A Multimodal Coupled Graph Attention Network for Joint Traffic Event Detection and Sentiment Classification

Yazhou Zhang, Prayag Tiwari, Qian Zheng, Abdulmotaleb El Saddik, *Fellow, IEEE*, and M. Shamim Hossain, *Senior Member, IEEE*

***Abstract*— Traffic events are one of the main causes of traffic accidents, leading to traffic event detection being a challenging research problem in traffic management and intelligent transportation systems (ITSs). The main gap in this task lies in how to extract and represent the valuable information from various kinds of traffic data. Considering the important role that social networks play in traffic data analysis, we argue that sentiment classification and traffic event detection are two closely related tasks in ITSs, where event and sentiment can reveal both explicit and implicit traffic accidents, respectively. Unfortunately, none of the recent approaches in traffic event detection have taken sentiment knowledge into view. This paper proposes a multimodal coupled graph attention network (MCGAT). It aims to construct a multimodal multitask interactive graphical structure where terms (sucha as words, and pixels) are treated as nodes, and their contextual and cross-modal correlations are formalized as edges. The key components are cross-modal and cross-task graph connection layers. The cross-modal graph connection layer captures the multimodal representation, where each node in one modality connects all nodes in another modality. The cross-task graph connection layer is designed by connecting the multimodal node in one task to two single nodes in another task. Empirical evaluation of two benchmarking datasets, such as MGTES and Twitter, shows the effectiveness of the proposed model over state-of-the-art baselines in terms of F1 and accuracy, with significant improvements of 2.4%, 2.4%, 2.7%, and 2.7%.**

***Index Terms*— Traffic event detection, sentiment classification, graph neural networks, graph embedding, intelligent transportation system.**

I. Introduction

An increasing number of car users actively take part in urban road networks, and thus traffic congestion inevitably emerges and is becoming a key problem in traffic management and intelligent transportation systems (ITSs). There are two kinds of reasons for traffic congestion, e.g., external and internal reasons, where the external reason includes accidental events, facility failure, weather conditions, etc., while internal reasons mainly derive from user's own state, e.g., his/her feelings and mindset. For example, the battle between two aggressive drivers often results in traffic disruption, as shown in Fig. 1. In contrast, an orderly and convenient traffic environment makes each participant comfortable. Hence, the analysis of both cases will benefit traffic congestion prediction and ITS. This brings forth two tasks: the first task is traffic event detection, and the second is sentiment classification [1].

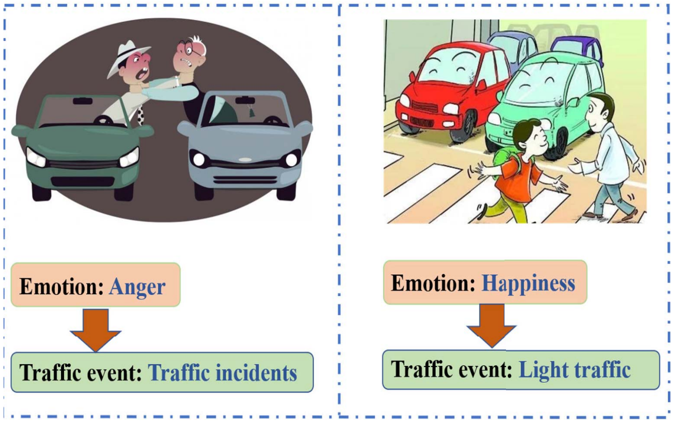
Emotion: Anger
Traffic event: Traffic incidents
Emotion: Happiness
Traffic event: Light traffic

Fig. 1. An example of how user sentiment affects traffic event detection.

Traffic events mainly consist of incidents, weather conditions, traffic congestion, and light traffic. Traffic detection refers to discovering, identifying and classifying various traffic factors expressed in traffic records using machine/deep learning, data mining. In view of its potential for traffic management and smart transportation, traffic event detection has attracted increased attention. As a result, a growing body of literature has been published to explore its key challenge, i.e., transportation data mining and representation. For instance, Angelica *et al.* [2] used a support vector machine (SVM) and word embedding to process traffic reviews collected from Twitter. Alomari and Mehmood [3] proposed a hyper classifier that combined naive Bayes (NB), SVM, and logistic regression (LR) to discover multiple types of road events, e.g., accidents, road work, and road closures.

There are two research gaps in the existing studies. The first is that most of the recent research has focused on traffic event

Manuscript received 8 December 2021; revised 12 May 2022, 24 June 2022, and 15 August 2022; accepted 23 August 2022. Date of publication 25 October 2022; date of current version 2 August 2023. This work was supported by the Deanship of Scientific Research at King Saud University, Riyadh, Saudi Arabia, through the Vice Deanship of Scientific Research Chairs: Research Chair of Pervasive and Mobile Computing. The Associate Editor for this article was W. Wei. *(Yazhou Zhang and Prayag Tiwari contributed equally to this work.) (Corresponding authors: M. Shamim Hossain; Qian Zheng.)*

Yazhou Zhang and Qian Zheng are with the Software Engineering College, Zhengzhou University of Light Industry, Zhengzhou 450002, China (e-mail: yzzhang@zzuli.edu.cn; zhengqian@zzuli.edu.cn).

Prayag Tiwari is with the School of Information Technology, Halmstad University, 30118 Halmstad, Sweden (e-mail: prayag.tiwari@hh.se).

Abdulmotaleb El Saddik is with the Multimedia Communications Research Laboratory (MCRLab), University of Ottawa, Ottawa, ON K1N 6N5, Canada, and also with the Computer Vision Department, Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates (e-mail: elsaddik@uottawa.ca).

M. Shamim Hossain is with the Research Chair of Pervasive and Mobile Computing, King Saud University, Riyadh 11543, Saudi Arabia, and also with the Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia (e-mail: mshossain@ksu.edu.sa).

Digital Object Identifier 10.1109/TITS.2022.3205477

1558-0016 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 08:18:37 UTC from IEEE Xplore. Restrictions apply.

detection, while the shared knowledge from internal reasons has been neglected [4], [5]. Many studies find that drivers who are in a negative emotional state have a much higher rate of traffic accidents than those who are in a normal mood [6]. The second is the lack of structured data representation, viz. graph embedding and learning. Each transportation network can be treated as a weighted and directed graph, in which nodes represent intersections of the road network and edges denote traffic routes. The current approaches have failed to take the correlation between textual and visual nodes to learn multimodal node representations.

Motivation. Hence, we summarize two key research questions:

RQ1: Does sentiment mining help traffic event detection?

RQ2: How can an efficient and effective multimodal graph learning model be developed?

We take two actions to answer the abovementioned two questions. To answer RQ1, we argue that traffic event detection and sentiment classification are two correlative tasks, and detecting sentiment will benefit traffic event detection. We aim to leverage the shared sentiment knowledge to enhance the traffic data representation. To answer RQ2, we construct two graphs based on the textual and visual traffic documents, where each term (e.g., word, region) is treated as a node, and the semantic dependency between terms is seen as an edge. Meanwhile, we aim to model multimodal node fusion and multitask correlation via node connection across different graphs.

Motivated by this, we propose a multimodal coupled graph attention network (MCGAT) for joint traffic event detection and sentiment classification, which leverages both multimodal representation and sentiment knowledge extraction. In particular, each textual document and its visual counterpart in traffic event detection and sentiment classification tasks are seen as graphs, in which terms (i.e., words, pixels) are treated as nodes, and their contextual and cross-modal correlations are formalized as edges. The key components are cross-modal and cross-task graph connection layers. The cross-modal graph connection layer captures the multimodal representation by building text and image subgraphs, where each node in one modality connects all nodes in another modality. The cross-task graph connection layer is designed by connecting the multimodal node in one task to two single nodes in another task. Four subgraphs are integrated into a unified graph architecture and updated jointly. After L iterations, we obtain the document and image representation and merge them together to gain a multimodal representation. WE forwards it through the softmax functions for traffic event detection and sentiment prediction. We clarify that the objectives of the proposed model is to answer two research questions R1 and R2, because of the design of multi-task joint learning and the introduction of graph structure learning.

In line with previous traffic event detection approaches [7], we create two social media-based traffic event datasets, i.e., MGTES and Twitter. Then, we designed and conducted comprehensive experiments to prove the effectiveness of the proposed MCGAT model by comparing it with a series of strong baselines, including two typical and representative machine learning approaches (i.e., SVM and random forest, RF), five single modal deep learning approaches, i.e., convolutional neural network (CNN), bidirectional gated recurrent unit (BiGRU), multihead attention-based gated recurrent unit (MAGRU) network, bidirectional encoder representations from transformers (BERT), graph attention networks (GAT), and three multimodal graph attention network (GAT) approaches, i.e., early-fusion, late-fusion and Dempster/Shafer (D-S) evidence fusion models. Our MCGAT achieves the F1 and accuracy classification scores of 2.4%, 2.4%, 2.7% and 2.7%.

To the best of our knowledge, our work is the first to combine traffic event detection and sentiment classification in a multimodal multitask scenario. Different from previous studies that only focused on traffic event detection while the shared knowledge from related tasks has been neglected, this study aims to incorporate sentimental knowledge into traffic event detection by proposing a multitask graph learning framework. In addition, previous studies have failed to take the correlation between textual and visual nodes to learn multimodal node representations. There is a great difference in that we create a multimodal multitask graph in which a cross-modal connection is designed to learn multimodal fusion.

The major innovations of the work can be summarized as follows.

- A multitask multimodal graph learning framework for traffic event and sentiment joint detection is proposed to simultaneously incorporate multimodal and multitask information into a joint learning model.
- We regard textual and visual semantic units as nodes, model their semantic dependencies as edges, and build a coupled multi-modal graph.
- We create a cross-modal and a cross-task connection layer by connecting each node in one modality or task to all nodes in the other modality or task.
- The effectiveness of the proposed model is checked by using it to detect traffic event and sentiment. Empirical experimental results on two datasets show that the proposed model substantially outperforms strong baselines.

The structure of this paper is as follows. We describe the related work in Section II. Section III shows the multimodal coupled GAT model. Next, experiments are performed in Section IV. We presents our conclusions in Section V.

II. Related Work

In this section, we describe the related studies of traffic event detection and sentiment analysis.

A. Traffic Event Detection

Traffic event detection is now considered a supervised multilabel classification task in ITSs. As the scale and distribution of traffic data have grown explosively, machine learning and deep learning-based approaches have become the mainstream paradigms.

