International Journal of Approximate Reasoning 187 (2025) 109536

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

A novel three-way based self-adaptive filtering model for sentiment analysis

Zhihui Zhang, Dun Liu [*], Rongping Shen

School of Economics and Management, Southwest Jiaotong University, Chengdu 610031, China

ARTICLE INFO

Keywords:
Three-way decision
Sentiment analysis
Feature representation
Information redundancy

ABSTRACT

In the era of social media and diverse communication platforms, understanding human emotion across various modalities has become a crucial challenge. While significant progress has been made in feature extraction and interaction techniques, several unresolved issues persist, particularly concerning the balance between these two aspects. A central question is whether all extracted features are of equal importance, or if some may contain redundant or noisy information that undermines effective modality interaction. To address these challenges, we propose a novel Three-Way Decision-Based Self-Adaptive Filtering Model (TWSAFM). Inspired by the three-way decision (TWD) theory, we introduce a self-adaptive filtering module that categorizes extracted modal features into three distinct domains: acceptable, rejectable, and reconsidering. This classification allows for separate processing of features, enabling the model to prioritize essential information while minimizing the impact of redundant and noisy data. Experimental validation on three benchmark datasets demonstrates that TWSAFM outperforms state-of-the-art methods in sentiment analysis tasks. Furthermore, training studies and parameter sensitivity analysis underscore the effectiveness of TWSAFM in efficiently filtering out irrelevant and noisy features, highlighting its robust contribution to enhancing feature interaction.

1. Introduction

With the rapid proliferation of user-generated content on social media platforms, sentiment analysis has evolved into an efficient tool for extracting and analyzing user emotions and opinions. Traditional sentiment analysis methods that are based on singular modality such as text [1], have faced challenges in analyzing nuanced emotional expressions, particularly in analyzing the sentiment embedded in non-textual forms such as voice tones or visual gestures. This limitation has prompted a shift exploration toward multimodal sentiment analysis, which integrates multiple modalities, including text, audio, and video, to achieve a more comprehensive understanding of sentiment [34].

While this field has been proven powerful in analyzing sentiment across modalities, it still faces a variety of challenges, which are mainly rooted in the inherent complexity of integrating different modalities. Rapid growth of multimodal user-generated content on social media has outpaced existing analytical methods, leading to redundant or noisy feature representations and opaque fusion steps. First, while multimodal data may carry information that is complementary and correlated to each other, there is still abundant uncorrelated or redundant noise information across different modalities, which may harm the accuracy and efficiency of the modality integration accordingly. Therefore, one of the challenges is managing the noise and redundancy in extracted multimodal features.

[*] Corresponding author.
E-mail addresses: zzh3442@163.com (Z. Zhang), newton83@163.com (D. Liu), shen_rongping@126.com (R. Shen).

https://doi.org/10.1016/j.ijar.2025.109536
Received 9 April 2025; Received in revised form 6 July 2025; Accepted 24 July 2025

Available online 5 August 2025

Some existing methods attempt to address this challenge by introducing a flexible constraint in the decoding stage to strengthen the relationships implicitly between multimodal feature extraction and feature interaction, so that task-relevant information remains [46]. However, only imposing constraints in decoded representation may not sufficiently eliminate the noise and redundancy. Furthermore, these methods may also lack explainability, as the reasons for discarding certain features are unclear.

In addition, another challenge in sentiment analysis is how to handle the heterogeneity across different modality representations in feature extraction. Different modalities may vary in characteristics such as data distributions, sampling rates, and data volumes, therefore synchronizing and fusing these representations is also a critical challenge [16]. To address this challenge, strategies such as hierarchical mutual information maximization [7] and cross-modal attention mechanisms [29] have been proposed to align modal representations. However, the alignment between different modalities in these strategies is implicit and relies on a large amount of training data.

Current fusion models (e.g., MulT, MISA) either treat all features equally, letting noise degrade downstream performance, or apply constraints only at decoding, sacrificing explainability. To address the aforementioned challenges, a three-way-based self-adaptive filtering model (TWSAFM) for multimodal sentiment analysis is proposed in this research. First, inspired by the three-way theory, a three-way self-adaptive filtering module is designed between modality embedding and interaction, which provides a more explainable and nuanced mechanism for filtering out redundancy and noise. In addition, a similarity-based attention mechanism is designed and incorporated during the fusion process, which alleviates the heterogeneity across modalities by aligning features through loss functions correlated to the task. In summary, the contributions of this research are outlined as follows.

- First, the three-way decision theory is incorporated into multimodal sentiment analysis for constructing a filtering module with explainability. By dividing the modality features into acceptable, rejectable, and reconsidering domains based on their estimated contribution. Then, features within each domain are processed differently so that features with high contribution can be prioritized and redundant features can be mitigated.
- Second, we additionally introduce a self-adaptive mechanism into the three-way filtering module. By introducing a self-adaptive mechanism, the parameters of domain division criteria are not intuitively specified, but serve as hyperparameters that can be learned adaptively in accordance with the task and datasets. Therefore, the generalization capability of the three-way filtering module is further enhanced.
- Third, a cross-modal attention mechanism with similarity constraints is proposed to handle the heterogeneity across different modalities. Specifically, a similarity constraint is introduced in the cross-modal attention mechanism, which enables the model to automatically identify and emphasize portions with analogous characteristics when processing information from different modalities, thereby enhancing the effectiveness of feature fusion.

The subsequent sections of this paper are organized as follows. In Section 2, we provide a comprehensive review of the existing literature on sentiment analysis and the three-way decision, highlighting the key advancements and the limitations of traditional approaches and why we introduce the three-way decision into our model. In Section 3, we introduce our proposed methodology, detailing the innovative triadic decision framework for feature selection and reconstruction. Section 4 presents the experimental setup, including the datasets utilized for evaluation, the performance metrics applied, and the specific implementation details of our model. In Section 5, we discuss the results of our experiments, providing insights into the model's performance across different datasets and tasks. Finally, in Section 6, we conclude the paper by summarizing our findings, discussing the implications of our work for future research, and suggesting potential directions for further exploration in the field of multimodal sentiment analysis.

