect on model performance (w/o label-contrastive and w/o self-contrastive).
- The omission of the Guided Semantic Fusion (GSF) module entails eliminating single-modal branches, thereby relying solely on multimodal features for detection (w/o GSF).
- Disabling the Dynamic Contribution Allocation layer (DCA), resulting in feature fusion through simple concatenation instead of a dynamic contribution assignment mechanism (w/o DCA).

Table 2 summarizes the performance impact of removing certain components from our model. The findings are as follows:

- Omitting text descriptions and frequency domain features results in reduced model performance, underscoring the value of integrating single-modal information for enhancement.
- Our model demonstrates superiority over variants without label-contrastive and self-contrastive methods, affirming the significance of graph contrastive learning in identifying shared features across news classes and capturing complex intra-modal relationships through diverse graph structures.
- The advantage of including GSF is evident as our model outperforms configurations without GSF, highlighting the importance of single-modal information in augmenting overall performance. These features not only complement multimodal inputs but also direct attention to salient details, thereby reinforcing the benefits of leveraging both types of features for improved detection capabilities.
- Directly merging single-modal with multimodal features without a mechanism to weigh their importance may hinder prompt prioritization between them. In contrast, our proposed configuration outperforms those without DCA, confirming that allocating feature contributions effectively enhances detection performance.

Table 2. Ablation studies

	Method	Accuracy	F1-score
2-class	w/o additional text	0.893	0.887
	w/o additional image	0.894	0.888
	w/o label-contrastive	0.893	0.886
	w/o self-contrastive	0.887	0.884
	w/o GSF	0.882	0.879
	w/o DCA	0.889	0.887
	GS^{2}F	**0.896**	**0.891**
3-class	w/o additional text	0.881	0.874
	w/o additional image	0.883	0.877
	w/o label-contrastive	0.878	0.869
	w/o self-contrastive	0.876	0.869
	w/o GSF	0.872	0.865
	w/o DCA	0.881	0.876
	GS^{2}F	**0.885**	**0.885**
6-class	w/o additional text	0.867	0.744
	w/o additional image	0.868	0.747
	w/o label-contrastive	0.865	0.737
	w/o self-contrastive	0.862	0.733
	w/o GSF	0.862	0.728
	w/o DCA	0.865	0.742
	GS^{2}F	**0.870**	**0.751**