1) Machine learning based approaches: Shallow machine learning approaches aim to extract traffic features and discover

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 08:18:37 UTC from IEEE Xplore. Restrictions apply.

traffic events via various kinds of classifiers and statistical rules. Their performance is often over borderline, since such methods benefit from the strong math foundation and machine learning technology foundation. For instance, Al Dhanhani *et al.* [8] first explored the potential of the shapelet transform in traffic event detection and showed comparable performance against other machine learning algorithms. Alomari and Mehmood [3] proposed a hyper classifier that combined NB, SVM, and LR to discover multiple types of road events, e.g., accidents, road work, and road closures. From the perspective of feature learning, Xu *et al.* [9] filtered noisy information from traffic events by adjusting the association rules among words and trained more effective traffic features [10]. Imamverdiyev and Sukhostat [11] chose to use an extreme learning machine (ELM) to detect incidents in network traffic [12]. Salas *et al.* [2] used SVM and word embedding to process traffic reviews collected from Twitter. Zulfikar *et al.* [13] treated street properties, e.g., the street name and address in Google Maps, as features and used an SVM to judge traffic congestion. Hodo *et al.* [14] compared the performance difference between SVM and shallow neuron networks and proved the priority of neural networks. Sodhro AH's team [15] discussed the current trends and future challenges in machine learning based communication networks, and proposed a range of strong approaches.

2) Deep learning-based approaches: In consideration of its outstanding representation and discernment, deep neural networks have been applied to traffic event detection. They achieve better experimental results than shallow machine learning approaches. Wan *et al.* [16] used a superframe segmentation approach based on feature fusion for intelligent traffic data analysis. Dabiri and Heaslip [17] collected more than 50,000 traffic tweets, used word embedding to represent each term, and performed traffic event classification using a recurrent neural network (RNN) and convolutional neural network. Zhang *et al.* [18] presented a hyper deep learning framework that combined a deep belief network (DBN) and long short-term memory (LSTM), obtaining an overall accuracy of 85%. Aboah [19] focused on visual traffic event classification and proposed a decision tree powered by deep learning. Zhang *et al.* [20] studied traffic data classification from acoustic information and compared seven typical deep learning approaches. They concluded that a spectral feature-based CNN achieved the best performance. Chen and Wang [21] incorporated sensor data and social tweets into their framework, and designed a multi-modal neural network (MMM) for traffic event detection.

In summary, the two aforementioned types of studies have made good progress in ITSs. However, different from them, we take the first step to incorporate sentimental knowledge into traffic event detection and use graph neural networks to learn node representation.

B. Multimodal Sentiment Analysis

Multimodal sentiment analysis aims to identify the polarity expressed in multimodal documents or data. For instance, Zhang *et al.* [5], [22], [23] introduced a quantum theory-like multimodal sentiment classification model. Liu *et al.* [24] designed a tensor product-based multimodal representation model for conversational sentiment analysis. Most recent studies are performed from a deep learning perspective, since deep learning paradigms can extract abstract and deep features [25], [26], [27]. Jiang *et al.* [28] discussed the importance of multimodal fusion approaches for emotion analysis. Zadeh *et al.* [29] proposed using tensor products to perform multimodal visual and vocal feature fusion. Huang *et al.* [30] introduced an attention-based multimodal fusion approach for learning effective representations. Hossain and Muhammad [31] used deep learning approach for multimodal emotion classification task. Yadav *et al.* [32] proposed a sentiment analysis framework based on the weighted word representation technique and trained five widely used classifiers. Inspired by a hierarchical attention network, they proposed a deep language-independent multilevel attention-based Conv-BiGRU network (MACBiG-Net) to model past and future context information for sentiment classification [33]. To effectively extract visual features, they presented an xception residual attention-based network (XRA-Net) [34]. Their team presented the affect-based movie trailer classification model and a combination of CNN and BiLSTM for emotion classification [35], [36], [37]. To explore the potential of RNNs in multimodal sentiment analysis, Agarwal and Yadav [38] designed and compared four typical variants of RNNs, GRNN, LRNN, GLRNN and UGRNN, on the CMU-MOSI dataset. Yadav and Vishwakarma [39] designed a residual attention-based deep learning network (RA-DLNet) to alleviate the problem of visual sentiment analysis. They evaluated their model on eight benchmarking datasets.

Moreover, the multimodal learning strategy has motivated a large number of works on multimodal sentiment classification,which are often focused on multimodal representation learning [40], [41], [42], [43], [44]. Huddar *et al.* [45] decided to employ RNN to model the speaker state and context information between sentences and conducted experiments on interactive emotional dyadic motion capture (IEMOCAP) dataset. Li *et al.* [46] presented a cross-modal graph attention model for multimodal emotion analysis [47]. Inspired by a multitask learning paradigm, Yu *et al.* [48] proposed the training multimodal and unimodal sentiment classification tasks to learn multimodal and unimodal representations. Zhang *et al.* [49], [50] opened a new road for conversational sentiment analysis, quantum sentiment analysis. They and their team aim to develop a novel multimodal conversational model based on quantum probability. They proposed a quantum-inspired interactive network (QIN) model textual sentiment analysis. In order to create a conversational graph-based CNN for sentiment analysis, Zhang *et al.* [51] considered each speaker and utterance as a node in each discussion.

Our work is the first that brings together sentiment analysis and traffic data detection and develops a multitask learning problem.

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 08:18:37 UTC from IEEE Xplore. Restrictions apply.

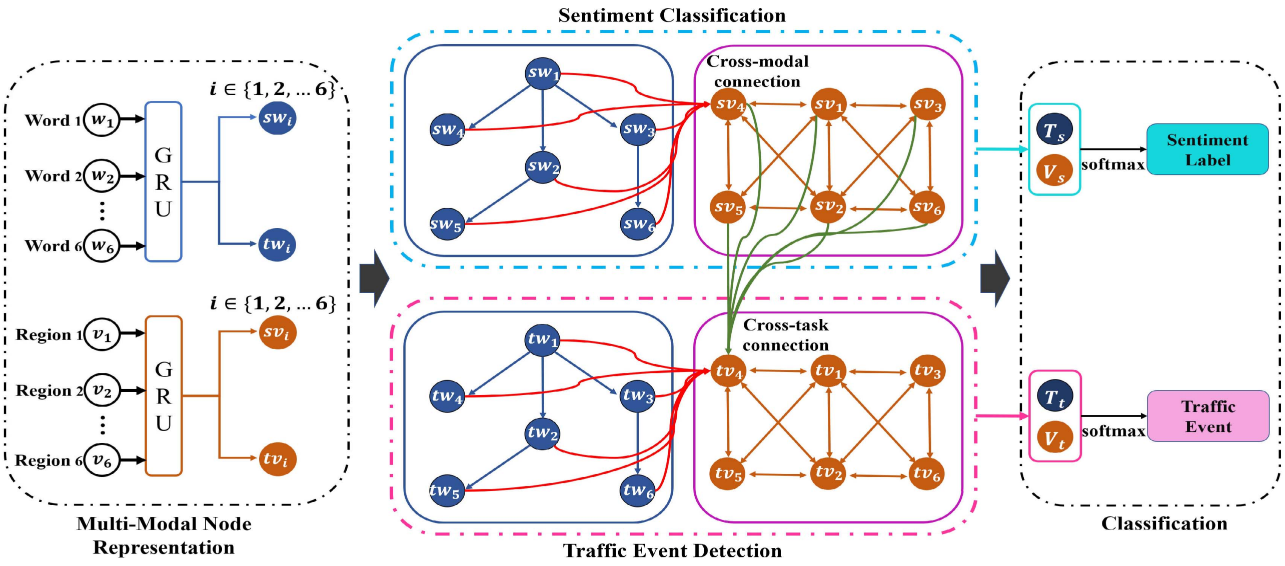

Fig. 2. The architecture of the multimodal coupled graph attention network.

III. The Proposed Methodology

To capture extra sentiment knowledge and learn graph embedding for multimodal traffic records, we propose a multimodal coupled graph attention network.

A. Problem Formulation and Overall Network

1) Problem Formulation: Assume that the dataset contains N multimodal traffic samples, where the i^{th} one is represented as $D^i = \left\{\left(T^i, V^i\right), Y^i\right\}$. T^i and V^i denote the corresponding textual and visual records, respectively, and $i \in [1, 2, \ldots, N]$. In this study, we regard traffic event detection as the main task, denoted as t, while sentiment classification is seen as the ancillary task, denoted as s. For example, T_t^i and T_s^i represent the i^{th} textual documents for two tasks. Then, $T_t^i \in \mathcal{H}^{l_{T_t} \times d_{T_t}}$, $V_t^i \in \mathcal{H}^{l_{V_t} \times d_{V_t}}$, $T_s^i \in \mathcal{H}^{l_{T_s} \times d_{T_s}}$, and $V_s^i \in \mathcal{H}^{l_{V_s} \times d_{V_s}}$. Here, l_{T_t}, l_{V_t}, l_{T_s} and l_{V_s} denote the sequence length of textual and visual utterances for two tasks, and d_t and d_v mean the dimensions of the textual and visual features for two tasks.

Now, the multitask learning problem can be formulated as:

$$\zeta = \sum_{\alpha} \prod_{i} p\left(Y_\alpha^i | T_\alpha^i, V_\alpha^i, \Theta_\alpha\right) \tag{1}$$

where Θ represents the parameter set, p denotes the prediction probability, t denotes the task of traffic event detection, s represents the task of sentiment analysis, and $\alpha \in \{t, s\}$. 1 shows that our goal is to determine the joint probability of the traffic event or sentiment label based on the a priori knowledge (e.g., the textual and visual documents) and parameter tuning for all samples in the dataset.

2) Overall Network: We construct four subgraphs based on textual documents and visual images for two tasks, i.e., a text subgraph for traffic event detection, a text subgraph for sentiment analysis, a visual subgraph for traffic event detection, and a visual subgraph for sentiment analysis. Then, these four subgraphs are incorporated into a unified graph architecture for modeling their correlations. The overall architecture of the MCGAT framework is shown in Fig. 2.

More specifically, MCGAT contains six core modules, which are a multimodal node representation layer, a cross-modal connection layer, a graph attention layer, a cross-task connection layer, a graph representation layer and a classification layer. (1) In the multimodal node representation layer, each textual word and visual region are represented as textual and visual embeddings, denoted as $\vec{e}_{tw_j}$, $\vec{e}_{tv_j}$, $\vec{e}_{sw_j}$ and $\vec{e}_{sv_j}$. (2) In the cross-modal connection layer, each node in one modality subgraph connects all nodes in another modality graph to learn cross-modal knowledge. For example, each word in the text subgraph is connected with all visual regions in the visual subgraph and vice versa. (3) In the graph attention layer, each textual or visual node representation will be updated via a multihead attention mechanism. (4) The cross-task connection layer is designed to model the inter-relatedness across traffic event detection and sentiment analysis. (5) In the graph representation layer, we obtain the document and image representation, i.e., T_t, V_t, T_s and V_s. (6) The final multimodal representations M_t and M_s for two tasks are calculated by merging textual and visual representations together and thus are fed into the softmax functions for traffic event detection and sentiment prediction. We will present each layer in the following sections.