2. Related work

With the rapid development of multimodal sentiment analysis, numerous approaches have been proposed to address a variety of challenges and thereby enhance overall performance. In addition, three-way decision theory has been widely applied in deep learning fields to improve algorithm explainability and performance. In the following, the existing literature related to multimodal sentiment analysis, three-way decision theory and three-way decision in sentiment analysis are reviewed respectively.

2.1. Sentiment analysis with deep learning approach

Over recent years, various sentiment analysis approaches have been proposed and achieved impressive performance [5]. Among these researches, addressing the challenges of information redundancy and modality alignment are two critical issues. For the challenge of information redundancy, the approaches can be generally classified into three categories: disentanglement, attention-driven fusion, and text-centric hierarchical frameworks. Disentanglement-based methods, such as Disentanglement Translation Network (DTN) and MISA, separate modality-common and modality-specific features to minimize redundancy while retaining discriminative properties [8,46]. In addition, the attention-driven approaches, such as Deep Multimodal Attentive Fusion (DMAF) and Efficient Multimodal Transformer with Dual-Level Feature Restoration (EMT-DLFR), employ selective attention and reconstruction mechanisms to filter out irrelevant information, thereby enhancing feature refinement during fusion [11,27]. Furthermore, text-centric hierarchical approaches, including Text-Centric Hierarchical Fusion Network (TCHFN), prioritize the text modality for its rich emotional information while aligning non-text modalities and filtering redundant features [10,32]. In the above approaches, redundant information is often treated uniformly across modalities, and the nuanced differences in their contributions to specific tasks do not get enough attention.

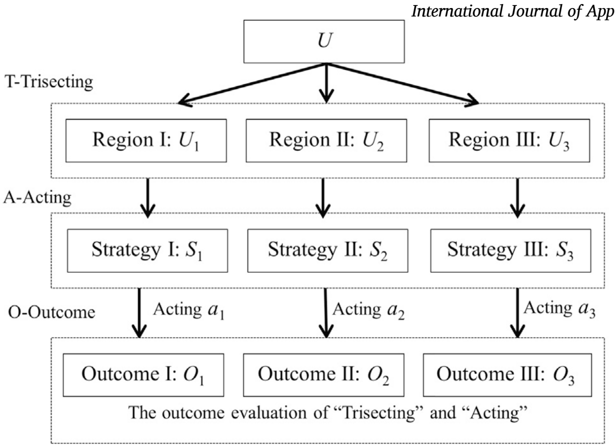


Fig. 1. TAO model of three-way decision.

For the challenge of modality alignment, most researches focus on correlating features from different modalities to improve model performance. Early methods used rigid word-level alignment, while the newer approaches such as the Multimodal Transformer (MulT) and CoolNet rely on soft alignment and cross-modal attention to dynamically adapt features across modalities [29,33]. Models like AcFormer further enhance alignment using contrastive learning to ensure consistency between modalities before fusion [48]. Despite the promising improvements these approaches have achieved in aligning different modalities, the performance of approaches in handling raw modalities and scaling to high-dimensional data remains to be explored.

2.2. Three-way decision in machine learning

The three-way decision (TWD), first proposed by Yao for interpreting rough set theory [37], is a decision-making theory that addresses uncertain information by introducing three types of decisions: acceptance, rejection, and deferment. The basic idea is to extend traditional binary classification into ternary classification by incorporating a deferral decision (i.e., making no decision) to reduce the risk of erroneous decisions [38]. For example, a doctor might assess a patient's health based on examination results, leading to three possible decisions: healthy, ill, or requiring further tests. Therefore, TWD is widely applied in scenarios where existing data is often insufficient and additional information is needed to form a decision.

The concept of TWD has been expanded in accordance with the artificial intelligence. Yao [39] expanded the concept of TWD by proposing the TAO model, which aligns with human cognition and thinking patterns. The TAO model, as shown in Fig. 1, structures the three-way decision process into three interconnected components: trisecting, acting, and outcome evaluation. Trisecting divides a whole into three meaningful parts to reduce complexity; acting devises strategies to process these parts through trisection-driven, action-driven, or iterative modes; and outcome evaluation measures the effectiveness of the combined effort. The model emphasizes the dynamic interplay between these components, offering flexibility, adaptability, and iterative refinement. It demonstrates the cognitive efficiency of "thinking in threes", making it a powerful tool for analyzing and solving complex problems.

Due to the capability of handling uncertainty, costs and risks during decision-making processes, the TAO model has been widely applied in the field of machine learning to improve model performance and explainability. For instance, Guo integrated this theory with deep learning models like encoder-decoder networks for portfolio management, allowing delayed decisions on uncertain stocks to reduce risks and optimize returns [6]. Min [23] introduced a tri-pattern discovery algorithm, demonstrating its potential to reduce data complexity and improve interpretability across domains like bioinformatics and text mining. In medical diagnostics, combining TWD with hybrid information systems and loss functions improves decision-making by resolving uncertainties in datasets [17]. It is also applied in ensemble oversampling for imbalanced keyword extraction, where TWD optimizes sampling strategies to enhance classification [18]. Additionally, sequential TWD models support dynamic decision-making in stock predictions and credit scoring, improving accuracy and reducing financial risks through multi-granularity and data fusion [25,36]. These applications highlight the adaptability of TWD in addressing uncertainty and boundary decisions across various domains.