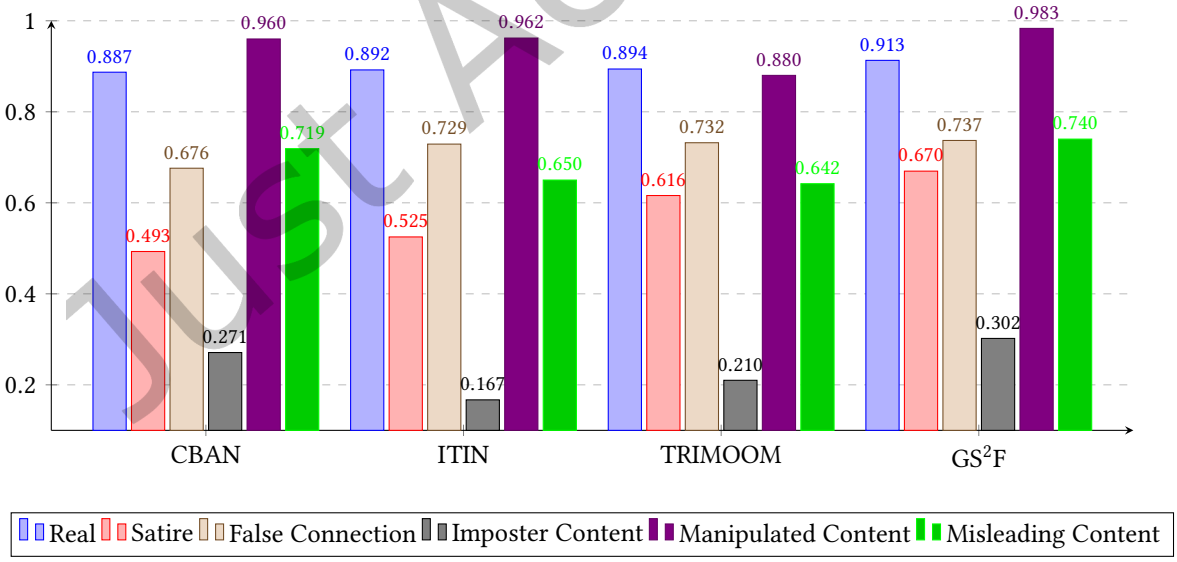


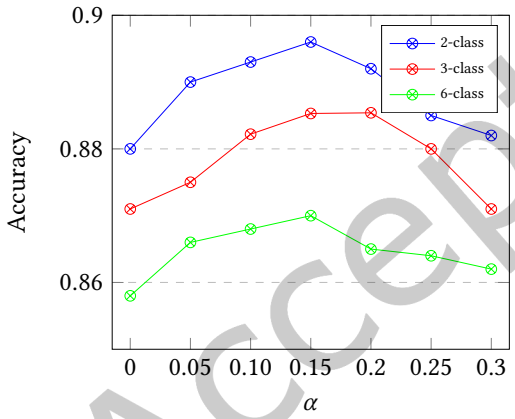
Fig. 3. The classification results of GS^{2}F and other models

4.6 Analysis of results

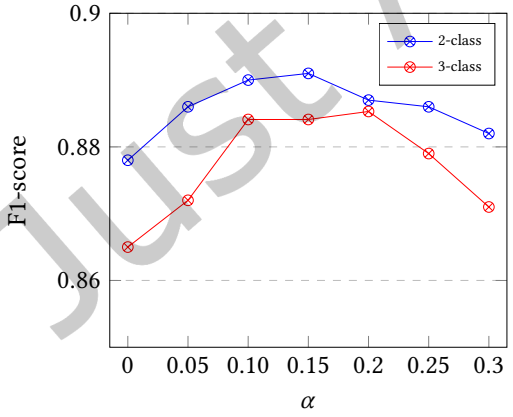
To investigate the detection performance of the model based on the 6-class classification of Fakeddit, we considered the detection effect for each class, as detailed in Fig. 3. In class detection, the accuracy of a class is equivalent to the recall value, signifying the proportion of correctly classified samples within this class.

(1) GS^{2}F performs more effectively in classifying Real and Manipulated Content, among various other classes. The likely explanation for this superiority is the heightened comprehensiveness of our model. Our model directs attention to salient details, thereby reinforcing the benefits of leveraging multiple types of features for improved detection capabilities.

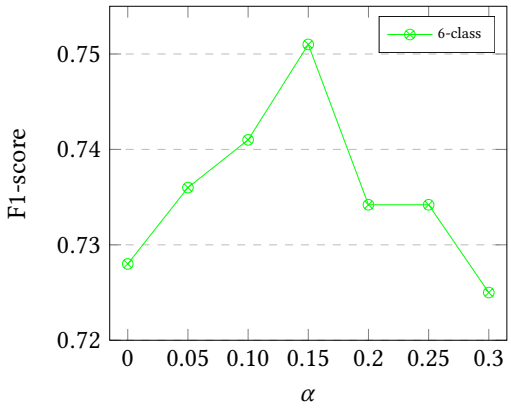
(2) Satire and Imposter Content exhibit a limited sample size, and the news within these classes is characterized by high secrecy. Most models produce poor predictions in these classes and therefore require more attention.



(a) Accuracy of GS^{2}F with different α



(b) F1-score of GS^{2}F with different α



(c) F1-score of GS^{2}F with different α

Fig. 4. Performance of GS^{2}F with different α values in graph contrastive learning.

4.7 Analysis of the weight parameter α

We evaluate the influence of the weight parameter α in graph contrastive learning on the performance of multimodal fake news detection models. The experimental outcomes for 2-class, 3-class, and 6-class detection tasks with varying values of α are depicted in Figure 4. The proposed model achieves peak performance at $\alpha = 0.15$, 0.2, and 0.15 for 2-class, 3-class, and 6-class detection respectively. Performance enhancement is initially observed with an increase in α. However, beyond a certain threshold, additional increases in α result in decreased effectiveness. This trend suggests that excessive weighting may cause the model to overly prioritize clustering behavior over accurate fake news identification. Consequently, it is crucial to judiciously select the weight parameter to ensure optimal training of the model.