B. Graph Construction

We detail the construction procedure of the proposed MCGAT network here. A textual traffic record, which contains L words, is represented as a directed graph, while its visual counterpart having M regions is represented as a strongly connected bidirectional graph, which can be written as:

$$\mathcal{G}_k = (\mathcal{V}, \mathcal{E}, \mathcal{W})_k \tag{2}$$

where $k \in \{tt, tv, st, sv\}$ represents text or image for traffic event detection and sentiment classification. For example, tt denotes the text graph for traffic event detection, and tv denotes the visual graph for traffic event detection. The vertices $v_j \in \mathcal{V}$ in $\mathcal{G}_k$.

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 08:18:37 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1 Training the MCGAT Model

```
Require: Graph sequence $\mathcal{G} = \{\mathcal{G}_{tt}, \mathcal{G}_{tv}, \mathcal{G}_{st}, \mathcal{G}_{sv}\}$ and node initial vectors $U_{tw}^{G^{(0)}}$, $U_{sw}^{G^{(0)}}$, $U_{tv}^{G^{(0)}}$, $U_{sv}^{G^{(0)}}$ based on Eq. 10 and 11.
Ensure: Multi-modal graph embeddings, $M_t$ and $M_s$.
for $i \in [1, N]$ do
    for $k \in [1, 4]$ do
        for Text: $j \in [1, L]$, Image: $j \in [1, M]$ do
            $U_{tw}^{G^{(\gamma)}} \leftarrow GAT\left(U_{tw}^{G^{(\gamma-1)}}, A_{tt}, \Theta_{tw}^{G^{(\gamma-1)}}\right)$
            $U_{sw}^{G^{(\gamma)}} \leftarrow GAT\left(U_{sw}^{G^{(\gamma-1)}}, A_{st}, \Theta_{sw}^{G^{(\gamma-1)}}\right)$
            $U_{tv}^{G^{(\gamma)}} \leftarrow GAT\left(U_{tv}^{G^{(\gamma-1)}}, A_{tv}, \Theta_{tv}^{G^{(\gamma-1)}}\right)$
            $U_{sv}^{G^{(\gamma)}} \leftarrow GAT\left(U_{sv}^{G^{(\gamma-1)}}, A_{sv}, \Theta_{sv}^{G^{(\gamma-1)}}\right)$
        end for
    end for
    $T_t \leftarrow Mean\left(U_{tw}^{G^{(\gamma)}}\right)$; $V_t \leftarrow Mean\left(U_{tv}^{G^{(\gamma)}}\right)$
    $T_s \leftarrow Mean\left(U_{sw}^{G^{(\gamma)}}\right)$; $V_s \leftarrow Mean\left(U_{sv}^{G^{(\gamma)}}\right)$
    $M_t \leftarrow [T_t; V_t]$; $M_s \leftarrow [T_s; V_s]$
end for
Minimize the cross-entropy loss $\zeta$ in Eq. 19.
```

1) Vertices: Each word in the text or each region in the image for two tasks is seen as a vertex, satisfying that $tw_j \in \mathcal{V}_{tt}$, $tv_j \in \mathcal{V}_{tv}$, $sw_j \in \mathcal{V}_{st}$ and $sv_j \in \mathcal{V}_{sv}$. Each node is represented as the feature vector using the multimodal node representation encoder. We treat this vector as the vertex feature. The first layer states vector for all nodes in four sub-graphs are obtained.

2) Edges: The construction process of edges $\mathcal{E}$ can be divided into two categories. (1) In text graphs for traffic event detection and sentiment classification, we exploit syntactical dependency relationships within a sentence to construct edges between vertices. Specifically, we build the dependency tree of the given sentence [52] to semantically connect nodes. (2) In image graphs, we argue that each visual region (vertex) is contextually dependent on all the other regions in an image. We thus build two fully connected bidirectional graphs for two tasks, where each node is connected to all the other nodes (including itself) with an edge. This action will lead to an edge scale of $O\left(M^2\right)$, which introduces expensive calculations for model training. To alleviate this problem, we construct the edges by only connecting each vertex to its adjacent neighbors, e.g., the distance between vertices is 1.

3) Edge Weights: The graph attention network is used as a fundamental part of our system to calculate the weights between a node and its neighbors. In many graph-structured data processing applications, GAT's ability to implicitly assign various weights to various nodes in a neighborhood has produced state-of-the-art outcomes. [53]. For the text for traffic event detection, given a L-node dependency graph $G_{tt}^{(0)} = \left\{\vec{e}_{tw_1}^{G_{tt}^{(0)}}, \vec{e}_{tw_2}^{G_{tt}^{(0)}}, \ldots, \vec{e}_{tw_j}^{G_{tt}^{(0)}}, \ldots, \vec{e}_{tw_L}^{G_{tt}^{(0)}}\right\}$, each node is represented by a word embedding vector $\vec{e}_{tw_j}^{G_{tt}^{(0)}} = (\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_d)$.

Then, GAT layer-wisely updates the representation of each node with learnable linear transformation and multihead attention and produces a new set of node representations $G^{(E)} = \left\{\vec{e}_{tw_1}^{G^{(E)}}, \vec{e}_{tw_2}^{G^{(E)}}, \ldots, \vec{e}_{tw_L}^{G^{(E)}}\right\}$, as the output by an E-layer GAT. The update process can be written as:

$$\vec{e}_{tw_j}^{G^{(\gamma)}} = \Big\Vert_{h=1}^{H} \sigma\left(\sum_{p \in \mathcal{N}_j} \alpha_{jp}^{h} \mathbf{W}^h \vec{e}_{tw_p}^{G^{(\gamma-1)}}\right) \tag{3}$$

$$\alpha_{jp}^{h} = \frac{exp\left(\mathcal{F}\left(\vec{\mathbf{a}}^T\left[\mathbf{W}\vec{e}_{tw_j}^{G^{(\gamma-1)}} || \mathbf{W}\vec{e}_{tw_p}^{G^{(\gamma-1)}}\right]\right)\right)}{\sum_{q \in \mathcal{N}_j} exp\left(\mathcal{F}\left(\vec{\mathbf{a}}^T\left[\mathbf{W}\vec{e}_{tw_j}^{G^{(\gamma-1)}} || \mathbf{W}\vec{e}_{tw_q}^{G^{(\gamma-1)}}\right]\right)\right)} \tag{4}$$

$$\mathcal{F}(a, b) = LeakyReLU\left(\vec{\mathbf{a}}^T\left[\mathbf{W_a}a || \mathbf{W_b}b\right]\right) \tag{5}$$

where $\gamma \in [1, 2, 3 \ldots, E]$. H is the number of heads, which will be set to eight. $||$ is the concatenation operation, σ denotes a nonlinear activation function, and α_{jp}^{h} is the attention score of node j to its neighbor node p in the h-th attention head in the same graph. $\mathbf{W}^h$ is the linear transformation weight. $\mathcal{N}_j$ denotes the set of the j node's neighbors. $\mathcal{F}(\cdot)$ is a LeakyReLU function that has a small negative slope of 0.2. $\vec{\mathbf{a}}^T$ is an attention context vector learned during training. $\mathbf{W}$ is a set of trainable parameters, where the goal is to transform features from low-order to high-order. It plays a similar role to a fully connected layer. Eq. 5 is proposed to merge node a and node b together and thus map it into a scalar by performing a nonlinear LeakyReLU transformation. This action is used to learn the attention score $\mathcal{F}(a, b)$ of the node pair (a, b). Based on the attention score $\mathcal{F}(a, b)$, we proposed Eq. 4 to normalize the attention scores of all neighbors of node j. Due to the varied structure of graphs, nodes can have a different number of neighbors. To have a common scaling across all neighborhoods, the attention scores need to be normalized. This makes coefficients easily comparable across different nodes. After normalization, α_{jp}^{h} is treated as the final attention weight. 3 shows that the embeddings from neighbors are aggregated together, scaled by the attention weights, to update the node representation through a nonlinear activation function.

For the sake of simplicity, we could also formulate this feature generation process as follows:

$$U_{tw}^{G^{(\gamma)}} = GAT\left(U_{tw}^{G^{(\gamma-1)}}, A_{tt}, \Theta_{tw}^{G^{(\gamma-1)}}\right) \tag{6}$$

where $U_{tw}^{G^{(l-1)}}$ is the state for all nodes at layer γ, $A_{tt} \in R^{L \times L}$ is the corresponding graph adjacency matrix, and $\Theta_{tw}^{G^{(\gamma-1)}}$ is the parameter set at layer $\gamma - 1$.

In the same manner, the feature generation process of the text node for sentiment classification and that of the visual nodes can also be depicted as:

$$U_{sw}^{G^{(\gamma)}} = GAT\left(U_{sw}^{G^{(\gamma-1)}}, A_{st}, \Theta_{sw}^{G^{(\gamma-1)}}\right) \tag{7}$$

$$U_{tv}^{G^{(\gamma)}} = GAT\left(U_{tv}^{G^{(\gamma-1)}}, A_{tv}, \Theta_{tv}^{G^{(\gamma-1)}}\right) \tag{8}$$

$$U_{sv}^{G^{(\gamma)}} = GAT\left(U_{sv}^{G^{(\gamma-1)}}, A_{sv}, \Theta_{sv}^{G^{(\gamma-1)}}\right) \tag{9}$$

where A_{st}, A_{tv} and A_{sv} represent three adjacency matrices.

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 08:18:37 UTC from IEEE Xplore. Restrictions apply.

C. Multimodal Node Representation

1) Text: For text in the task of traffic event detection, there are L words in the i^{th} target sentence, i.e., $T_t^i = \{tw_1, tw_2, \ldots, tw_L\}$. Each word $tw \in \mathcal{R}^{d_t}$ is represented by pretrained BERT vectors [54], [55]. The sentence is then passed through a gated recurrent unit (GRU) to learn sentence representation $F_{tt} = [\vec{e}^{G^{(0)}}_{tw_1}, \vec{e}^{G^{(0)}}_{tw_2}, \ldots, \vec{e}^{G^{(0)}}_{tw_L}]$. The initialization process for textual records can be formulated as:

$$\begin{aligned} F_{tt} &= GRU\left(BERT(T_t^i)\right) \\ F_{st} &= GRU\left(BERT(T_s^i)\right) \end{aligned} \tag{10}$$

where $F_{st} = [\vec{e}^{G^{(0)}}_{sw_1}, \vec{e}^{G^{(0)}}_{sw_2}, \ldots, \vec{e}^{G^{(0)}}_{sw_L}]$.