2.3. Three-way decision in sentiment analysis

The application of TWD in sentiment analysis has demonstrated its effectiveness in handling uncertain and dynamic information across various domains. For instance, the dual-channel multimodal sentiment analysis framework utilizes TWD to mitigate inconsistencies in text and visual-audio modalities, improving sentiment prediction accuracy by introducing a third decision option that enhances flexibility in classification [31]. Similarly, the feature fusion approach for Chinese irony detection leverages TWD to address the imbalance and sparsity of features, enabling a two-stage classification process that reduces errors from low-confidence predictions [12]. In dynamic text sentiment classification, a temporal-spatial three-way granular computing framework dynamically updates sentiment classifications by adapting to evolving data and granularity, making it suitable for real-time applications [35]. Furthermore,

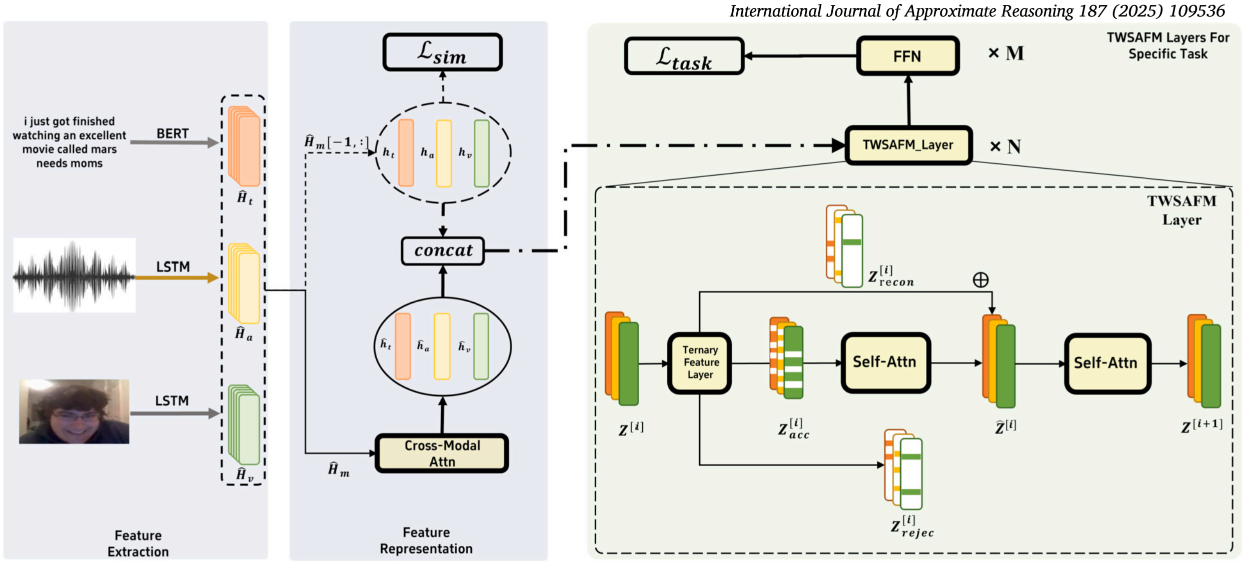


Fig. 2. Overall architecture of TWSAFM.

the analysis of drug reviews incorporates three-way decision into fusion models, integrating deep and traditional learning methods to handle low-confidence predictions effectively and boost accuracy [3].

Although there have been several researches on sentiment analysis and TWD, few research pay attention to leveraging TWD to manage the noise and redundancy in multimodal data. By doing so, it may hold the potential to improve the interpretability and efficiency of sentiment classification models in complex, real-world scenarios.

3. Methods

In this section, we propose a novel framework TWSAFM aimed at addressing the above-mentioned challenges in sentiment analysis, including information redundancy and modality heterogeneity. Specifically, the proposed approach integrates two key components: a TWD theory inspired self-adaptive filtering module, and a cross-modal attention module for efficient feature alignment. The self-adaptive filtering module categorizes fused modality features into acceptable, rejectable, and reconsidering groups, thereby allowing for adaptive filtering and prioritization of relevant information. Meanwhile, the cross-modal attention module aligns heterogeneous data streams by imposing similarity constraints, thereby facilitating seamless multimodal interaction. In-depth descriptions of TWSAFM are now provided, including the model's architecture, feature extraction process, fusion strategies and filtering techniques, and we comprehensive investigate the contributions of different modules and parameters.

3.1. *Task setup*

Each video segment in the benchmark dataset, consisting of a sequence of video frames, is associated with an overall emotional label. Based on this, we develop a model that leverages textual, acoustic, and visual signals from the video segments to identify emotional content. Features from these various modalities are extracted from each video segment to be used as inputs for the model.

Given the input sequences $X_m \in \mathbb{R}^{l_m \times d_m}$ for modality m, where l_m is the sequence length and d_m is the feature dimension of the modality, our goal is to create a unified representation that can be used for accurate sentiment prediction. The modalities considered in our model are text(t), acoustic(a) and visual(v).

3.2. *Overall architecture*

The overall architecture of the TWSAFM is illustrated in Fig. 2, which includes three main modules: multimodal feature extraction, multimodal feature representation, and TWSAFM module for adaptive feature filtering. First, in the stage of multimodal feature extraction, BERT and LSTM models are adapted to extract features from different raw modalities respectively. Then, in the stage of multimodal feature representation, we first impose constraints on the extracted multimodal features to identify common features across different modalities. Additionally, a cross-modal attention mechanism is introduced in this stage, where the interaction effect across different modalities on the target modality, based on enhanced information, is leveraged for a more precise feature representation. In the third stage, the multimodal features are further filtered within the TWSAFM module, where the feature filtering is conducted adaptively, and then applied to the final task. The overall architecture of the proposed TWSAFM is described above, and the details of three modules are now introduced.

3.3. Feature extraction

In the feature extraction stage, specialized models tailored for each modality are used to extract multimodal features respectively.