4.8 Visualization analysis

To demonstrate the effectiveness of our model in multimodal fake news detection, we apply t-distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction to visualize the class separability achieved by our model. Figure 5, comprising subfigures (a), (b), and (c), displays the results of ablation studies where we sequentially removed key components: graph contrastive learning, the guided semantic fusion module, and analyzed the complete model. Distinct classes are indicated by different colors in these visualizations. Notably, subfigure (c) shows enhanced class separation compared to subfigure (a), highlighting the crucial role that graph contrastive learning plays in enabling our model to capture common discriminative features for news classification. Furthermore, when compared with a variant lacking the guided semantic fusion module, our full model exhibits improved separability. Subfigure (b) reveals some overlap between real and fake news features which suggests that an excessive reliance on correlated multimodal features might degrade classification performance. Our comprehensive model leverages a rich set of discriminative features from both single-modal and multimodal interactions, significantly boosting its classification capabilities.

To directly evaluate our classification performance, we utilize the Calinski-Harabasz index, a well-established clustering metric. A higher index indicates clearer cluster separation and thus better classification. Table 3 shows that the model without graph contrastive learning achieves higher Calinski-Harabasz scores than the model without the guided semantic fusion module, reflecting superior classification. Furthermore, our complete model's scores marginally outperform both aforementioned configurations, highlighting the benefits of integrating graph contrastive learning and guided semantic fusion modules into our framework.

Table 3. Calinski-Harabasz index of different methods

Methods	Calinski-Harabasz index
w/o label-self-contrastive	4884.62
w/o GSF	4574.58
GS^2F	5634.43

4.9 Case study

We randomly select some samples to assess the impact of various kinds of features. Our model demonstrates high accuracy in classifying these samples. Fig. 6 showcases four samples, which illustrate the contribution weights assigned by our model's dynamic contribution assignment layer for each feature. Samples (a) and (b) highlight the importance of single-modal information with greater contributions allocated accordingly. For instance, in sample (a), an image shows a grasshopper on a car window while accompanying text erroneously describes a giant

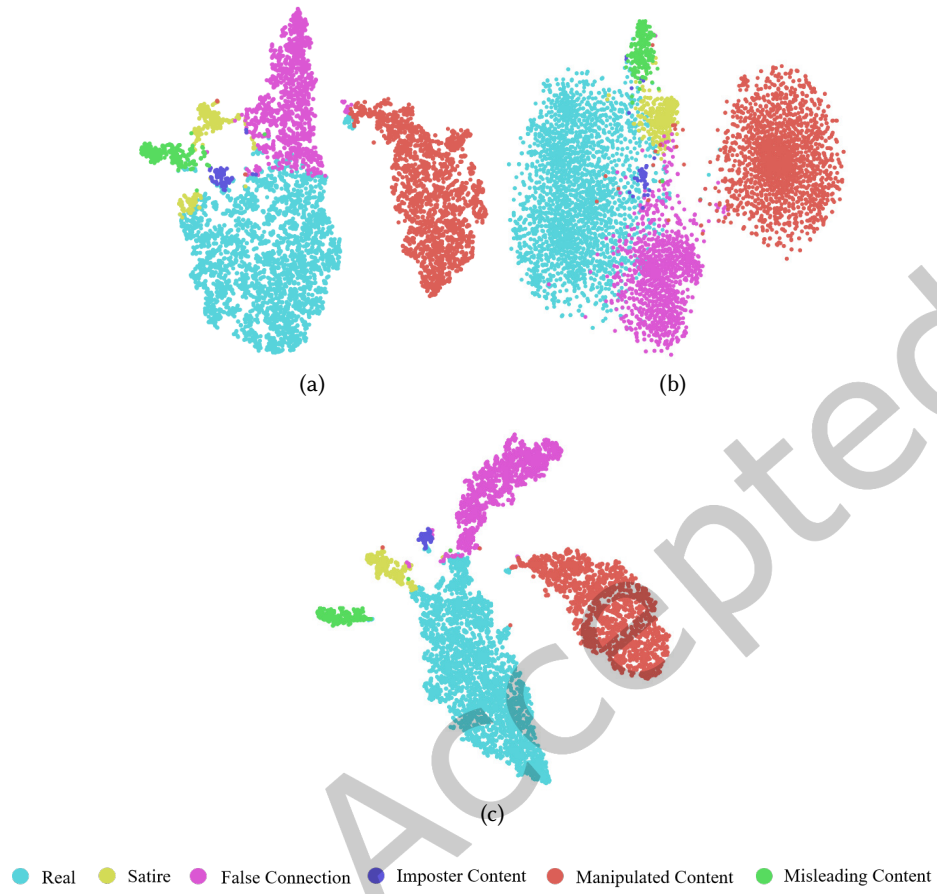


Fig. 5. T-SNE visualization depicting the 6-class detection in the Fakeddit dataset. (a) signifies the exclusion of graph contrastive learning, (b) indicates the removal of the guided semantic fusion module, isolating the single-modal information aspect, and utilizing only multimodal fusion features for prediction, while (c) denotes our complete model GS^{2}F.

