Cross-modal Contrastive Fusion Network For Sentiment Analysis With Dynamic Semantic Diffusion

Haiyun Ma[1*] and Zhonglin Zhang[2]

[1] School of Electronic Information and Electrical Engineering,Tianshui Normal University, Tianshui, 741001, China
[2] School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China
[*] Corresponding author. E-mail:tsmhy1999@163.com

Received: Jun. 25, 2025; Accepted: Aug. 03, 2025

As the importance of public engagement monitoring grows in the face of complex social challenges, analyzing social media data from multiple perspectives has become crucial for understanding diverse public sentiments. Current methods often fall short in effectively supporting decision-making due to their inability to dynamically adapt to the evolving nature of social media discussions. They rely on static strategies that fail to capture the intricate correlations between features across different views, making it difficult to identify sentiment patterns that emerge through complex dependencies in user-generated content. To address these shortcomings, we propose a novel deep multi-view contrastive fusion network (SMOM) designed for comprehensive public opinion monitoring in social media. SMOM features a view-specific feature extractor that captures inherent information within each view. It then employs cross-view contrastive learning to maximize mutual information between view-specific representations, ensuring consistency between views and bridging semantic gaps from an information theory perspective. Furthermore, SMOM implements structure-driven adaptive fusion by combining gate strategies and graph neural networks, enabling the adaptive integration of complementary information. These components work together seamlessly to uncover sentiment patterns, achieving thorough and accurate monitoring of public opinions in social media. Experimental evaluations on social media datasets demonstrate SMOM's superior performance in detecting nuanced public sentiments.

Keywords: Social media sentiment analysis; invariant representation learning; structure-driven adaptive fusion.

http://dx.doi.org/10.6180/jase.202604_29(4).0016

1. Introduction

The daily-generated heterogeneous content on social media platforms, including textual debates, visual protests, and video testimonials, forms a multi-view ecosystem that captures evolving public attitudes across various domains [1–5]. While this multi-view social media data offers significant opportunities for sentiment analysis, it also presents challenges such as dynamism, unstructured nature, and multimodality, which complicate data processing and analysis. In complex sentiment analysis scenarios, extracting valuable information from multi-source heterogeneous social media data and achieving a comprehensive understanding of public opinion have become critical issues for both academia and industry. To tackle these challenges, advanced techniques in data modeling, mining, fusion, and processing are rapidly evolving. These technologies enhance the efficient utilization of data resources and support decision-making through practices like social network analysis [6–8]. Among them, artificial intelligence-driven multimodal fusion technologies have shown great potential in social media sentiment analysis. By integrating diverse modalities such as text, images, and audio, and capturing their deep interactions, these technologies can more accurately reveal public sentiments and opinions, providing robust support for policymaking and market research.

In multi-view social media opinion monitoring, mainstream methods can be categorized into statistical learning-based methods [9, 10] and deep learning-based methods [11, 12]. Statistical learning-based methods typically rely on manually designed features and traditional machine learning models to extract and integrate features from multi-view data for opinion analysis. For instance, for the textual view, statistical features such as Term Frequency-Inverse Document Frequency and n-grams are commonly used; for the user view, metadata such as user activity and influence are extracted; for the network view, relationships between users are modeled [13, 14]. These methods unify multi-view features through concatenation or weighted integration into a feature space, and then use classifiers such as Support Vector Machines or Logistic Regression for sentiment analysis or topic detection. Deep learning-based methods leverage the automatic feature learning capabilities of neural networks to capture deep semantic relationships and contextual information from multi-view data. These methods often employ different network architectures to encode data from each view separately and then integrate multi-view information effectively using attention mechanisms or feature fusion modules.

Current multi-view approaches still exhibit significant shortcomings in aligning and fusing multi-view data. On the one hand, existing methods often focus solely on shallow feature alignment or simple feature concatenation, lacking deep modeling of the consistency and dynamics between views. This limitation makes it challenging for models to fully capture the intrinsic relationships among views, especially in dynamic opinion monitoring scenarios where the correlations between views are highly time-sensitive and complex. For instance, text may directly convey users' opinions, while images or videos often contain implicit emotional or contextual information. Without deep modeling, these methods fail to extract comprehensive cross-view associative features, limiting the ability to understand social media data holistically. On the other hand, most existing methods employ static fusion mechanisms, where all view features are fused according to predetermined, fixed patterns. This approach neglects the dynamic changes of view-specific features in semantic spaces. In reality, different views exhibit nonlinear and dynamic behaviors in distinct semantic spaces. For example, in opinion analysis, the relationships between text and images may evolve over time or shift with the contextual nuances of events. Static fusion methods are unable to dynamically adjust weights or update representations, resulting in a lack of adaptability to dynamically changing environments. Furthermore, fixed fusion patterns may lose critical semantic details during the integration process, introduce redundancy, or even amplify noisy data, ultimately reducing the efficiency and effectiveness of multi-view information utilization and fusion.