- **Textual Modality:** To ensure a fair comparison with prior research, we utilized the same BERT model [14] as in [7,15,26,27] for processing textual information. Specifically, we use the head information $\boldsymbol{h_t}$ from the final layer as the feature for the entire sentence. The remaining intermediate variables $\hat{H}_t$ that are also utilized as intermediate features for extracting the final features.
- **Acoustic & Visual Modality:** We used LSTM [9] to extract features from these two inputs [41], obtaining the corresponding final layer features $\boldsymbol{h_a}, \boldsymbol{h_v}$ and intermediate variables $\hat{H}_a$ and $\hat{H}_v$.

$$
\begin{aligned}
&\boldsymbol{h}_t, \hat{H}_t = BERT\{X_t; \theta_t^{BERT}\}, \boldsymbol{h_t} \in \mathbb{R}_t^d, \hat{H}_t \in \mathbb{R}^{l_t \times d_t}, \\
&\boldsymbol{h_m}, \hat{H}_m = LSTM\{X_m; \theta_m^{LSTM}\}, \boldsymbol{h_m} \in \mathbb{R}_m^d, \hat{H}_m \in \mathbb{R}^{l_m \times d_m}, m \in \{a, v\}.
\end{aligned}
\tag{1}
$$

Subsequently, to extract common features across different modalities which are relevant to the task, similarity constraints are imposed each modality features in the final layer. While this strategy has been applied in extracting common representations [47], we aim to enhance it by applying constraints and utilizing the intermediate variables that capture similar features. This combination with cross-modal attention helps in understanding how each modal extracts common representations within their feature space.

3.4. Cross-modal attention

In our model, the cross-modal attention mechanism is employed to capture the interactions between features from different modalities, thereby generating more representative multimodal features. Inspired by the MulT model [29], we design a cross-modal attention-based module to handle unaligned multimodal time series data. The cross-modal attention mechanism can be implemented through the following steps.

First, we employ the temporal convolution and employ positional embedding to make input sequences have sufficient awareness of its neighborhood elements and the same length. Then we get $H'_m \in \mathbb{R}^{l_m \times d}, m \in \{t, a, v\}$. We take the enhancement of textual information with audio and visual data as an example to describe the role of the cross-modal attention mechanism. For simplicity, we define:

$$
\begin{aligned}
&Q_t = H'_t W_{Q_t}, Q_t \in \mathbb{R}^{l_t \times d_k}, \\
&K_m = H'_m W_{K_m}, K_m \in \mathbb{R}^{l_m \times d_k}, \\
&V_m = H'_m W_{V_m}, V_m \in \mathbb{R}^{l_m \times d_k}, m \in \{a, v\}.
\end{aligned}
\tag{2}
$$

The single cross-modal attention from audio to text $\hat{\boldsymbol{h}}_{(\boldsymbol{a} \to \boldsymbol{t})}$ and visual to text $\hat{\boldsymbol{h}}_{(\boldsymbol{v} \to \boldsymbol{t})}$ is computed as:

$$
\hat{h}^{[0]}_{m \to t} = softmax(\frac{Q_t K_m}{\sqrt{d_k}}) V_m, m \in \{a, v\}.
\tag{3}
$$

After l blocks, we can get $\hat{\boldsymbol{h}}_{\boldsymbol{t}}^{[\boldsymbol{l}]} = Concat(\hat{h}^{[l]}_{a \to t} : \hat{h}^{[l]}_{v \to t})$. Similarly can get the $\hat{\boldsymbol{h}}_{\boldsymbol{a}}^{[\boldsymbol{l}]}$ and $\hat{\boldsymbol{h}}_{\boldsymbol{v}}^{[\boldsymbol{l}]}$.

While cross-modal attention effectively captures inter-modal interactions and produces enhanced feature representations, these representations may still contain redundancy and noise. Filtering and prioritizing these features is therefore critical before final sentiment prediction. To address this, we introduce the Three-Way Self-Adaptive Filtering Module (TWSAFM). TWSAFM employs a three-way decision-based self-adaptive approach to prioritize salient features and mitigate irrelevant information.

3.5. Three-way decision based self-adaptive filtering layers

After feature extraction and representation, we obtained feature representations $\boldsymbol{h_m} \in \mathbb{R}^d$ as well as enhanced information for each modality $\hat{\boldsymbol{h}}_{\boldsymbol{m}} \in \mathbb{R}^d$ derived from the others. We concatenate these to form three vectors $\boldsymbol{z_m} \in \mathbb{R}^{2d}$ for use in subsequent tasks. Then we concatenate them and get the all-modal feature $\mathbf{Z} \in \mathbb{R}^{3 \times 2d}$.

Before utilizing the feature in downstream tasks, we perform the feature filtering module inspired by TWD as showed in Fig. 3. Based on Fig. 3, the integration of TWD with the ternary feature model brings a semantic-oriented perspective to the classification of features. The ternary division derived from thresholds θ_1, θ_2 aligns with the three regions of the positive domain (R_1), boundary domain (R_2), and negative domain (R_3), which correspond to actions of acceptance, reconsideration, and rejection respectively. This mapping to semantic interpretations mirrors the practical applications of three-way decision models, where decisions are guided by nuanced evaluations. By categorizing features into these domains, we aim to dynamically allocate processing priorities. Features deemed significant (positive domain) are accepted for immediate processing, marginally relevant features (boundary domain) are reconsidered for future utility, while noise or irrelevant features (negative domain) are discarded. This structured approach enables the model to optimize feature relevance and apply tailored operations to mixed features, emphasizing their hierarchical importance.

Thus, we aspire for machines to filter information prior to processing, akin to human cognition. The three-way decision framework offers a fitting perspective by categorizing features into three classes: acceptable (important features), rejectable (useless features or noise), and reconsidering (currently unimportant but still containing useful information).