grasshopper attacking a city—here, text features are dominant due to the exaggerated description, relegating other features to lesser roles. In sample (b), where image information has been manipulated, our model successfully identifies and concentrates on visual cues to assign increased contribution weights. Our findings indicate that textual misinformation is more obscure than visual tampering, resulting in more uncertain contributions from the model. Samples (c) and (d) illustrate scenarios where text and image modalities are not easily separable, emphasizing the necessity for integrating feature information to detect discrepancies between textual and visual content. The model dynamically assigns weights across different features to direct focus toward relevant feature information in each feature, thereby enhancing its classification performance.

5 Conclusion

In this paper, we introduce a multimodal fake news detection framework GS^{2}F, leveraging a graph-based structure and guided semantic fusion network. We construct a multimodal graph to facilitate multi-granularity interactions

Fig. 6. The weights from the dynamic contribution assignment layer for 4 samples.

and employ graph contrastive learning for capturing complex relationships. The framework enhances initial single-modal features via a semantic fusion module. Instead of direct concatenation for classification, we fuse multimodal and single-modal features through a novel feature-based contribution distribution layer, thereby improving the model's discriminative power. Extensive experiments on a representative dataset confirm our model's superior performance. In future work, we will explore more effective methods to improve the detection performance of fake news by incorporating a wider range of feature relationships between different modes. Additionally, we will consider integrating video and news transmission information into fake news detection.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62376062), the Ministry of Education of Humanities and Social Science Project (No. 23YJAZH220, No. 24YJAZH244), the Philosophy and Social Sciences 14th Five-Year Plan Project of Guangdong Province (No. GD23CTS03, No. GD21CTS02), and the Guangdong Basic and Applied Basic Research Foundation of China (No. 2023A1515012718).

References

[1] Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2019. Sentiment aware fake news detection on online social networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2507–2511.

[2] Herley Shaori Al-Ash and Wahyu Catur Wibowo. 2018. Fake news identification characteristics using named entity recognition and phrase detection. In *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*. IEEE, 12–17.

[3] Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications* 10, 1 (2019), 7.
[4] Michele Cantarella, Nicolò Fraccaroli, and Roberto Volpe. 2023. Does fake news affect voting behaviour? *Research Policy* 52, 1 (2023), 104628.
[5] Tong Chen, Xue Li, Hongzhi Yin, and Jun Zhang. 2018. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2018 Workshops, BDASC, BDM, ML4Cyber, PAISI, DaMEMO, Melbourne, VIC, Australia, June 3, 2018, Revised Selected Papers 22*. Springer, 40–52.
[6] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*. 2897–2905.
[7] Tsun-hin Cheung and Kin-man Lam. 2022. Crossmodal bipolar attention for multimodal classification on social media. *Neurocomputing* 514 (2022), 1–12.
[8] Deepjyoti Choudhury and Tapodhir Acharjee. 2023. A novel approach to fake news detection in social networks using genetic algorithm applying machine learning classifiers. *Multimedia Tools and Applications* 82, 6 (2023), 9029–9045.
[9] Limeng Cui, Suhang Wang, and Dongwon Lee. 2019. Same: sentiment-aware multi-modal embedding for detecting fake news. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*. 41–48.
[10] Mudit Dhawan, Shakshi Sharma, Aditya Kadam, Rajesh Sharma, and Ponnurangam Kumaraguru. 2024. Game-on: Graph attention network based multimodal fusion for fake news detection. *Social Network Analysis and Mining* 14, 1 (2024), 114.
[11] Xishuang Dong, Uboho Victor, and Lijun Qian. 2020. Two-path deep semisupervised learning for timely fake news detection. *IEEE Transactions on Computational Social Systems* 7, 6 (2020), 1386–1398.
[12] Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 1625–1628.
[13] Boyang Fu and Jie Sui. 2022. Multi-modal affine fusion network for social media rumor detection. *PeerJ Computer Science* 8 (2022), e928.
[14] Saqib Hakak, Mamoun Alazab, Suleman Khan, Thippa Reddy Gadekallu, Praveen Kumar Reddy Maddikunta, and Wazir Zada Khan. 2021. An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems* 117 (2021), 47–58.
[15] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 7837–7851.
[16] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.
[17] Syed Ali Khayam. 2003. The discrete cosine transform (DCT): theory and application. *Michigan State University* 114, 1 (2003), 31.
[18] Bo Li and Olan Scott. 2020. Fake news travels fast: Exploring misinformation circulated around Wu Lei's coronavirus case. *International Journal of Sport Communication* 13, 3 (2020), 505–513.
[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.
[20] Yujia Li, Richard Zemel, Marc Brockschmidt, and Daniel Tarlow. 2016. Gated Graph Sequence Neural Networks. In *Proceedings of ICLR'16*.
[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.
[22] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. 3818–3824.
[23] Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. 2020. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 international conference on multimedia retrieval*. 16–25.
[24] Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 6149–6157.
[25] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 231–240.
[26] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1212–1220.
[27] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1212–1220.