To this end, a novel deep multi-view contrastive fusion network is proposed for social media opinion monitoring (SMOM), which contains view-specific feature extractor, cross-view contrastive learning, structure-driven adaptive fusion. Specifically, the view-specific feature extractor is designed to fully explore inherent information in each view where a bi-directional gated recurrent network is used to learn context information of the text view and the multi-layer perceptron networks are used to extract information of the acoustic and image views. Meanwhile, to bridge the semantic gaps between different views, the cross-view contrastive learning is devised via maximizing the mutual information between view-specific representations, which provides a theoretical framework for ensuring consistency between views. Furthermore, the structure-driven adaptive fusion is introduced via integrating the gate strategy and the graph neural network to achieve information fusion in a data-driven manner. The three components collaborate seamlessly, leveraging their respective strengths to effectively achieve comprehensive and accurate opinion monitoring. Finally, experiment results of two opinion monitoring datasets show the advantage and effectiveness of SMOM.

This paper is structured as follows: Section II delves into the methodological framework of SMOM. Section III conducts a systematic evaluation of SMOM. Section IV concludes key contributions of SMOM.

2. Method

Mathematically, consider a social media conversation $X = [x_1, \ldots, x_N]$ with N samples, where the i-th sample x_i consists of the acoustic view x_i^a, the image x_i^v and text view x_i^t. Meanwhile, let $U = \{u_1, \ldots, u_M\}$ $(U \geq 2)$ denotes M users in social media networks. Each sample x_i is expressed via the user $U_{\phi(x_i)}$, where ϕ denotes the corresponding relationship between samples and users. The goal of the social media monitoring is to predict the emotion of each sample. To this end, a novel deep multi-view contrastive fusion network is proposed for social media monitoring, which contains view-specific feature extractor, cross-view contrastive learning, structure-driven adaptive fusion, as shown in Fig. 1. The detailed methodology of three components is provided as follows.