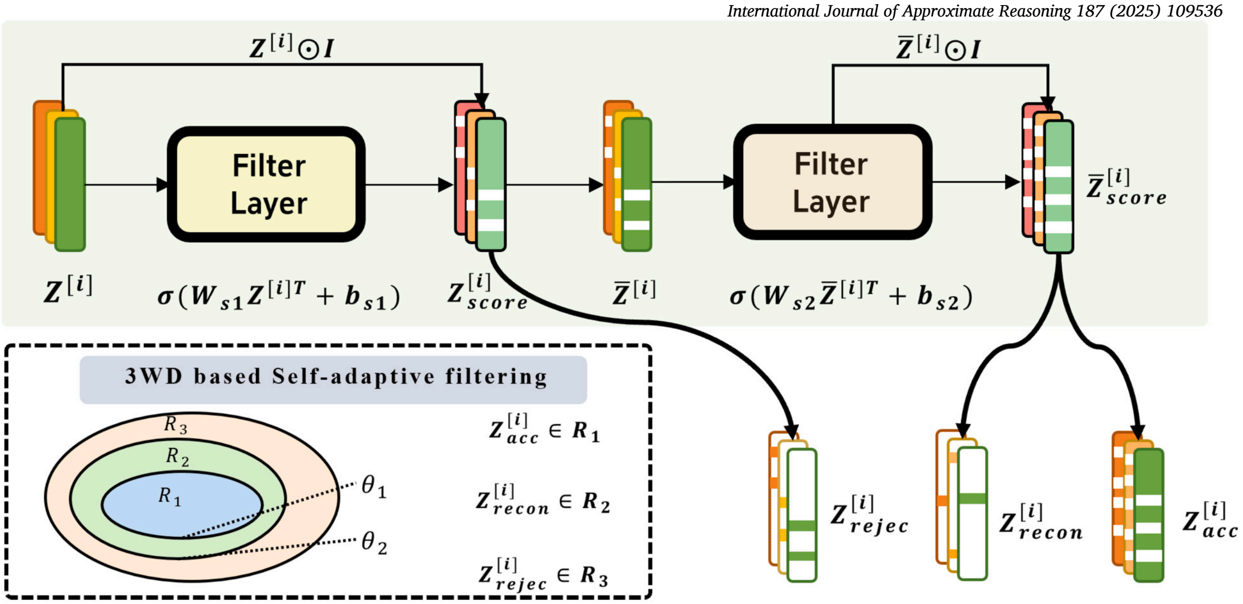


Fig. 3. Ternary feature layer.

3.5.1. *Ternary feature layer*

Applying this idea to machine learning, we categorize feature information into three classes: acceptable, rejectable, and reconsidering. This approach aligns perfectly with the concept of TWD. And the selection operation applies a filter **F** to the *i*th TWSAFM layer's input $\mathbf{Z}^{[\mathbf{i}]}$ such that:

$$\mathbf{Z}_{\mathbf{score}} = Score(\mathbf{Z}^{[\mathbf{i}]}) = \sigma(W_s \mathbf{Z}^{\top} + b_s). \tag{4}$$

Where $\sigma(\cdot)$ is the activate function, $W_s \in \mathbb{R}^{k \times 2d}$ and b_s are learnable parameters, and the k is the number of scoring units. This process requires matrix multiplication, followed by the activation function. The time complexity of this operation is $O(k \cdot d)$, which is relatively efficient and scales linearly with the size of the input feature vector.

The selection function **F** partitions the feature into three subsets based on learned thresholds θ_1 and θ_2.

$$\mathbf{F}(\mathbf{Z}_{\mathbf{score}}, \mathbf{Z}) = \begin{cases} \mathbf{Z}_{\mathbf{acc}} = \mathbf{Z} \odot I_{acc}, & if(\mathbf{Z}_{\mathbf{score}} \geq \theta_1) \\ \mathbf{Z}_{\mathbf{recon}} = \mathbf{Z} \odot I_{recon}, & if(\theta_1 > \mathbf{Z}_{\mathbf{score}} \geq \theta_2) \\ \mathbf{Z}_{\mathbf{reject}} = \mathbf{Z} \odot I_{reject}, & if(\mathbf{Z}_{\mathbf{score}} < \theta_2) \end{cases} \tag{5}$$

where I_{acc}, I_{recon} and I_{reject} are indicator matrices defined element-wise as: $[I_{acc}]_i = I([\mathbf{Z}_{\mathbf{score}}]_i \geq \theta_1), [I_{recon}]_i = I(\theta_1 > [\mathbf{Z}_{\mathbf{score}}]_i \geq \theta_2), [I_{reject}]_i = I([\mathbf{Z}_{\mathbf{score}}]_i < \theta_2)$ and $I(\cdot)$ is the indicator function:

$$I = \begin{cases} 1, & if\ condition\ is\ true \\ 0, & otherwise \end{cases} \tag{6}$$

This operation is element-wise, involving comparison and multiplication with indicator matrices. Its time complexity is $O(d)$, where d is the number of features. This operation is very efficient since it only requires element-wise operations, making it computationally inexpensive. Here, θ_1 and θ_2 are adaptive thresholds learned during training, and $\odot$ denotes element-wise multiplication. This mechanism allows the model to selectively focus on important features while attenuating or discarding less relevant ones.

Importantly, the thresholds θ_1 and θ_2 are not manually set but are implicitly learned through a series of linear transformations and non-linear activations. Given a hidden feature vector $h \in \mathbb{R}^d$, the model computes an intermediate score via $s = \tanh(Wh + b)$, where W and b are the parameters of the score layer. This score is further processed by a linear mask layer, followed by a layer normalization and a ReLU activation, yielding selection logits:

$$m = ReLU(LN(W_m \cdot s + b_m)). \tag{7}$$

A binary mask $M \in \{0,1\}^d$ is then constructed by applying a fixed threshold (zero in practice): entries with $m_i > 0$ are selected. This mask is first used to distinguish the accepted and reconsidered features $Z_{acc+recon}$ from the discarded ones (Z_{reject}). A second pass using the same mechanism is applied to $Z_{\text{acc+recon}}$ to further divide it into Z_{acc} and Z_{recon}, in accordance with Eq (5).