2) Image: For the image, we also assume that there are M visual regions in the i^{th} target image, i.e., $V_t^i = \{tv_1, tv_2, \ldots, tv_M\}$. Each input image is scaled to 480 × 360, in which each visual region is a 24 × 18 block. We extract the visual features via the pretrained residual neural network (ResNet) network [56]. We thus feed each visual region into the GRU unit to obtain its representation $F_{tv} = [\vec{e}^{G^{(0)}}_{tv_1}, \vec{e}^{G^{(0)}}_{tv_2}, \ldots, \vec{e}^{G^{(0)}}_{tv_M}]$. The initialization process for visual documents can be formulated as:

$$\begin{aligned} F_{tv} &= GRU\left(ResNet(V_t^i)\right) \\ F_{sv} &= GRU\left(ResNet(V_s^i)\right) \end{aligned} \tag{11}$$

where $F_{sv} = [\vec{e}^{G^{(0)}}_{sv_1}, \vec{e}^{G^{(0)}}_{sv_2}, \ldots, \vec{e}^{G^{(0)}}_{sv_M}]$.

D. Cross-Modal and Cross-Task Connection Layers

The core modules of the proposed MCGAT are a cross-modal connection layer and a cross-task connection layer, which are used to model the multimodal interaction and multitask correlation into a unified graph architecture. Hence, we have four subgraphs in the proposed framework. Suppose that there are L nodes in the text graphs and M nodes in the image graphs. We have $2L + 2M$ nodes (vertices) in the graphical whole architecture.

Moreover, there are two kinds of edges. (1) **Cross-modal connection.** Each node in one modality connects all nodes in another modality graph to learn cross-modal knowledge. For example, the j^{th} visual region in the image graph is connected to all words $[tw_1, tw_2, tw_3 \ldots, tw_L]$ in the text subgraph they belong to the same task, and vice versa. The visual node representation update process of the γ^{th} layer can be formulated as:

$$\vec{e}^{G^{(\gamma)}}_{tv_j} = \overset{H}{\underset{h=1}{\Vert}} \sigma\left(\sum_{p\in\mathcal{N}_j} \alpha^h_{jp}\mathbf{W}^h\vec{e}^{G^{(\gamma-1)}}_{tv_p} + \sum_{u\in L}\alpha^h_{ju}\mathbf{W}^h\vec{e}^{G^{(\gamma-1)}}_{tw_u}\right) \tag{12}$$

$\sum_{p\in\mathcal{N}_j} \alpha^h_{jp}\mathbf{W}^h\vec{e}^{G^{(\gamma-1)}}_{tv_p}$ represents the intramodal connection, and $\sum_{u\in L}\alpha^h_{ju}\mathbf{W}^h\vec{e}^{G^{(\gamma-1)}}_{tw_u}$ denotes the cross-modal connection.

In contrast, the textual node representation update process of the γ^{th} layer can also be formulated as:

$$\vec{e}^{G^{(\gamma)}}_{tw_j} = \overset{H}{\underset{h=1}{\Vert}} \sigma\left(\sum_{p\in\mathcal{N}_j} \alpha^h_{jp}\mathbf{W}^h\vec{e}^{G^{(\gamma-1)}}_{tw_p} + \sum_{u\in M}\alpha^h_{ju}\mathbf{W}^h\vec{e}^{G^{(\gamma-1)}}_{tv_u}\right) \tag{13}$$

(2) **Cross-task connection.** Two correlative tasks are constructed as two interactive subgraphs, where each node in one task connects all nodes in another task graph for the same modality. This action could explicitly leverage the mutual interaction knowledge from other tasks. For example, the j^{th} region, e.g., tv_j, for the task of traffic event detection is connected to all visual regions $[sv_1, sv_2, sv_3 \ldots, sv_M]$ in another task graph they belong to the same modality, and vice versa. The visual node representation update process of the γ^{th} layer can be formulated as:

$$\vec{e}^{G^{(\gamma)}}_{tv_j} = \overset{H}{\underset{h=1}{\Vert}} \sigma\left(\sum_{p\in\mathcal{N}_j} \alpha^h_{jp}\mathbf{W}^h\vec{e}^{G^{(\gamma-1)}}_{tv_p} + \sum_{u\in M}\alpha^h_{ju}\mathbf{W}^h\vec{e}^{G^{(\gamma-1)}}_{sv_u}\right) \tag{14}$$

where $\sum_{p\in\mathcal{N}_j} \alpha^h_{jp}\mathbf{W}^h\vec{e}^{G^{(\gamma-1)}}_{tv_p}$ represents the intramodal connection, and $\sum_{u\in M}\alpha^h_{ju}\mathbf{W}^h\vec{e}^{G^{(\gamma-1)}}_{sv_u}$ denotes the cross-task connection.

Similarly, the visual node representation update process for the task of sentiment classification can be formulated as:

$$\vec{e}^{G^{(\gamma)}}_{sv_j} = \overset{H}{\underset{h=1}{\Vert}} \sigma\left(\sum_{p\in\mathcal{N}_j} \alpha^h_{jp}\mathbf{W}^h\vec{e}^{G^{(\gamma-1)}}_{sv_p} + \sum_{u\in M}\alpha^h_{ju}\mathbf{W}^h\vec{e}^{G^{(\gamma-1)}}_{tv_u}\right) \tag{15}$$

As already mentioned, there are L words in the textual document and M visual blocks in the image, then we have L textual vertices and M visual vertices for both traffic event detection and sentiment classification, in total $2L \times 2M$ nodes. In the directed graphs, each edge has a direction from v_i to v_j and is written as an ordered pair $\langle v_i, v_j\rangle$. Based on this, we build an adjacency matrix, denoted as $A \in R^{(2L\times 2M)}$, where $A_{i,j} = 1$ if and only if the ordered pair $\langle v_i, v_j\rangle$ is in the set of edges. For example, $A_{i,j} = 1$ if vertices v_i and v_j belong to the different tasks for the same modality, and $A_{i,j} = 1$ if vertices v_i and v_j belong to the different modalities for the same task. In addition, $A_{i,j} = 1$ if vertices v_i and v_j are semantically connected (e.g., syntax dependency in the dependency tree) in the same sub-graph. Otherwise, $A_{i,j} = 0$.

E. Multimodal Graph Embedding

After an E layer updating of MCGAT, all nodes in four sub-graphs have their final embeddings, e.g., $\vec{e}^{G^{(E)}}_{tw_{j\in[1,2,\ldots,L]}}$, $\vec{e}^{G^{(E)}}_{tv_{j\in[1,2,\ldots,M]}}$, $\vec{e}^{G^{(E)}}_{sw_{j\in[1,2,\ldots,L]}}$ and $\vec{e}^{G^{(E)}}_{sv_{j\in[1,2,\ldots,M]}}$. The sentence and image representations for two tasks are obtained by averaging all textual and visual nodes' representations, i.e., T_t, V_t, T_s and V_s, which can be written as:

$$T_t = Avg\left(\sum_{j=1}^{L}\vec{e}^{G^{(E)}}_{tw_j}\right), \quad V_t = Avg\left(\sum_{j=1}^{M}\vec{e}^{G^{(E)}}_{tv_j}\right)$$

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 08:18:37 UTC from IEEE Xplore. Restrictions apply.

$$T_s = Avg\left(\sum_{j=1}^{L} \vec{e}_{sw_j}^{\,G^{(E)}}\right), \quad V_s = Avg\left(\sum_{j=1}^{M} \vec{e}_{sv_j}^{\,G^{(E)}}\right) \tag{16}$$

Then, the final multimodal representations M_t and M_s for two tasks are obtained by merging textual and visual representations:

$$\begin{aligned} M_t &= [T_t; V_t] \\ M_s &= [T_s; V_s] \end{aligned} \tag{17}$$

where $M_t, M_s \in R^{L+M}$.

F. Modal Prediction

We then use two softmax decoders to perform traffic event detection and sentiment prediction, where the multimodal representations M_t and M_s of the i^{th} sample are treated as inputs:

$$\begin{aligned} \hat{y}_t^i &= softmax\left(W^t M_t + b^t\right) \\ \hat{y}_s^i &= softmax\left(W^s M_s + b^s\right) \end{aligned} \tag{18}$$

where $\hat{y}_t^i$ and $\hat{y}_s^i$ are the predicted distributions for two tasks.

Model training. In our framework, we use cross entropy as the objective functions ζ_t, ζ_s for traffic event detection and sentiment classification and jointly minimize them with different weights. The reason is that traffic event detection is regarded as the main task in this paper.

$$\begin{aligned} \zeta_t &= -\frac{1}{N}\sum_i \sum_z y_{z,t}^i log \hat{y}_{z,t}^i + \lambda_r \|\theta\|^2 \\ \zeta_s &= -\frac{1}{N}\sum_i \sum_z y_{z,s}^i log \hat{y}_{z,s}^i + \lambda_r \|\theta\|^2 \\ \zeta &= w_t \zeta_t + w_s \zeta_s \end{aligned} \tag{19}$$

where $y_{z,t}^i$ and $y_{z,s}^i$ denote the ground truth, and w_t and w_s are weights.

IV. Experiments

A. Experimental Settings

We will answer the following two questions from an experimental viewpoint:

RQ1: Does sentiment mining help traffic event detection?

RQ2: How can the effectiveness of the proposed MCGAT be checked?

1) Datasets: Since joint traffic event detection and sentiment classification is a new research topic, there is a lack of a benchmarking dataset. In line with previous traffic event detection approaches [7], [57], [58], we also choose to create two social media-based traffic event datasets. (1) We create a weakly labeled multimodal traffic and sentiment dataset. Similar to the work [22], a lot of keywords with obvious traffic events are designed to query Getty Images and obtain multimodal samples. For example, we use "traffic crashes" and "sad" to query Getty Images and annotate the retrieved textual and visual documents with "traffic incident" and "negative" labels. We construct a new multimodal Getty Image traffic event and sentiment (MGTES) dataset containing 1,500 positive and 3,500 negative traffic samples.