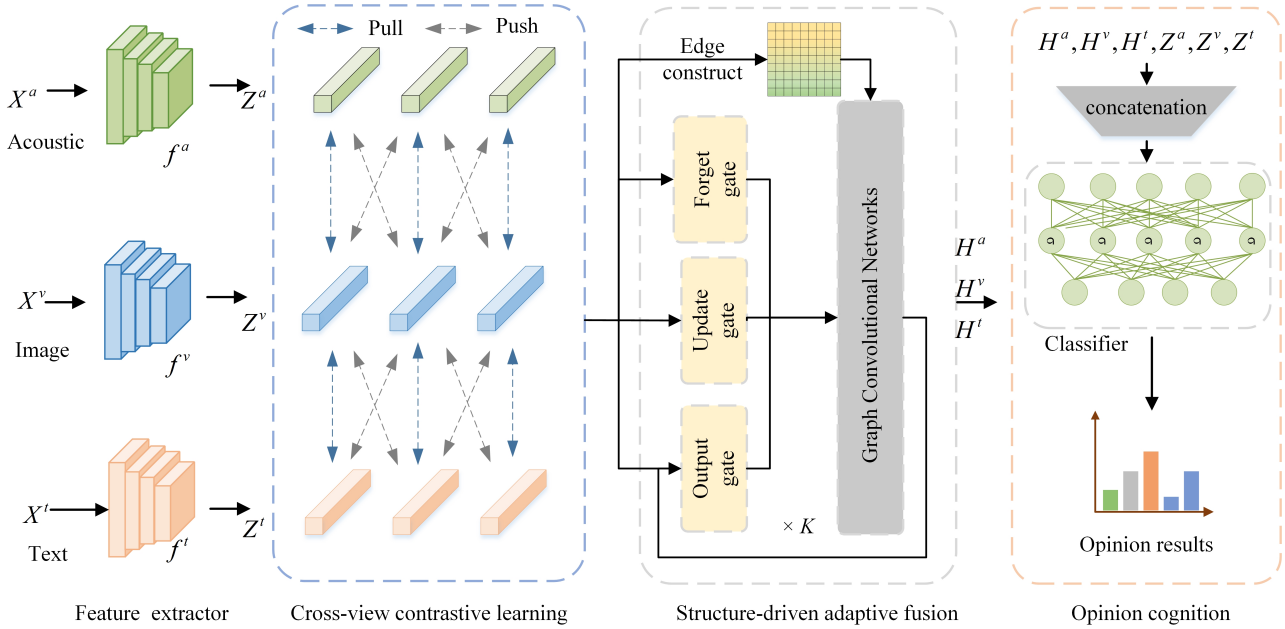


Fig. 1. The overall architecture of SMOM. Given a social media conversation X with the acoustic view X^a, the image view X^v and text view X^t, SMOM utilizes view-specific feature extractors to learn the corresponding representations Z^a, Z^v, and Z^t, respectively. Then, SMOM performs the cross-view contrastive learning between view-specific representations. Meanwhile, SMOM utilizes the structure-driven adaptive fusion function to obtain the corresponding fusion representations H^a, H^v, and H^t, respectively. Finally, SMOM leverages the classifier to obtain the opinions of each user.

2.1. View-specific feature extractor

In a multi-view context, different views carry their own unique structures and characteristics. For example, the text view typically contains rich semantic information, conveying deep language meanings, emotional tendencies, and contextual relationships, which help in understanding the background or context of an event. On the other hand, the image view primarily includes spatial information, such as objects, shapes, colors, and their spatial relationships within an image, which is crucial for scene or action recognition and analysis. The information from these different views is often complementary, meaning that understanding a complex situation solely from one view is often insufficient. In this context, designing modality-specific feature extractors becomes an essential strategy to fully mine and leverage this complementary information. These extractors are tailored to model the unique data structures and informational characteristics of each view, allowing each view to learn the most representative private representations. By doing so, the information from different views is extracted independently and in a targeted manner, avoiding interference and excessive blending between modalities, while retaining the key features of each view. This process of learning private representations helps in more accurately capturing the inherent structure and potential relationships of each view.

Given the sample $x_i = \{x_i^a, x_i^v, x_i^t\}$, the view-specific representations are obtained via the view-specific feature extractors:

$$\begin{aligned} s_i^a &= f^a(x_i^a) \\ s_i^v &= f^v(x_i^v) \\ s_i^t &= f^t(x_i^t) \end{aligned} \tag{1}$$

where $f^a, f^v, and f^t$ denote the view-specific feature extractors, respectively. $s_i^a, s_i^v, and s_i^t$ denote the view-specific representations of samples. Meanwhile, the view-specific representations of users are obtained via the view-specific feature extractors:

$$\begin{aligned} \bar{s}_i^a &= f^a(u_i^a) \\ \bar{s}_i^v &= f^v(u_i^v) \\ \bar{s}_i^t &= f^t(u_i^t) \end{aligned} \tag{2}$$

where u_i denotes the user belonging to the i-th sample. Then, a fusion operation is utilized to aggregate the sample information and user information within each view,

$$z_i^\gamma = s_i^\gamma + \eta \bar{s}_i^\gamma, \gamma \in a, v, t \tag{3}$$

where η denotes aggregation degree of user information.

2.2. Cross-view contrastive learning

Due to differences in structure, content, and the way information is presented across modalities, directly combining their representations may lead to information loss or inaccuracies. Therefore, Cross-view contrastive alignment is devised to establish a shared representation space where the private features of each view are consistent and complementary. This alignment helps enhance the performance of multi-view learning tasks by ensuring that the unique characteristics of each view are effectively captured and leveraged in the shared space.

Specifically, the cross-view contrastive alignment directly maximizes the mutual information between view-specifically representations views:

$$L_{cc} = -\sum_{i}^{N} \left(I\left(Z_i^{\gamma_1}, Z_i^{\gamma_2}\right) + \left(H\left(Z_i^{\gamma_1}\right) + H\left(Z_i^{\gamma_2}\right)\right)\right). \quad (4)$$

where γ_1 and γ_2 belong to $\{a, v, t\}$ and $\gamma_1 \neq \gamma_2$. $I\left(Z_i^{\gamma_1}, Z_i^{\gamma_2}\right)$ and H denote the mutual information and information entropy.