During training, the decision boundaries implied by θ_1 and θ_2 evolve automatically as the model updates the parameters W, b, W_m, and b_m through backpropagation. This enables the model to dynamically adjust its selection behavior based on gradient feedback from the ternary loss components (e.g., cosine similarity for Z_{acc}, penalty for Z_{reject}), allowing adaptive and data-driven partitioning of the feature space without the need for manual threshold tuning.

3.5.2. Feature selection and reconstruction

After feature selection, we can perform task-specific operations on the different types of features based on the requirements of the subsequent tasks. Naturally, features $\mathbf{Z}_{\mathbf{reject}}$ will be discarded to reduce noise. Feature selection helps to reduce the impact of irrelevant or noisy features, thereby decreasing the variance of gradients during model training. Among the remaining features, the more representative $\mathbf{Z}_{\mathbf{accept}}$ are directly fed into the attention mechanism for thorough mixing. The attention mechanism is defined as:

$$\begin{aligned}\mathbf{Q} = \mathbf{W}_{\mathbf{Q}}\mathbf{Z}_{\text{acc}}, \quad \mathbf{K} =& \mathbf{W}_{\mathbf{K}}\mathbf{Z}_{\text{acc}}, \quad \mathbf{V} = \mathbf{W}_{\mathbf{V}}\mathbf{Z}_{\text{acc}}, \\ \mathbf{A} =& \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right), \\ \mathbf{Z}' =& \mathbf{A}\mathbf{V},\end{aligned} \tag{8}$$

where $\mathbf{W}_{\mathbf{Q}}, \mathbf{W}_{\mathbf{K}}, \mathbf{W}_{\mathbf{V}} \in \mathbb{R}^{d_{\text{model}} \times 2d}$ are learnable projection matrices, and d_k is the dimensionality of the key vectors.

The $\mathbf{Z}_{\mathbf{recon}}$ features are then combined with the attention-processed acceptable features through a residual connection, aiming to activate this type of information at a higher dimensional level, we choose not to allow this subset of features to directly participate in the initial calculations for two primary reasons. First, the hierarchical nature of multimodal data fusion necessitates a clear separation of immediate and latent features to mitigate the risk of feature entanglement. Allowing $\mathbf{Z}_{\mathbf{recon}}$ features to directly influence initial processing could lead to premature optimization, where critical interdependencies among modalities are overlooked [28]. Moreover, $\mathbf{Z}_{\mathbf{recon}}$ features often represent nuanced or context-dependent information, which, while valuable, may introduce noise if processed prematurely. By deferring their integration, the model can fine-tune its foundational layers, thereby enabling the residual mechanism to highlight these subtle patterns more effectively [21].

$$\begin{aligned}\mathbf{Q}' = \mathbf{W}'_{\mathbf{Q}}(\mathbf{Z}'_{\text{acc}} + \mathbf{Z}_{\mathbf{recon}}), \quad \mathbf{K} =& \mathbf{W}_{\mathbf{K}}(\mathbf{Z}'_{\text{acc}} + \mathbf{Z}_{\mathbf{recon}}), \quad \mathbf{V} = \mathbf{W}_{\mathbf{V}}(\mathbf{Z}'_{\text{acc}} + \mathbf{Z}_{\mathbf{recon}}), \\ \mathbf{A}' =& \text{Softmax}\left(\frac{\mathbf{Q}'\mathbf{K}'^{\top}}{\sqrt{d_k}}\right), \\ \mathbf{Z}^{[\mathbf{i+1}]} =& \mathbf{A}'\mathbf{V}'.\end{aligned} \tag{9}$$

In Eq (9), the output $\mathbf{Z}^{[\mathbf{i+1}]}$ represents the ith layer of TWSAFM outputs, the $(i+1)$th layer of TWSAFM inputs. After several TWSAFM layers, we obtain fully mixed features. These features are then passed through Feed-Forward Network (FFN) layers to produce the final results, as illustrated in Fig. 2. Correspondingly, we can calculate the proportion of each feature category based on cosine similarity, which allows us to assign different computations to various classifications in subsequent task functions, enabling the machine to comprehend the significance of each operation.

3.6. Loss function

In our proposed model, the loss function must account not only for the standard task loss but also for the losses incurred during the feature selection and reconstruction processes. This ensures that the optimization objectives across all stages are aligned. Our proposed loss function includes the following components.

Similarity Loss $\mathcal{L}_{sim}$: This loss ensures that the high level representations extracted from different modalities are consistent. Given the final layer features $\boldsymbol{h}_t, \boldsymbol{h}_a$ and $\boldsymbol{h}_v$, we define the similarity loss as:

$$\mathcal{L}_{sim} = \sum_{m,n\in\{T,A,V\},m\neq n} (1 - cos(h_m, h_n)) \tag{10}$$

This term encourages the model to learn similar representations for the same input across different modalities, facilitating a more effective fusion of multimodal information.

Task Loss with TWSAFM $\mathcal{L}$: The TWSAFM model involves a tri-phase feature selection mechanism that categorizes features into three classes, and we use the $\alpha_{acc}, \alpha_{recon}$ and α_{dis} to denote the percentage of features relative to the original features.

$$\mathcal{L} = (\gamma_{acc}\alpha_{acc} + \gamma_{recon}\alpha_{recon} + \gamma_{dis}\alpha_{dis})\mathcal{L}_{task} \tag{11}$$

where γ_1, γ_2 and γ_3 are hyperparameters that balance the importance of acceptable, reconsidering, and rejectable features respectively. And the $\mathcal{L}_{task}$ can represent the loss function for any task, such as cross-entropy or MAE.