TABLE I

Dataset Statistics

Dataset	*Task*	*Classes*	*Num.*	*RC(%)*
MGTES	Sentiment	Positive	1500	50.0
		Negative	3500	50.0
	Traffic Event	Incidents	1900	38.0
		Weather Condition	1200	24.0
		Traffic congestion	550	11.0
		Light traffic	1350	27.0
Twitter	Sentiment	Positive	840	38.2
		Negative	1360	61.8
	Traffic Event	Incidents	880	40.0
		Weather Condition	570	25.9
		Traffic congestion	760	34.5
		Light traffic	550	25.0

(2) We create another dataset by collecting data from Twitter. We use Twint,[1] which is a Twitter API, to exact traffic-related image-text posts generated from May 1, 2018, to May 1, 2020. We also apply preprocessing approaches to clean and filter noisy tweets. Finally, our dataset contains 2,200 multimodal samples. Then, we recruited three volunteers to annotate the samples. Each tweet is manually labeled with its corresponding traffic event types and sentiment polarity. We conducted an assessment of the reliability by calculating Kappa scores ($\kappa = 0.56$), as shown in Table I.

2) Evaluation metrics: precision (P), *recall* (R) and *micro-F1* (M_i-F1) and *balanced accuracy* are used as evaluation metrics in our experiments

3) Hyperparameter: The dimensionality of word embeddings is 768. All weight matrices are initialized with zeros. We use the Adam algorithm to train the network, and the number of epochs is set to 100. The dropout rate is set to 0.3, and the batch size is 64. We set a parameter pool, which contains a learning rate in {0.001,0.005,0.01}. We argue that we empirically set the parameter pool since these learning rates are widely used in various deep learning models.

4) Preprocessing: For the images, the overly large images (i.e., size exceeding 1,000 pixels*1,000 pixels) are all resized to the same size. For the text-based document, we first clean up all the texts by looking for unreadable characters and automatically fixing spelling errors. Using the NLTK library in Python, the repetitive and stop words are eliminated [59]. We also use it to perform word tokenization.

5) Experimental environment: In this work, we we mainly use two software packages, PyTorch 1.12 and Python 3.8, which are installed on Ubuntu 20.04, to support our implementation. All experiments are performed on a Dell server with the following characteristics: 128 GB of RAM, GeForce GTX 3080Ti and 10-core CPU processor. It takes about 120 minutes for the state-of-the-art system (i.e., MCGAT) to train its best performance.

B. Baselines

We list a range of strong baselines for comparison and include two popular machine learning approaches, five single modal deep learning approaches and three multimodal graph convolutional network (GCN) approaches.

[1] https://github.com/twintproject/twint

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 08:18:37 UTC from IEEE Xplore. Restrictions apply.

TABLE II

COMPARISON WITH BASELINES. BEST PERFORMANCES ARE IN **BOLD**

Dataset	Baselines	Traffic Event Detection				Sentiment Classification			
		P(%)	R(%)	$\mathrm{M_i}$-F1(%)	Acc.(%)	P(%)	R(%)	$\mathrm{M_i}$-F1(%)	Acc.(%)
MGTES	SVM	52.17	53.24	53.19	53.24	62.55	62.34	62.41	62.37
	RF	51.32	52.89	52.76	52.91	63.44	63.57	63.50	63.57
	CNN	53.77	53.94	53.85	53.94	63.66	63.72	63.71	63.72
	BiGRU	53.40	53.57	53.56	53.72	63.78	64.02	64.02	64.11
	MAGRU	54.55	54.61	54.60	54.69	64.54	64.67	64.62	64.71
	BERT	57.29	57.46	57.44	57.55	68.11	68.32	68.27	68.44
	ResNet	52.33	52.46	52.39	52.45	62.26	62.45	62.37	62.42
	GAT	56.43	56.68	56.59	56.69	66.89	67.14	67.11	67.24
	Early-fusion	59.35	59.61	59.54	59.63	70.45	70.57	70.53	70.60
	Late-fusion	60.77	60.72	60.72	60.71	72.02	72.34	72.21	72.38
	D-S fusion	60.15	60.24	60.18	60.24	72.57	72.72	72.64	72.72
	Text-MCGAT	59.22	59.34	59.25	59.37	69.46	69.71	69.65	69.72
	Image-MCGAT	53.46	53.22	53.22	53.31	63.75	63.92	63.92	63.92
	MCGAT	**62.21**	**62.30**	**62.25**	**62.35**	**75.23**	**75.27**	**75.25**	**73.30**
	△SOTA	(+2.3%)	(+2.4%)	(+2.4%)	(+2.4%)	(+3.6%)	(+3.5%)	(+3.5%)	(+3.6%)
Twitter	SVM	59.57	59.63	59.63	59.66	66.23	66.41	66.40	66.42
	RF	59.74	59.91	59.89	59.89	67.17	67.23	67.21	67.25
	CNN	60.05	60.00	59.97	60.03	68.45	68.56	68.54	68.62
	BiGRU	59.65	59.72	59.68	59.72	67.43	67.63	67.60	67.63
	MAGRU	61.34	61.47	61.41	61.48	68.52	68.64	68.59	68.67
	BERT	63.77	63.84	64.81	64.83	71.32	71.39	71.34	71.40
	ResNet	59.33	59.57	59.45	59.55	66.66	66.72	66.57	66.72
	GAT	62.51	62.37	62.41	62.37	70.13	70.29	70.22	70.29
	Early-fusion	65.11	65.34	65.30	65.34	74.22	74.41	74.38	74.41
	Late-fusion	65.58	65.69	65.64	65.69	74.57	74.74	74.70	74.74
	D-S fusion	65.34	65.48	65.44	65.46	74.43	74.59	74.51	74.59
	Text-MCGAT	64.55	64.57	64.56	64.57	73.75	73.96	73.88	73.96
	Image-MCGAT	61.74	61.88	61.82	61.88	68.52	68.67	68.60	68.67
	MCGAT	**67.42**	**68.07**	**67.79**	**68.09**	**76.09**	**76.18**	**76.13**	**76.19**
	△SOTA	(+2.8%)	(+2.7%)	(+2.7%)	(+2.7%)	(+2.0%)	(+2.1%)	(+2.1%)	(+2.0%)

SVM represents the textual tweets using GloVe vectors. The kernel function of SVM is set to "RBF."

RF models the textual tweets using GloVe vectors and classifies them using a random forest classifier. We set the number of trees to 100.

CNN [60] consists of two convolutional layers and a fully connected layer, which is trained on top of word embeddings for traffic event classification.

BiGRU [61] uses a bidirectional GRU network for learning the hidden states of documents and then uses a softmax function to judge traffic events.

MAGRU [62] adopts the multihead attention mechanism to extract the most significant textual features and feeds them to a softmax function for traffic event detection.

BERT [63] learns the sentence representation using BERT and feeds it into an SVM classifier.

ResNet learns the visual image representation using the pretrained ResNet network and feeds it into an SVM classifier.

GAT [53] In the conventional syntactic dependency graph, GAT employs the GloVe vectors as the node's local feature vector and represents the final sentence using the mean of all node vectors.

Early fusion uses BERT to extract the textual features and uses GAT to extract visual features and connect them together as the multimodal representation. It belongs to an early fusion approach.

Late fusion adopts BERT to extract the textual features and uses GAT to extract visual features and feeds them into two softmax functions to obtain their local decisions. Then, such decision vectors are merged as the decision fused features and forwarded through an SVM classifier.

D-S fusion takes a similar strategy as the late-fusion approach but uses D-S evidence theory as the decision fusion rule.

C. Results and Analysis

The experimental performance of all baselines is shown in Table II.

1) Traffic event detection: In this study, we regard traffic event detection as the main task. (1) We analyze the performance on MGTES. From Table II, we can see that two machine learning approaches, i.e., SVM and RF, perform poorly against other baselines. Shallow network learning cannot extract features as effectively as deep learning. In addition, SVM obtains slight improvements over RF. As two simple deep learning baselines, CNN and BiGRU exceed the abovementioned approaches and obtain good results. This indicates the superior potential of deep neural networks over machine learning approaches. Moreover, CNN slightly outperforms BiGRU for the task of traffic event detection on both datasets. We argue that CNN and BiGRU have different focuses, where CNN focuses on modeling spatial correlation in data, while BiGRU pays more attention to modeling sequence information. After investigating two datasets, we know that traffic records from the Getty Image and Twitter datasets are more diverse and informal. The length of the traffic

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 08:18:37 UTC from IEEE Xplore. Restrictions apply.

record is very short where the sequence dependency across words is relatively weak. The spatial correlation across words plays a more important role than the sequence information. Hence, CNN performs better than BiGRU by a small margin. By introducing a multihead attention mechanism into the model design, MAGRU can learn the most important features and perform better than CNN and BiGRU. Its f1 and accuracy scores reach 54.6% and 54.69%, respectively. BERT, which is a popular pretrained language model, takes a further step by emphasizing the importance of the attention mechanism and designs a multilayer attentive network. It outperforms MAGRU by a large margin. One possible reason is that the pretrained textual features provided by BERT have stronger discrimination ability. ResNet performs very poorly since visual semantic analysis involves a higher-level abstraction than text analysis. How to fill the "semantic gap" is still an unsolved issue. However, ResNet still overcomes SVM and RF. GAT achieves the second-best performance among all baselines, whose f1 and accuracy scores are 56.59% and 56.59%. It can learn both syntax information and context knowledge from long distances. By leveraging multimodal information, early fusion, late fusion and D-S fusion are better than single BERT because multimodal information can provide more accurate descriptions than unimodal information. Among them, late fusion and D-S fusion outperform early fusion, which implies that decision-level fusion is more effective than feature-level fusion. However, the former introduces expensive computations. Late fusion's f1 and accuracy scores reach 60.72% and 60.71%, respectively.

(2) We observe a similar phenomenon from the Twitter dataset. SVM and RF still obtain the worst experimental results among all baselines. However, CNN and BiGRU obtain comparable results against SVM and RF because tweet texts are informal, short and more difficult to analyze. Hence, performing convolutional operations and modeling long-term dependency are not useful. BERT and GAT perform better than CNN and MAGRU but are weaker than the three multimodal approaches. Text-MCGAT and Image-MCGAT do not perform very well, especially the poor performance of Image-MCGAT, demonstrating that text and visual modalities cannot be treated independently for multimodal traffic event detection. Among them, Text-MCGAT achieves comparatively good F1 scores, as we expected. Textual information has been proven to be the most important part of multimodal affection analysis. Image semantic understanding is more complex than text understanding, which involves a higher level of abstraction. In this case, the proposed MCGAT model remarkably overcomes all baselines and achieves state-of-the-art performance with f1 and accuracy scores of 67.79% and 68.09%, respectively. We attribute the main improvements to both cross-modal connection and cross-task connection, which ensures that MCGAT can model intermodality fusion and multitask interaction. For illustration, a complete ROC curve is shown in Fig. 3.