To solve $I\left(Z_i^{\gamma_1}, Z_i^{\gamma_2}\right)$, the joint probability distribution of Z^{γ_1} and Z^{γ_2} are formulated as follows:

$$\mathcal{P}(Z^{\gamma_1}, Z^{\gamma_2}) = \frac{1}{N} \sum_{i=1}^{N} Z_i^{\gamma_1} \left(Z_i^{\gamma_2}\right)^{\top} \quad (5)$$

For discrete distributions, the mutual information and information entropy are given as below:

$$I(Z^{\gamma_1}, Z^{\gamma_2}) = \sum_{d=1}^{D} \sum_{d'=1}^{D} P_{dd'} \ln \frac{P_{dd'}}{P_d \cdot P_{d'}} \quad (6)$$

$$\begin{aligned} H(Z^{\gamma_1}) &= -\sum_{d=1}^{D} P_d \ln P_d \\ H(Z^{\gamma_2}) &= -\sum_{d'=1}^{D} P_{d'} \ln P_d \end{aligned} \quad (7)$$

where P_d and $P_{d'}$ denote the marginal probability distributions $\mathcal{P}(z^{\gamma_1} = d)$ and $\mathcal{P}(z^{\gamma_2} = d')$, which could be obtained by summing over the d-th row and d'-th column of P, respectively. Based on above discussion, the final form of the cross-view contrastive learning is as below:

$$\begin{aligned} L_{cc} &= -\sum_{i}^{N} \left(I\left(Z_i^{\gamma_1}, Z_i^{\gamma_2}\right) + \left(H\left(Z_i^{\gamma_1}\right) + H\left(Z_i^{\gamma_2}\right)\right)\right) \\ &= -\left(\sum_{d=1}^{D} \sum_{d'=1}^{D} P_{dd'} \ln \frac{P_{dd'}}{P_d \cdot P_{d'}} \right. \\ &\quad \left. - \left(\sum_{d=1}^{D} P_d \ln P_d + \sum_{d'=1}^{D} P_{d'} \ln P_{d'} \right) \right) \\ &= -\left(\sum_{d=1}^{D} \sum_{d'=1}^{D} P_{dd'} \ln \frac{P_{dd'}}{P_d \cdot P_{d'}} \right. \\ &\quad \left. + \left(\sum_{d=1}^{D} \sum_{d'=1}^{D} P_{dd'} \ln \frac{1}{P_d} + \sum_{d'=1}^{D} \sum_{d=1}^{D} P_{dd'} \ln \frac{1}{P_{d'}} \right) \right) \\ &= -\sum_{d=1}^{D} \sum_{d'=1}^{D} P_{dd'} \left(\ln \frac{P_{dd'}}{P_d \cdot P_{d'}} + \left(\ln \frac{1}{P_d} + \ln \frac{1}{P_{d'}} \right) \right) \\ &= -\sum_{d=1}^{D} \sum_{d'=1}^{D} P_{dd'} \ln \frac{P_{dd'}}{P_d \cdot P_{d'}}. \end{aligned} \quad (8)$$

In summary, the cross-view contrastive learning framework provides a powerful method for learning modality-aligned representations by leveraging mutual information and entropy, ensuring that the private features of each modality are captured and aligned in a shared space to enhance the performance of multi-view learning tasks.

2.3. Structure-driven adaptive fusion

The purpose of the structure-driven adaptive fusion strategy is to dynamically model the contextual information of multi-view conversations, fully capturing the complementarity between views while reducing redundant information, thereby enhancing the model's understanding of the social network context. In the multi-view social network sentiment recognition task, different modalities such as text, visual, and audio provide unique information, and the dynamic variation of this information in semantic space can be crucial for sentiment recognition. Specifically, the dynamic fusion module introduces Graph Convolutional Networks and gating mechanisms to combine inter-view and intra-view contextual information, dynamically aggregating multi-view features. This approach not only preserves view-specific information but also effectively learns the complex interactions between views, thus enhancing view complementarity in the shared semantic space while avoiding the accumulation of redundant information. Ultimately, this fusion method dynamically updates and filters information across multiple semantic layers, enabling the model to better extract the essence of multiview data, thereby improving performance in social network sentiment recognition tasks. This design offers significant advantages in balancing information flow and integrating view features,

particularly when deeply modeling the context and sentiment of social network discussions, demonstrating outstanding performance.

Graph construction: An undirected graph is conducted to express structure correlation between samples, i.e., $G = (V, E)$, where V denotes $3N$ node sets and E denotes the edge set. There exist two rules to model the edge. (1) Any two nodes belonging to the same view in the same sample are connected. (2) Nodes belonging to the same samples but from different views are connected.

$$A_{ij} = 1 - \frac{\arccos(\frac{x_i \cdot x_j}{\|x_i\|\|x_j\|})}{\pi}, \tag{9}$$

where A_{ij} is edge weight between the i-th node and the j-th node node.