3.7. Three-way based method's impact for embedding features

In multimodal deep learning, particularly in tasks like sentiment analysis, existing methods often focus on two critical stages: feature embedding and modality interaction. Feature embedding is where different modalities, such as text and audio, are transformed into a shared representation space. However, it is important to note that not all embedded features are equally useful for the task at

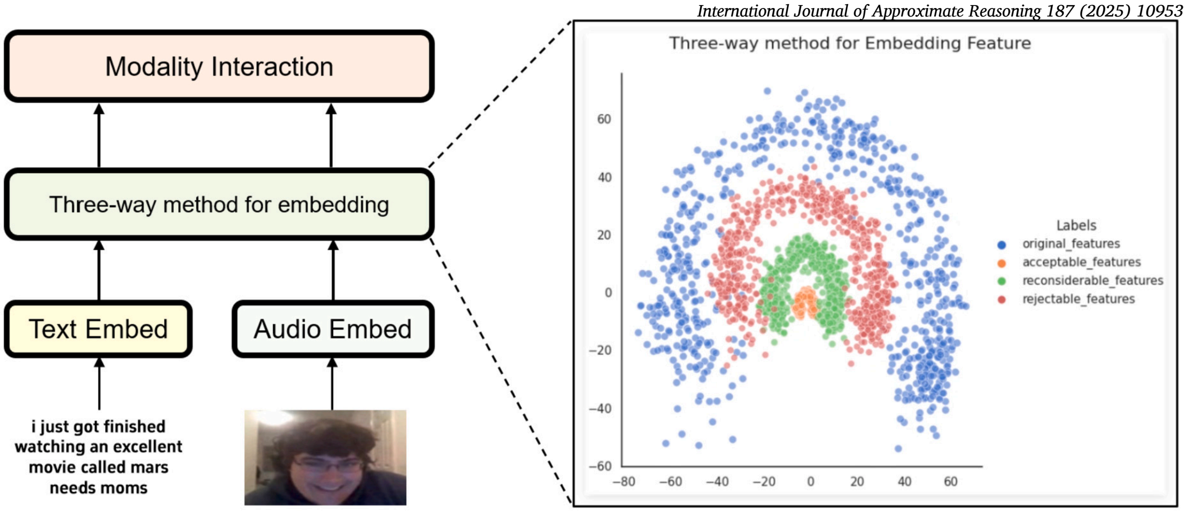


Fig. 4. Three-way based feature filtering for training data. The blue scattered "original features" are classified by a unified filtering module. By distinguishing these features, the model ensures that only the most essential and useful features are kept (the central orange features) while irrelevant ones are rejected or reconsidered, ultimately improving the model's efficiency and performance for modality interaction accordingly. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

hand. Many features may be irrelevant, redundant, or even harmful to the overall performance of the model. The model mentioned in this section, which shows a impressive impact for embedded features in Fig. 4. It shows that when we obtain the initial phase of original embedded features, these features are scattered across the space. After effective training, our TWSAFM appropriately categorizes the features, extracting the most relevant and important ones from each feature (orange). It discards the outermost features (red) and retains the reconsidering features (green) between them. Different types of features are then enhanced in the subsequent modality interaction phase with varying importance and operations. By adopting this approach, the model ensures that only the most relevant and impactful features are utilized, which helps reduce information redundancy and improves overall efficiency. Ultimately, the three-way decision method plays a crucial role in ensuring that features are used optimally, enhancing the performance and interpretability of multimodal models.

4. Experiment

In this section, we introduce our experimental settings including experimental datasets, baselines for comparison, evaluation metrics and the experimental environment.

4.1. Dataset

In our following discussions, three well-established sentiment analysis datasets based on videos: CMU-MOSI [44], CMU-MOSEI [45], CH-SIMS [40] are utilized to evaluate the effectiveness of the proposed model. These datasets involve tasks related to sentiment analysis and emotion detection. The detailed information of the datasets is described in combination with Table 1.

CMU-MOSI is one of the most widely used datasets in the field of multimodal sentiment analysis. It consists of video clips from YouTube, where speakers express their opinions on various movie themes through monologues. Each video clip is annotated with sentiment labels according to the Stanford Sentiment Treebank method. The sentiment scores range from -3 to 3, with -3 indicating strong negative sentiment and 3 indicating strong positive sentiment. The dataset is divided into training, validation, and test sets, containing 1,284, 229, and 686 utterance segments, respectively.

CMU-MOSEI dataset is currently the largest dataset for sentiment analysis and emotion recognition. It is an expanded version of the CMU-MOSI dataset, featuring a more diverse range of samples, speakers, and topics. It contains over 23,500 video segments of sentences from more than 1,000 YouTube speakers, with a balanced gender distribution. All sentence segments are randomly selected from monologue videos on various topics. These videos have been transcribed and properly punctuated for analysis.

CH-SIMS is a Chinese single and multimodal sentiment analysis dataset, that contains 2281 refined video clips with both multimodal and independent single-modal annotations. It allows researchers to study the interactions between modes, or to perform single-modal sentiment analysis using independent single-modal annotations. Each video segment is manually annotated with a sentiment intensity score defined from -1 (strongly negative) to 1 (strongly positive).

4.2. Baseline for comparison

To ensure a fair and comprehensive evaluation, the baseline algorithms are selected based on their relevance to sentiment analysis, popularity in related tasks, and effectiveness as reported in previous studies. These methods include both traditional and state-of-the-art models, enabling a thorough comparison across diverse approaches:

Table 1
The statistics for datasets.

Dataset	Train	Valid	Test	All
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16326	1871	4659	22856
CH-SIMS	1368	456	457	2281

TFN. Tensor Fusion Network [42] uses a new multi-modal fusion method to model inter-modal dynamics, enabling end-to-end learning of intra-modal and inter-modal dynamics.

LMF. The Low-rank Lultimodal Fusion method [19], which performs multimodal fusion using low-rank tensors to improve efficiency for sentiment analysis.

MFN. A neural architecture for multi-view sequential learning called the Memory Fusion Network (MFN) [43] explicitly accounts for both interactions in a neural architecture and continuously models them through time.

MFM. The Multimodal Factorization Model [30] learns representations that can be factorized into multimodal discriminative and modality-specific generative features.