2) Sentiment Classification: Sentiment classification is treated as the secondary task, but we also share traffic event information with it to check whether traffic knowledge could help sentiment analysis. Table II shows that SVM, RF, CNN and BiGRU achieve comparable f1 and accuracy scores. CNN outperforms BiGRU on the Twitter dataset but is weaker on the MGTES dataset. The reason is that sentiment is quite distinct from traffic events, which involve a higher level of abstraction and subjectivity. Sentiment classification relies more on the sequence information than the spatial correlation. Textual samples from the MGTES dataset have longer lengths than those from the Twitter dataset. Hence, BiGRU gains better classification performance on the MGTES dataset. MAGRU obtains slight improvement over the above four approaches. The reason is that a multihead attention mechanism is introduced to assign greater weights to important words. BERT and GAT outperform all other baselines by a large margin since both are state-of-the-art deep neural networks. By incorporating multimodal information, three multimodal GAT approaches achieve the best performance in terms of all metrics. Again, decision-level fusion approaches exceed the feature-level fusion approach. One possible reason is that feature concatenation cannot provide more effective representation. Text-MCGAT and Image-MCGAT did not perform well, as we expected. This shows that it is necessary to

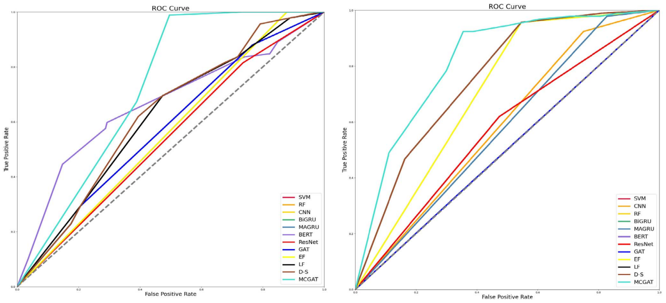

Fig. 3. The ROC curves on both datasets.

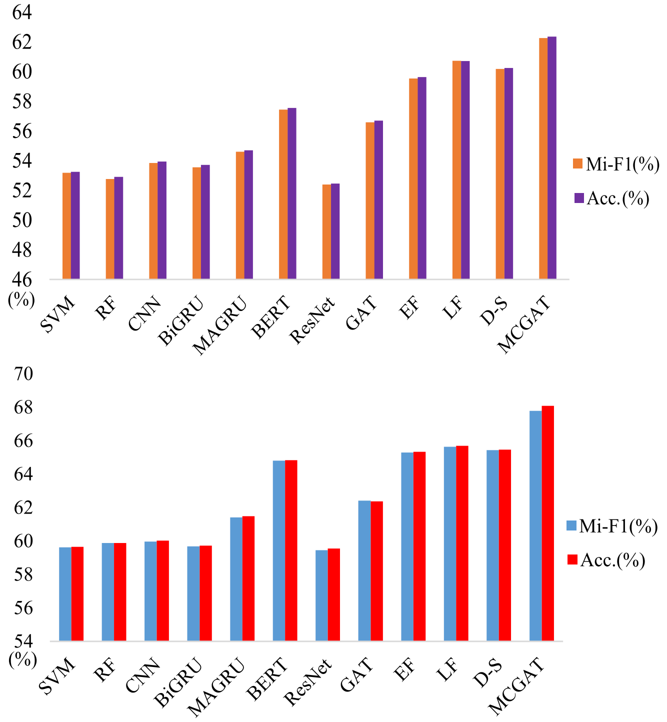

Fig. 4. The performance of all models on the MGTES and Twitter datasets.

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 08:18:37 UTC from IEEE Xplore. Restrictions apply.

TABLE III
SEMANTICMCGAT V/S MCGAT

Dataset	Models	Traffic		Sentiment	
		F1(%)	Acc.(%)	F1(%)	Acc.(%)
MGTES	SemanticMCGAT MCGAT	62.77 62.25	62.84 62.35	76.32 75.25	74.77 73.30
Twitter	SemanticMCGAT MCGAT	68.42 67.79	69.74 68.09	76.94 76.13	77.35 76.19

model multimodal sentiment information, as human language is multimodal. A similar phenomenon occurs for Text-MCGAT and Image-MCGAT, where Text-MCGAT performs better and Image-MCGAT performs worse. This again proves the importance of textual information for sentiment classification. Finally, our MCGAT achieves state-of-the-art results among all baselines with f1 and accuracy scores of 75.25%, 73.3% and 76.13%, 76.19% on the two datasets. This remarkable improvement benefits from the proposed four subgraphs. We will explore their contribution to the ablation test. All these experiments prove the feasibility of multitask learning.

D. Selection of Visual Vertices

In the proposed MCGAT model, we split the whole image into 20 blocks and take each of them as a vertex in the visual graph. We clarify that choosing a pixel or a visual block as the vertex in a graphical neural network is a widely used and natural choice. However, we also attempt to explore the role of semantic blocks via the image segmentation approach in improving performance. Hence, we introduce Google DeepLab v3 [64], which is a strong semantic segmentation architecture to split the whole image, treat each visual region as a vertex in the visual graph, and learn its representation via the pretrained ResNet model, termed SemanticMCGAT. We compare such an extended variant with the standard MCGAT model in Table III.

SemanticMCGAT slightly outperforms the standard MCGAT model for both traffic detection and sentiment detection tasks. One possible reason is that the visual blocks in SemanticMCGAT capture more semantic information than those in MCGAT. The success of this simple strategy demonstrates the enormous potential of the proposed MCGAT model. It may achieve better classification performance if superior visual semantic modeling attempts have been made, which will be left to our future work.

E. STL v/s MTL Setup

To prove the advantage of multitask learning (MTL) over single-task learning (STL), we compare MTL with STL in Table IV. We observe the following: (1) for two setups, the multimodal setup overcomes the unimodal setup, which shows that multimodal setups provide richer features; and (2) MTL obtains higher performance than STL on two datasets. The reason is that explicitly modeling the correlation between two tasks with a graph connection network can encourage the model to effectively leverage the shared knowledge of one task for another task. Moreover, we also see that text models are better than visual models for both setups, which implies that text information plays a more important role than visual documents. Now, the first and second research questions have been answered.

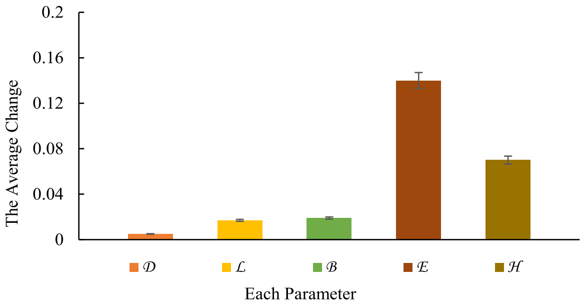

Fig. 5. Sensitivity study.

F. Ablation Test

To explore the attribution of each component in MCGAT, we choose to design three submodels. (1) *No cross-modal connection* replaces the cross-modal connection layer with multi-modal feature concatenation. (2) *No cross-task connection* removes the cross-task connection layer from the proposed MCGAT model. (3) *No GAT* replaces the GAT architecture with the standard bidirectional RNN.

From Table V, we notice that *No GAT* achieves comparable results against MCGAT, showing that the graph network contributes slightly to the performance of MCGAT. *No cross-task connection* performs the worst among the baselines, which indicates that cross-task connection plays the most role in MCGAT. This proves the importance of modeling multitask correlation. By analyzing the *no cross-modal connection*'s scores, we know that the cross-modal connection layer also has a great influence on the whole performance of MCGAT. Each component is necessary for MCGAT.

G. Sensitivity Analysis

In this section, we perform a sensitivity test to explore how much each hyperparameter affects the model performance. We adjust one hyperparameter and leave the other parameters unchanged. Then, we show the change in model performance in Fig. 5.

We mainly analyze five core parameters in the proposed MCGAT model, i.e., dropout rate (denoted as $\mathcal{D}$), learning rate (denoted as $\mathcal{L}$), batch size (denoted as $\mathcal{B}$), the number of epochs (denoted as $\mathcal{E}$) and the number of heads (denoted as $\mathcal{H}$). Fig.5 shows the average change in model performance.

Fig. 5 shows that the number of epochs has the greatest influence on model performance, e.g., the performance difference between 100 epochs and 50 epochs is substantial. The average change in model performance reached almost 15%. The second important parameter is the number of heads in the attention mechanism, where the average change reaches 8%. The learning rate and batch size are seen as important parameters, which may have serious consequences for the model performance.

H. Case Study

To explore the potential of the proposed cross-modal connection and cross-task connection modules, we perform a case study and exhibit a few multimodal cases, as shown in Fig. 6.

For traffic event detection, we note that the proposed cross-modal connection mechanism can effectively solve the

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 08:18:37 UTC from IEEE Xplore. Restrictions apply.