Adaptive fusion: A gating strategy based on the graph convolution operation is proposed to fuse multi-view context information in the social media network. The fusion process is defined as:

$$\begin{aligned} \Gamma_u^{(k)} &= \sigma(W_u^g \cdot [g^{(k-1)}, H'^{(k-1)}] + b_u^g) \\ \Gamma_f^{(k)} &= \sigma(W_f^g \cdot [g^{(k-1)}, H'^{(k-1)}] + b_f^g) \\ \Gamma_o^{(k)} &= \sigma(W_o^g \cdot [g^{(k-1)}, H'^{(k-1)}] + b_o^g) \end{aligned} \tag{10}$$

where $\Gamma_u^{(k)}$ controls the flow of relevant information, $\Gamma_f^{(k)}$ decides which part of the previous memory should be discarded, and $\Gamma_o^{(k)}$ determines what information is passed to the next layer. The operations are based on a gating mechanism, where $g^{(k-1)}$ is the previous gate output.$H'^{(k-1)}$ is the memory from the previous layer. W_Γ^g, b_Γ^g are the learnable parameters. Then, we compute the candidate memory $\tilde{C}^{(k)}$ for the current layer:

$$\tilde{C}^{(k)} = \tanh(W_C^g \cdot [g^{(k-1)}, H'^{(k-1)}] + b_C^g) \tag{11}$$

This candidate memory is then merged with the previous memory using the update and forget gates:

$$C^{(k)} = \Gamma_f^{(k)} \odot C^{(k-1)} + \Gamma_u^{(k)} \odot \tilde{C}^{(k)} \tag{12}$$

The forget gate $\Gamma_f^{(k)}$ controls how much of the previous memory $C^{(k-1)}$ is retained. The update gate $\Gamma_u^{(k)}$ determines how much of the new candidate memory $\tilde{C}^{(k)}$ is added.

$$g^{(k)} = \Gamma_o^{(k)} \odot \tanh(C^{(k)}) \tag{13}$$

This output $g^{(k)}$ will be used for the graph convolution operation.

Graph convolution with renormalization: After computing the gating mechanism and context memory, a graph convolution operation is applied to update the node representations. The operation involves a graph convolution matrix $\tilde{P}$, which is derived using a renormalization trick:

$$\tilde{P} = \tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} \tag{14}$$

where $\tilde{A}$ is the adjacency matrix of the graph, representing the relationships between nodes. $\tilde{D}$ is the degree matrix of the graph.

Now, we apply the graph convolution operation to aggregate information from neighboring nodes in the graph:

$$H^{(k)} = \text{ReLU}\big(((1-\alpha)\tilde{P}H^{(k-1)} + \alpha H^{(0)})((1-\beta_{k-1})I_n + \beta_{k-1}W^{(k-1)})\big) \tag{15}$$

whereα and β_k are hyperparameters controlling the balance between the previous layer's information and the new layer's information. $H^{(0)}$ is initialized with the multimodal input features Z^a, Z^v, Z^t. I_n is the identity matrix. Finally, the output of the k-th layer is computed as:

$$H'^{(k)} = H^{(k)} + g^{(k)} \tag{16}$$

This output $H'^{(k)}$ is used in subsequent layers or for classification tasks.

2.4. Opinion cognition

After passing through the stack of K layers, the model generates refined representations of the three views (acoustic, image, and text) for each sample i. These refined representations are denoted as h_i^a, h_i^v, and h_i^t, respectively. These refined features capture both view-specific information as well as inter-view interactions learned through the deep fusion process. Once the multi-view representations are computed, they are concatenated and passed into a classifier to predict the emotion of each user. The classifier is formulated as follows:

$$\hat{y}_i = \text{Softmax}(W_c[z_i^a; z_i^v; z_i^t; h_i^a; h_i^v; h_i^t] + b_c) \tag{17}$$

where W_c is the weight matrix for the final classification layer. b_c is the bias term. $[\cdot]$ represents the concatenation. The Softmax output, $\hat{y}_i$, represents the predicted probabilities for each emotion class for the i-th sample. The classifier ultimately assigns the class with the highest probability as the predicted opinion for that sample.

To train the model, the cross-entropy loss is minimized via measuring the difference between the true labels and the predicted probabilities. The total training loss $\mathcal{L}$ is given by:

$$L_{ce} = -y \log \hat{y} + \|\Theta\|_2 \tag{18}$$

The cross-entropy term penalizes incorrect predictions by applying a logarithmic penalty proportional to the predicted probability for the true class. The L2-regularization

Table 1. Comparison results between SMOM and baselines on the IEMOCAP and MELD datasets.