[28] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 153–162.
[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
[30] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*. 2931–2937.
[31] Julio CS Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Supervised learning for fake news detection. *IEEE Intelligent Systems* 34, 2 (2019), 76–81.
[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).
[33] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 658–666.
[34] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.
[35] Vivek K Singh, Isha Ghosh, and Darshan Sonagara. 2021. Detecting fake news stories via multimodal analysis. *Journal of the Association for Information Science and Technology* 72, 1 (2021), 3–17.
[36] Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. 2020. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 13915–13916.
[37] Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Leveraging intra and inter modality relationship for multimodal fake news detection. In *Companion Proceedings of the Web Conference 2022*. 726–734.
[38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.
[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[40] Bin Wang, Yong Feng, Xian-cai Xiong, Yong-heng Wang, and Bao-hua Qiang. 2023. Multi-modal transformer using two-level visual features for fake news detection. *Applied Intelligence* 53, 9 (2023), 10429–10443.
[41] Haizhou Wang, Sen Wang, and YuHu Han. 2022. Detecting fake news on Chinese social media based on hybrid feature fusion method. *Expert Systems with Applications* 208 (2022), 118111.
[42] Jinxia Wang, Stanislav Makowski, Alan Cieślik, Haibin Lv, and Zhihan Lv. 2023. Fake news in virtual community, virtual society, and metaverse: A survey. *IEEE Transactions on Computational Social Systems* (2023).
[43] Longzheng Wang, Chuang Zhang, Hongbo Xu, Yongxiu Xu, Xiaohan Xu, and Siqi Wang. 2023. Cross-modal Contrastive Learning for Multimodal Fake News Detection. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5696–5704.
[44] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*. 849–857.
[45] Junfei Wu, Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. 2023. Adversarial contrastive learning for evidence-aware fake news detection with graph neural networks. *IEEE Transactions on Knowledge and Data Engineering* (2023).
[46] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*. 2560–2569.
[47] Shufeng Xiong, Guipei Zhang, Vishwash Batra, Lei Xi, Lei Shi, and Liangliang Liu. 2023. TRIMOON: Two-Round Inconsistency-based Multi-modal fusion Network for fake news detection. *Information fusion* 93 (2023), 150–158.
[48] Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2019. Jointly embedding the local and global relations of heterogeneous graph for rumor detection. In *2019 IEEE international conference on data mining (ICDM)*. IEEE, 796–805.
[49] Wenjia Zhang, Lin Gui, and Yulan He. 2021. Supervised contrastive learning for multimodal unreliable news detection in covid-19 pandemic. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3637–3641.
[50] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*. 3465–3476.
[51] Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. 2022. MFAN: Multi-modal Feature-enhanced Attention Networks for Rumor Detection. IJCAI.
[52] Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multi-modal Fake News Detection on Social Media via Multi-grained Information Fusion. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*. 343–352.

[53] Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian. 2022. Multimodal sentiment analysis with image-text interaction network. *IEEE Transactions on Multimedia* (2022).

Received 17 March 2024; revised 29 October 2024; accepted 6 December 2024