TABLE IV

COMPARISON BETWEEN SINGLE-TASK LEARNING (STL) AND MULTI-TASK (MTL) LEARNING FRAMEWORKS

Task	Setups	T		V		T+V	
		$\mathrm{M_i}$-F1(%)	Acc.(%)	$\mathrm{M_i}$-F1(%)	Acc.(%)	$\mathrm{M_i}$-F1(%)	Acc.(%)
Traffic event MGTES	STL	56.59	56.69	52.17	52.26	61.32	61.41
	MTL	59.25	59.37	53.22	53.31	62.25	62.35
Traffic event Twitter	STL	62.41	62.37	59.56	59.63	66.13	66.24
	MTL	64.56	64.57	61.82	61.88	67.79	68.09

TABLE V

ABLATION TEST

Dataset	Models	Metrics	
		F1(%)	Acc.(%)
MGTES	No cross-modal connection	61.77	61.82
	No cross-task connection	61.32	61.41
	No GAT	62.14	62.20
	MCGAT	62.25	62.35
Twitter	No cross-modal connection	66.87	66.94
	No cross-task connection	66.13	66.24
	No GAT	67.06	67.22
	MCGAT	67.79	68.09

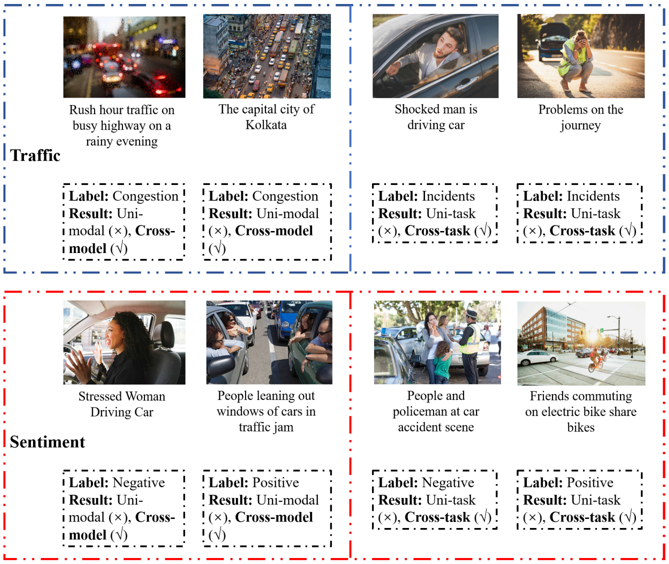
Rush hour traffic on busy highway on a rainy evening
The capital city of Kolkata
Shocked man is driving car
Problems on the journey
Traffic
Label: Congestion **Result:** Uni-modal (×), **Cross-model** (√)
Label: Congestion **Result:** Uni-modal (×), **Cross-model** (√)
Label: Incidents **Result:** Uni-task (×), **Cross-task** (√)
Label: Incidents **Result:** Uni-task (×), **Cross-task** (√)
Stressed Woman Driving Car
People leaning out windows of cars in traffic jam
People and policeman at car accident scene
Friends commuting on electric bike share bikes
Sentiment
Label: Negative **Result:** Uni-modal (×), **Cross-model** (√)
Label: Positive **Result:** Uni-modal (×), **Cross-model** (√)
Label: Negative **Result:** Uni-task (×), **Cross-task** (√)
Label: Positive **Result:** Uni-task (×), **Cross-task** (√)

Fig. 6. Multimodal samples where cross-modal and cross-task setups perform better than unimodal and uni-task setups.

differences between different modalities. It can help the model obtain more differentiated information and improve the complementarity between modalities by using one modal information to update another modal representation. In addition, for a few samples where the textual and visual records cannot directly describe traffic events, our cross-task approach could leverage the shared and related sentimental knowledge to help make predictions. For example, "shocked man is driving car" does not present a traffic incident, while his angry emotion is likely to lead to a traffic incident. By leveraging such extra knowledge, the proposed model makes correct decisions.

For sentiment classification, we have similar findings. The visual image of the first sample presents the fact that the woman is smiling while its text counterpart shows clear negative sentiment. The proposed cross-model connection can take multimodal interaction into consideration. Moreover, even though the visual and textual documents do not present any emotion-related scenes or words, our model could infer correct sentiment by using the shared knowledge from traffic event detection. Such samples show the interaction between multimodal fusion, multitask correlation and the two tasks.

V. CONCLUSION AND FUTURE STUDIES

Traffic event and sentiment joint detection is a new research task in ITSs and is very important to improve the intelligence level of traffic services. However, none of the recent approaches in traffic event detection have taken multitask learning under consideration. To fill this gap, we present a multimodal coupled graph attention network (MCGAT), which leverages both multimodal and affective shared knowledge, to identify traffic events from the social network. A cross-modal connection layer is designed to learn effective multimodal representations, and a cross-task connection layer is proposed to capture sentiment knowledge from another task. Empirical evaluation on two benchmarking datasets shows the effectiveness of the proposed model over state-of-the-art baselines. In addition, the proposed MCGAT model is a universal paradigm that can also be used to other multimodal multitask joint analysis tasks, e.g., multimodal sentiment and sarcasm detection, and multimodal sentiment and gesture recognition. The model construction and training are similar to traffic event and sentiment joint detection.

Our study also has a few limitations. The proposed model naively splits the image into blocks instead of using a state-of-the-art image segmentation approach to select specific semantic blocks, which limits the potential of the proposed model. In addition, we adopt the hard parameter sharing mechanism to design cross-task connections, while the soft parameter sharing mechanism has not yet been considered. We will solve such limitations in future.

Since there is lack of benchmarking multi-task learning datasets in ITS, future studies will focus on creating a multimodal multi-task dataset, which involves text, image and audio modalities. In view of the popularity of the pretrained language model, we also plan to build a multimodal traffic event-enriched transformer.

REFERENCES

[1] Z. Cui, K. Henrickson, R. Ke, and Y. Wang, "Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4883–4894, Nov. 2020.

[2] A. Salas, P. Georgakis, C. Nwagboso, A. Ammari, and I. Petalas, "Traffic event detection framework using social media," in *Proc. IEEE Int. Conf. Smart Grid Smart Cities (ICSGSC)*, Jul. 2017, pp. 303–307.

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 08:18:37 UTC from IEEE Xplore. Restrictions apply.

[3] E. Alomari, R. Mehmood, and I. Katib, "Road traffic event detection using Twitter data, machine learning, and apache spark," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug. 2019, pp. 1888–1895.

[4] Z. Yu *et al.*, "Mobility-aware proactive edge caching for connected vehicles using federated learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 8, pp. 5341–5351, 2021.

[5] Y. Qia *et al.*, "Privacy-preserving blockchain-based federated learning for traffic flow prediction," *Future Gener. Comput. Syst.*, vol. 117, pp. 328–337, Apr. 2021.

[6] Y. Zhang *et al.*, "Learning interaction dynamics with an interactive LSTM for conversational sentiment analysis," *Neural Netw.*, vol. 133, pp. 40–56, Jan. 2021.

[7] Y. Seliverstov, S. Seliverstov, I. Malygin, and O. Korolev, "Traffic safety evaluation in northwestern federal district using sentiment analysis of internet users' reviews," *Transp. Res. Proc.*, vol. 50, pp. 626–635, 2020.

[8] A. AlDhanhani, E. Damiani, R. Mizouni, and D. Wang, "Analysis of shapelet transform usage in traffic event detection," in *Proc. IEEE Int. Conf. Cognit. Comput. (ICCC)*, Jul. 2018, pp. 41–48.

[9] S. Xu, S. Li, R. Wen, and W. Huang, "Traffic event detection using Twitter data based on association rules," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vols. IV–2/W5, pp. 543–547, May 2019.

[10] K. Lin, J. Luo, L. Hu, M. Hossain, and A. Ghoneim, "Localization based on social big data analysis in the vehicular networks," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 1932–1940, Aug. 2017.

[11] Y. Imamverdiyev and L. Sukhostat, "Anomaly detection in network traffic using extreme learning machine," in *Proc. IEEE 10th Int. Conf. Appl. Inf. Commun. Technol. (AICT)*, Oct. 2016, pp. 1–4.

[12] R. W. Liu, J. Nie, S. Garg, Z. Xiong, Y. Zhang, and M. S. Hossain, "Data-driven trajectory quality improvement for promoting intelligent vessel traffic services in 6G-enabled maritime IoT systems," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5374–5385, Apr. 2021.

[13] M. T. Zulfikar *et al.*, "Detection traffic congestion based on Twitter data using machine learning," *Proc. Comput. Sci.*, vol. 157, pp. 118–124, 2019.

[14] E. Hodo, X. Bellekens, E. Iorkyase, A. Hamilton, C. Tachtatzis, and R. Atkinson, "Machine learning approach for detection of nonTor traffic," in *Proc. 12th Int. Conf. Availability, Rel. Secur.*, Aug. 2017, pp. 1–6.

[15] N. A. Zardari, R. Ngah, O. Hayat, and A. H. Sodhro, "Adaptive mobility-aware and reliable routing protocols for healthcare vehicular network," *Math. Biosciences Eng.*, vol. 19, no. 7, pp. 7156–7177, 2022.

[16] S. Wan, X. Xu, T. Wang, and Z. Gu, "An intelligent video analysis method for abnormal event detection in intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4487–4495, Jul. 2021.

[17] S. Dabiri and K. Heaslip, "Developing a Twitter-based traffic event detection model using deep learning architectures," *Exp. Syst. Appl.*, vol. 118, pp. 425–439, Mar. 2019.

[18] Z. Zhang, Q. He, J. Gao, and M. Ni, "A deep learning approach for detecting traffic accidents from social media data," *Transp. Res. C, Emerg. Technol.*, vol. 86, pp. 580–596, Jan. 2018.

[19] A. Aboah, M. Shoman, V. Mandal, S. Davami, Y. Adu-Gyamfi, and A. Sharma, "A vision-based system for traffic anomaly detection using deep learning and decision trees," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4207–4212.

[20] X. Zhang, Y. Chen, M. Liu, and C. Huang, "Acoustic traffic event detection in long tunnels using fast binary spectral features," *Circuits, Syst., Signal Process.*, vol. 39, no. 6, pp. 2994–3006, 2020.

[21] Q. Chen and W. Wang, "Multi-modal neural network for traffic event detection," in *Proc. IEEE 2nd Int. Conf. Electron. Commun. Eng. (ICECE)*, Dec. 2019, pp. 26–30.

[22] Y. Zhang *et al.*, "A quantum-inspired multimodal sentiment analysis framework," *Theor. Comput. Sci.*, vol. 752, pp. 21–40, Dec. 2018.

[23] Y. Zhang *et al.*, "CFN: A complex-valued fuzzy network for sarcasm detection in conversations," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 12, pp. 3696–3710, Dec. 2021.

[24] Y. Liu, Y. Zhang, Q. Li, B. Wang, and D. Song, "What does your smile mean? Jointly detecting multi-modal sarcasm and sentiment using quantum probability," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, 2021, pp. 871–880.

[25] Y. Zhang, P. Tiwari, L. Rong, R. Chen, N. A. AlNajem, and M. S. Hossain, "Affective interaction: Attentive representation learning for multi-modal sentiment classification," *ACM Trans. Multimedia Comput., Commun., Appl.*, Mar. 2022.

[26] S. U. Amin *et al.*, "Deep learning for EEG motor imagery classification based on multi-layer CNNs feature fusion," *Future Gener. Comput. Syst.*, vol. 101, pp. 542–554, Dec. 2019.

[27] A. Kumar and G. Garg, "Sentiment analysis of multimodal Twitter data," *Multimedia Tools Appl.*, vol. 78, pp. 1–17, Mar. 2019.

[28] Y. Jiang, W. Li, M. S. Hossain, M. Chen, and M. Al-Hammadi, "A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition," *Inf. Fusion*, vol. 53, pp. 209–221, Jan. 2020.

[29] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. EMNLP*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1103–1114.