	IEMOCAP				MELD			
Method	ACC	Pre	Rel	F1	ACC	Pre	Rel	F1
GCNet	0.4582	0.4545	0.4715	0.4787	0.3771	0.3709	0.3943	0.4043
CLF	0.5246	0.5511	0.5535	0.5195	0.4466	0.4612	0.4714	0.4555
SCM	0.6214	0.6222	0.6323	0.6355	0.5487	0.5398	0.5719	0.5369
MGL	0.6474	0.6478	0.6551	0.6599	0.5696	0.5367	0.5914	0.5455
Panosent	0.6566	0.6500	0.6633	0.6658	0.5812	0.5262	0.5854	0.5465
SMOM	**0.6770**	**0.6811**	**0.6747**	**0.6747**	**0.5931**	**0.5494**	**0.5931**	**0.5474**

term helps maintain small weights, preventing the model from overfitting to noise in the training data.

The final loss is defined as follows:

$$L = L_{ce} + wL_{cc} \tag{19}$$

where w denotes the balance parameter between losses.

3. Results and discussion

3.1. Setup

Dataset and Metric: In the experiments, to validate the performance of the SMOM model for opinion monitoring, two multi-view opinion monitoring datasets (IEMOCAP and MELD) are utilized: IEMOCAP is a video dataset containing dyadic conversations involving 10 unique social users. It comprises a total of 7,433 samples, each annotated with one of six emotion labels. The dataset captures rich multimodal information, including speech, text, and visual cues, making it an ideal benchmark for multi-view opinion monitoring tasks. MELD is a video dataset derived from multi-party conversations featuring 304 unique social users. It includes a total of 13,708 samples, each annotated with one of five emotion labels. The dataset is sourced from the Friends TV series and contains conversations where two or more participants interact, offering a realistic and complex setting for analyzing opinion dynamics across multiple views. These datasets provide a diverse and challenging environment to assess the effectiveness of the SMOM model in detecting and monitoring opinions with varying emotional contexts and multi-modal inputs. Meanwhile, accuracy (ACC), precision (Pre), recall (Rel), and f1-score (F1) are used to assess the performance of the SMOM [14].

Implementation Details: The experimental setup includes Python 3.6.10 as the programming environment, with PyTorch 1.4.0 as the primary deep learning framework. PyTorch-Geometric 1.4.3 is used for graph-based computations, while Torch-Scatter 2.0.4 supports efficient tensor operations. Scikit-learn 0.21.2 is employed for additional machine learning tasks, and CUDA 10.1 enables accelerated computations on compatible NVIDIA GPUs, ensuring efficient training and inference.

3.2. Comparison with baselines

Comparison baselines: Five deep multi-view opinion monitoring methods are compared on two datasets about four metrics, to demonstrate SMOM performance, containing GCNet [1], CLF [2], SCM [6], MGL [7], and Panosent [8].

Comparison results: Table 1 and 2 present the performance comparison between the SMOM method and several baseline methods on the IEMOCAP and MELD datasets. On the MELD dataset, SMOM achieves an accuracy of 0.6088, which is 2.76% higher than Panosent (0.5812). Its precision, recall, and F1 score reach 0.5669, 0.6088, and 0.5828, respectively, also surpassing all other methods. These results indicate that SMOM demonstrates not only superior accuracy but also strong overall performance across other evaluation metrics.

The superiority of SMOM can be attributed to three main reasons. First, SMOM employs view-specific feature extractors designed to independently capture the unique characteristics of text, image, and audio modalities. For example, the text modality extracts rich semantic and contextual information, the image modality captures spatial structures and visual cues, and the audio modality provides speech emotion and rhythm information. Additionally, by incorporating a user-specific adaptive aggregation mechanism, the model effectively integrates interactions between samples and users, enhancing feature diversity and expressiveness while maintaining relevance. Second, SMOM leverages a cross-view contrastive learning strategy, which maximizes mutual information across different modalities to ensure feature consistency in the shared space while preserving the unique properties of each modality. This strategy, optimized through information entropy and mutual information, prevents ambiguity during the fusion process and enhances the complementarity of modalities, thereby improving feature discrimination and representation capabilities. Finally, the structure-driven dynamic adaptive fusion strategy is a critical factor in SMOM's per-

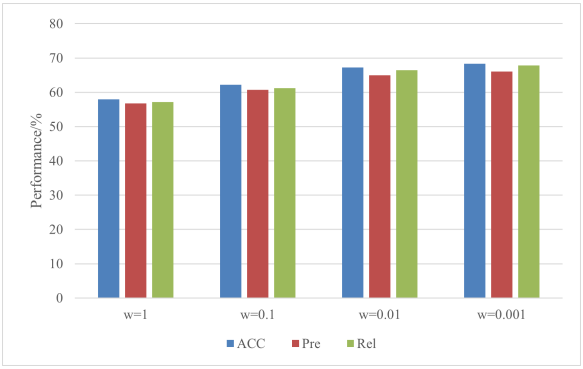
Table 2. Comparison results between SMOM and baselines on seven classes of the IEMOCAP dataset about ACC.