[30] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, "Image–text sentiment analysis via deep multimodal attentive fusion," *Knowl.-Based Syst.*, vol. 167, pp. 26–37, Mar. 2019.

[31] M. S. Hossain and G. Muhammad, "Emotion recognition using secure edge and cloud computing," *Inf. Sci.*, vol. 504, pp. 589–601, Dec. 2019.

[32] A. Yadav and D. Kumar Vishwakarma, "A weighted text representation framework for sentiment analysis of medical drug reviews," in *Proc. IEEE 6th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2020, pp. 326–332.

[33] A. Yadav and D. K. Vishwakarma, "A language-independent network to analyze the impact of COVID-19 on the world via sentiment analysis," *ACM Trans. Internet Technol.*, vol. 22, no. 1, pp. 1–30, Feb. 2022.

[34] A. Yadav, A. Agarwal, and D. K. Vishwakarma, "XRA-Net framework for visual sentiments analysis," in *Proc. IEEE 5th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2019, pp. 219–224.

[35] Y. Liu *et al.*, "Deep anomaly detection for time-series data in industrial IoT: A communication-efficient on-device federated learning approach," *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6348–6358, 2021.

[36] M. S. Hossain and G. Muhammad, "Emotion-aware connected healthcare big data towards 5G," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2399–2406, Aug. 2018.

[37] A. Yadav, "A multilingual framework of CNN and Bi-LSTM for emotion classification," in *Proc. 11th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, 2020, pp. 1–6.

[38] A. Agarwal, A. Yadav, and D. K. Vishwakarma, "Multimodal sentiment analysis via RNN variants," in *Proc. IEEE Int. Conf. Big Data, Cloud Comput., Data Sci. Eng. (BCD)*, May 2019, pp. 19–23.

[39] A. Yadav and D. K. Vishwakarma, "A deep learning architecture of RA-DLNet for visual sentiment analysis," *Multimedia Syst.*, vol. 26, no. 4, pp. 431–451, Aug. 2020.

[40] G. Muhammad, M. S. Hossain, and N. Kumar, "EEG-based pathology detection for home health monitoring," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 2, pp. 603–610, Feb. 2020.

[41] P. Tiwari, H. Zhu, and H. M. Pandey, "DAPath: Distance-aware knowledge graph reasoning based on deep reinforcement learning," *Neural Netw.*, vol. 135, pp. 1–12, Mar. 2021.

[42] Y. Zhang, R. Wang, M. S. Hossain, M. F. Alhamid, and M. Guizani, "Heterogeneous information network-based content caching in the Internet of Vehicles," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 10216–10226, Oct. 2019.

[43] Y. Zhang *et al.*, "Stance level sarcasm detection with BERT and stance-centered graph attention networks," *ACM Trans. Internet Technol.*, to be published. Apr. 2022.

[44] S. Qian, T. Zhang, C. Xu, and M. S. Hossain, "Social event classification via boosted multimodal supervised latent Dirichlet allocation," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 11, no. 2, pp. 1–22, Jan. 2015.

[45] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Attention-based multi-modal sentiment analysis and emotion detection in conversation using RNN," in *Int. J. Interact. Multimedia Artif. Intell.*, vol. 6, no. 6, pp. 1–10, 2021.

[46] X. Li, J. Li, Y. Zhang, and P. Tiwari, "Emotion recognition from multi-channel EEG data through a dual-pipeline graph attention network," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2021, pp. 3642–3647.

[47] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio-visual emotional big data," *Inf. Fusion*, vol. 49, pp. 69–78, Sep. 2019.

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 08:18:37 UTC from IEEE Xplore. Restrictions apply.

[48] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," 2021, *arXiv:2102.04830*.

[49] Y. Zhang, Z. Zhao, P. Wang, X. Li, L. Rong, and D. Song, "ScenarioSA: A dyadic conversational database for interactive sentiment analysis," *IEEE Access*, vol. 8, pp. 90652–90664, 2020.

[50] P. Wang, Y. Zhang, X. Li, Y. Hou, and D. Song, "Does tang poetry affect human emotional state? A pilot study by EEG," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2017, pp. 2044–2047.

[51] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, "Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5415–5421.

[52] C. Zhang, Q. Li, and D. Song, "Aspect-based sentiment classification with aspect-specific graph convolutional networks," 2019, *arXiv:1909.03477*.

[53] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.

[54] X. Li *et al.*, "EEG based emotion recognition: A tutorial and review," *ACM Comput. Surveys*, early access, Mar. 2022.

[55] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1624–1639, Dec. 2011.

[56] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the ResNet model for visual recognition," *Pattern Recognit.*, vol. 90, pp. 119–133, Jun. 2019.

[57] J. Cao *et al.*, "Web-based traffic sentiment analysis: Methods and applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 844–853, Apr. 2014.

[58] Y. Zhang, Y. Li, R. Wang, M. S. Hossain, and H. Lu, "Multi-aspect aware session-based recommendation for intelligent transportation services," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4696–4705, Jul. 2021.

[59] Y. Zhang, D. Song, X. Li, and P. Zhang, "Unsupervised sentiment analysis of Twitter posts using density matrix representation," in *Proc. Eur. Conf. Inf. Retr.* Grenoble, France: Springer, 2018, pp. 316–329.

[60] Y. Zhang *et al.*, "A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis," *Inf. Fusion*, vol. 62, pp. 14–31, Oct. 2020.

[61] Y. Zhang, D. Song, P. Zhang, X. Li, and P. Wang, "A quantum-inspired sentiment representation model for Twitter sentiment analysis," *Appl. Intell.*, vol. 49, no. 8, pp. 3093–3108, Aug. 2019.

[62] A. Kumar, V. T. Narapareddy, V. A. Srikanth, A. Malapati, and L. B. M. Neti, "Sarcasm detection using multi-head attention based bidirectional LSTM," *IEEE Access*, vol. 8, pp. 6388–6397, 2020.

[63] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[64] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with Atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

Yazhou Zhang received the Ph.D. degree from the College of Intelligence and Computing, Tianjin University, Tianjin, China, in 2020. He is currently a Lecturer with the Software Engineering College, Zhengzhou University of Light Industry, Zhengzhou, China. His research interests include opinion mining (or sentiment analysis), data fusion, and quantum cognition. He is currently working on developing quantum inspired sentiment analysis models and their application to problems like conversational sentiment analysis and information fusion.

Prayag Tiwari received the Ph.D. degree from the University of Padova, Italy. He is currently working as an Assistant Professor at Halmstad University, Sweden. Previously, he worked as a Post-Doctoral Researcher at the Aalto University, Finland, and Marie Curie Researcher at the University of Padova. He has several publications in top journals and conferences. His research interests include machine learning, deep learning, quantum machine learning, NLP, healthcare, and the IoT.

Qian Zheng received the Ph.D. degree from Southern Medical University. She is currently a Lecturer with the Software Engineering College, Zhengzhou University of Light Industry, Zhengzhou, China. Her research interests include opinion mining (or sentiment analysis), medical image fusion, and medical image analysis.

Abdulmotaleb El Saddik (Fellow, IEEE) is the Acting Chair of the Computer Vision Department at MBZUAI, United Arab Emirates and a Distinguished University Member at the University of Ottawa, Canada. He is an internationally recognized scholar who has made strong contributions to the knowledge and understanding of intelligent multimedia computing, communications, and applications. He has coauthored ten books and more than 600 publications and chaired more than 50 conferences and workshops and has supervised more than 150 researchers. He has received research grants and contracts totaling more than 20 million dollars. He is the author of the book *Haptics Technologies: Bringing Touch to Multimedia*. His research focus is on the establishment of digital twins to enhance the quality of life of citizens using AI, the IoT, SN, AR/VR, haptics and 5G to allow people to interact in real-time with one another as well as with their smart digital representations in the metaverse in a secure manner. He is a fellow of the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada. He is an ACM Distinguished Scientist and has received several awards, including the Friedrich Wilhelm Bessel Award from the German Humboldt Foundation and the IEEE Instrumentation and Measurement Society Technical Achievement Award. He also received the IEEE Canada C. C. Gotlieb (Computer) Medal and the A. G. L. McNaughton Gold Medal for important contributions to the field of computer engineering and science and the IEEE TCSC Achievement Award for Excellence in Scalable Computing. He is the Editor-in-Chief of the *ACM Transactions on Multimedia Computing, Communications and Applications* (ACM TOMM).

M. Shamim Hossain (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Ottawa, ON, Canada, in 2009. He is currently a Professor with the Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. He is also an Adjunct Professor with the School of Electrical Engineering and Computer Science, University of Ottawa, ON, Canada. He has authored and coauthored more than 350 publications, including refereed journals (280+ SCI/ISI-indexed articles, 150+ IEEE/ACM transactions/journal articles, 23+ ESI highly cited papers, and two hot papers), conference papers, books, and book chapters. His research interests include cloud networking, smart environment (smart city, smart health), AI, deep learning, edge computing, the Internet of Things (IoT), multimedia for health care, and multimedia big data. He has served as the co-chair, the general chair, the workshop chair, the publication chair, and a TPC member in several IEEE and ACM conferences. He is the Chair of the IEEE Special Interest Group on Artificial Intelligence (AI) for Health with the IEEE ComSoc eHealth Technical Committee. He serves as the Co-Chair for the 2nd IEEE GLOBECOM 2022 Workshop on Edge-AI and IoT for Connected Health. He is the Technical Program Co-Chair of ACM Multimedia 2023. He is currently the Chair of the Saudi Arabia Section of the Instrumentation and Measurement Society Chapter. He was a recipient of a number of awards, including the Best Conference Paper Award, the 2016 *ACM Transactions on Multimedia Computing, Communications and Applications* (TOMM) Nicolas D. Georganas Best Paper Award, the 2019 King Saud University Scientific Excellence Award (Research Quality), and the Research in Excellence Award from the College of Computer and Information Sciences (CCIS), King Saud University (three times in a row). He is on the Editorial Board of the IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT (TIM), IEEE TRANSACTIONS ON MULTIMEDIA (TMM), and *ACM Transactions on Multimedia Computing, Communications, and Applications* (TOMM), IEEE MULTIMEDIA, IEEE NETWORK, IEEE WIRELESS COMMUNICATIONS, IEEE ACCESS, *Journal of Network and Computer Applications* (Elsevier), *International Journal of Multimedia Tools and Applications* (Springer), and *Games for Health Journal*. He is a Distinguished Member of the ACM. He is an IEEE Distinguished Lecturer (DL).

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 14,2025 at 08:18:37 UTC from IEEE Xplore. Restrictions apply.