Method	Happy	Sad	Neutral	Angry	Excited	Frustrated
GCNet	0.3355	0.6182	0.4215	0.3956	0.5263	0.4558
CLF	0.3483	0.6112	0.4162	0.3974	0.5288	0.4577
SCM	0.4214	0.7222	0.5323	0.6355	0.7311	0.6065
MGL	0.4311	0.7829	0.5554	0.6346	0.7766	0.6479
Panosent	0.4649	0.8266	0.5851	0.6497	0.7500	0.6477
SMOM	**0.4773**	**0.8595**	**0.6017**	**0.6667**	**0.7807**	**0.6614**

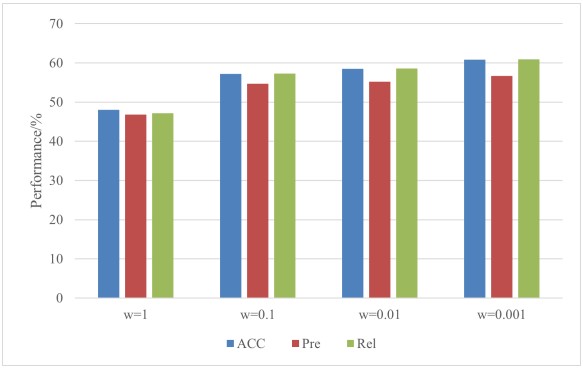
formance improvement. This strategy uses graph convolutional networks to construct an interaction graph between samples and modalities, dynamically modeling the complex relationships between modality features while employing gating mechanisms to finely control the updating and filtering of information. Specifically, the graph convolutional network captures high-order semantic relationships between samples through adjacency matrices and node update rules, while the gating mechanism dynamically selects key information via update, forget, and output gates, reducing the interference of redundant signals. This strategy minimizes noise while fully utilizing the complementary nature of multi-modal data, significantly enhancing the model's ability to capture emotion and semantics in complex social network contexts. Overall, SMOM achieves innovative designs in multi-modal information extraction, modality alignment, and dynamic context fusion, making it a standout performer in emotion analysis tasks, consistently outperforming other methods.

3.3. Parameter analysis

The parameter w plays a crucial role in balancing the cross-entropy loss and the view contrastive loss in SMOM, significantly impacting model performance across the two datasets. As shown in Fig. 2, the model's performance in terms of ACC, Pre, and Rel improves as w decreases from 1 to 0.001. For instance, for the IEMOCAP dataset, the impact of w is even more pronounced. At $w = 1$, the accuracy is 0.58, precision is 0.568, and recall is 0.572. By reducing w to 0.1, the accuracy improves to 0.622, with recall increasing to 0.612. The trend continues as w decreases further: at $w = 0.01$, accuracy increases to 0.672, and at $w = 0.001$, the model achieves its best performance, with accuracy reaching 0.683, precision 0.661, and recall 0.678. These results indicate that smaller values of w allow the view contrastive loss to play a more dominant role during training, leading to better alignment of multi-view features and improved overall performance. However, overly large values of w can cause the model to focus excessively on the cross-entropy loss, neglecting the benefits of cross-view feature alignment. Therefore, tuning w is essential for achieving optimal performance on different datasets.



(a) Parameter analysis on IEMOCAP



(b) Parameter analysis on MELD

Fig. 2. Parameter analysis of SMOM on two different datasets.

4. Conclusion

In this work, a novel deep multi-view contrastive fusion network (SMOM) is proposed for sentiment analysis in social media. SMOM independently extracts rich features from text, image, and acoustic modalities, bridges semantic gaps between views by maximizing mutual information, and dynamically models inter-view and intra-view interactions through graph convolutional networks and gating mechanisms. Overall, SMOM addresses limitations of static fusion strategies and captures complex relationships

among views for multi-view social media opinion monitoring. However, challenges remain in handling multi-modal data with varying characteristics and ensuring robustness across diverse platforms. In future work, we plan to further enhance the adaptability and efficiency of SMOM. For example, we will explore its application in real-time public opinion monitoring during major social events, where the model can analyze multi-modal social media data to promptly identify emerging public concerns and trends. Additionally, we will investigate the use of SMOM in brand sentiment analysis for businesses, allowing them to monitor social media discussions about their products or services and adjust marketing strategies accordingly.

5. Acknowledgment

This work was supported by the Natural Science Foundation of Gansu Province of China (21JR7RE174), 2024 Research Project on Educational and Teaching Reform at the School Level of Tianshui Normal University (JGG-24241435).

References

[1] Z. Lian, L. Chen, L. Sun, B. Liu, and J. Tao, (2023) *"Gcnet: Graph completion network for incomplete multimodal learning in conversation"* **IEEE Transactions on pattern analysis and machine intelligence *45***(7): 8419–8432. DOI: 10.1109/TPAMI.2023.3234553.

[2] G. Tu, B. Liang, R. Mao, M. Yang, and R. Xu. "Context or knowledge is not always necessary: A contrastive learning framework for emotion recognition in conversations". In: *Findings of the association for computational linguistics: ACL 2023*. 2023, 14054–14067.

[3] J. Gao, P. Li, A. A. Laghari, G. Srivastava, T. R. Gadekallu, S. Abbas, and J. Zhang, (2024) *"Incomplete multiview clustering via semidiscrete optimal transport for multimedia data mining in IoT"* **ACM Transactions on Multimedia Computing, Communications and Applications *20***(6): 1–20. DOI: 10.1145/3625548.

[4] J. Gao, M. Liu, P. Li, A. A. Laghari, A. R. Javed, N. Victor, and T. R. Gadekallu, (2023) *"Deep Incomplete Multiview Clustering via Information Bottleneck for Pattern Mining of Data in Extreme-Environment IoT"* **IEEE Internet of Things Journal *11***(16): 26700–26712. DOI: 10.1109/JIOT.2023.3325272.

[5] J. Gao, M. Liu, P. Li, J. Zhang, and Z. Chen, (2024) *"Deep Multiview Adaptive Clustering With Semantic Invariance"* **IEEE Transactions on Neural Networks and Learning Systems *35***(9): 12965–12978. DOI: 10.1109/TNNLS.2023.3265699.

[6] H. Yang, X. Gao, J. Wu, T. Gan, N. Ding, F. Jiang, and L. Nie. "Self-adaptive context and modal-interaction modeling for multimodal emotion recognition". In: *Findings of the association for computational linguistics: ACL 2023*. 2023, 6267–6281.

[7] T. Meng, F. Zhang, Y. Shou, H. Shao, W. Ai, and K. Li, (2024) *"Masked graph learning with recurrent alignment for multimodal emotion recognition in conversation"* **IEEE/ACM Transactions on Audio, Speech, and Language Processing**: DOI: 10.1109/TASLP.2024.3434495.

[8] Y. Tewel, O. Kaduri, R. Gal, Y. Kasten, L. Wolf, G. Chechik, and Y. Atzmon, (2024) *"Training-free consistent text-to-image generation"* **ACM Transactions on Graphics (TOG) *43***(4): 1–18. DOI: 10.1145/3658157.

[9] M. Luo, H. Fei, B. Li, S. Wu, Q. Liu, S. Poria, E. Cambria, M.-L. Lee, and W. Hsu. "Panosent: A panoptic sextuple extraction benchmark for multimodal conversational aspect-based sentiment analysis". In: *Proceedings of the 32nd ACM International Conference on Multimedia*. 2024, 7667–7676. DOI: 10.1145/3664647.36807.

[10] S. Zou, X. Huang, and X. Shen. "Multimodal prompt transformer with hybrid contrastive learning for emotion recognition in conversation". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, 5994–6003. DOI: 10.1145/3581783.3611805.

[11] S. Liu, P. Gao, Y. Li, W. Fu, and W. Ding, (2023) *"Multimodal fusion network with complementarity and importance for emotion recognition"* **Information Sciences *619***: 679–694. DOI: 10.1016/j.ins.2022.11.076.

[12] Q. Cheng, K. Wen, and X. Gu, (2022) *"Vision-language matching for text-to-image synthesis via generative adversarial networks"* **IEEE Transactions on Multimedia *25***: 7062–7075. DOI: 10.1109/TMM.2022.3217384.

[13] K. Yang, H. Xu, and K. Gao. "Cm-bert: Cross-modal bert for text-audio sentiment analysis". In: *Proceedings of the 28th ACM international conference on multimedia*. 2020, 521–528. DOI: 10.1145/3394171.3413690.

[14] T. Yu, H. Gao, T.-E. Lin, M. Yang, Y. Wu, W. Ma, C. Wang, F. Huang, and Y. Li. "Speech-text pre-training for spoken dialog understanding with explicit cross-modal alignment". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, 7900–7913.