MISA. This method factorizes modalities into Modality-Invariant and -Specific Representations (MISA) [8] using a combination of specially designed losses and then performs multimodal fusion on these representations.

MMIM. MultiModal InfoMax (MMIM) [7] proposes a hierarchical mutual information maximization framework to guide the model to learn shared representations from all modalities.

Bert-Mag. This method proposes an attachment to BERT and XLNet called Multimodal Adaptation Gate (MAG) [24]. MAG allows BERT and XLNet to accept multimodal nonverbal data during fine-tuning.

Self-MM. Self-Supervised Multi-task Multimodal (Self-MM) sentiment analysis network [41] designs a unimodal label generation module based on self-supervised learning to explore unimodal supervision.

MulT. Multimodal Transformer (MulT) [29] utilizes directional pairwise cross-modal attention to capture intermodal correlations in unaligned multimodal sequences.

4.3. *Evaluation metrics*

Sentiment intensity prediction is commonly used in the evaluation of sentiment analysis methods, which fundamentally is a regression problem, and evaluation metrics include the mean absolute error (MAE) and the Pearson correlation coefficient (Corr). In addition, continuous sentiment intensity scores are usually transformed into discrete categories for evaluating classification accuracy. Consistent with prior practice [8,27], we present seven-class accuracy (Acc-7), five-class accuracy (Acc-5), binary accuracy (Acc-2), and F1-score for both the CMU-MOSI and CMU-MOSEI datasets. It is worth mentioning that there are two distinct approaches exist for the binary classification task: negative/non-negative [45] and negative/positive [29]. Accordingly, we provide the Acc-2 and F1-score for each method, using a segmentation marker (-/-) to distinguish them, where the score on the left represents negative/non-negative, and the one on the right corresponds to negative/positive. For CH-SIMS, as outlined in [40], we report five-class accuracy (Acc-5), three-class accuracy (Acc-3), and binary accuracy (Acc-2). For all metrics except MAE, higher values signify better performance.

4.4. *Raw feature extraction*

To ensure a fair comparison, we utilize the standard low-level features provided by each benchmark, which are also employed by the current state-of-the-art methods.

Text Modality: With the advancement of technology, pre-trained models like BERT have increasingly become the go-to feature extractors for text segments. Consistent with recent studies [16,27], we leverage the pre-trained BERT model to encode raw textual data. Specifically, we use the bert-base-uncased model for the CMU-MOSI and CMU-MOSEI datasets, and the bert-base-chinese model for the CH-SIMS dataset.

Visual features: The CMU-MOSI and CMU-MOSEI datasets rely on Facet3 to extract 35 facial action units, capturing muscle movements associated with emotional expressions. For the CH-SIMS dataset, OpenFace 2.0 [2], a facial behavior analysis toolkit, is utilized to obtain 17 facial action units, 68 facial landmarks, and additional features related to head and eye movements.

Audio features: For the CMU-MOSI and CMU-MOSEI datasets, each dataset uses the COVAREP [4] framework to extract low-level audio statistics. These features include 12 Mel-frequency cepstral coefficients (MFCCs), pitch, voiced/unvoiced segments (VUV), and glottal source parameters, among others. For CH-SIMS, we use Librosa [22] to extract logarithmic fundamental frequency, 12 constant-Q chromatograms, and 20 MFCCs.

5. Results and analysis

5.1. *Experiment results on datasets*

The experiment results of the proposed TWSAFM model are shown in Tables 2 to 4. It is observed that the TWSAFM model demonstrates promising performance across the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets, which demonstrates its ability to handle information redundancy, and align heterogeneous modalities, therefore improving the model’s performance.

Table 2
Performance Comparison on CMU-MOSI.

Model	A2(↑)	A7↑	MAE↓	Corr↑	F1↑
TFN[a]	– / 77.1	–	0.870	0.700	– / 77.9
LMF[a]	– / 76.4	32.8	0.912	0.668	– / 75.7
MFN[a]	– / 77.4	34.1	0.965	0.632	– / 77.3
MFM[a]	– / 81.7	35.4	0.877	0.706	– / 81.6
MISA[a]	81.8 / 83.4	42.3	0.783	0.761	81.7 / 83.6
MMIM[a]	84.1 / 86.1	46.7	0.700	0.800	84.0 / 86.0
MulT[a]	– / 83.0	40.0	0.871	0.698	– / 82.8
TFN	77.6 / 79.0	34.7	0.958	0.662	77.5 / 79.0
LMF	75.4 / 76.1	33.1	0.971	0.640	75.4 / 76.1
MFN	78.9 / 81.0	38.8	0.874	0.705	80.2 / 81.1
MFM	77.7 / 79.0	35.7	0.929	0.665	77.7 / 79.0
MISA	82.4 / 83.8	42.1	0.742	0.778	82.4 / 83.9
MMIM	82.6 / 82.6	42.9	0.748	0.779	82.6 / 83.5
Bert-Mag	82.9 / 84.6	44.5	0.743	0.784	**82.9** / 84.6
SELF-MM	82.8 / 84.6	**46.7**	0.729	0.778	82.8 / 84.6
MulT	80.0 / 81.9	36.2	0.908	0.695	79.9 / 81.8
TWSAFM	**82.9** / **85.2**	44.9	**0.718**	**0.786**	82.8 / **85.3**

[a] Represents the model's performance is cited from the [15].

Table 3
Performance Comparison on CMU-MOSEI.

Model	A2(↑)	A7(↑)	MAE(↓)	Corr(↑)	F1(↑)
MulT[a]	– / 82.5	51.8	0.580	0.703	– / 82.3
MFM[a]	– / 84.4	51.3	0.568	0.717	– / 84.3
MISA[a]	83.6 / 85.5	52.2	0.555	0.756	83.8 / 85.3
MMIM[a]	82.2 / 86.0	54.2	0.526	0.772	82.7 / 85.9
TFN	82.2 / 82.3	51.2	0.572	0.722	82.1 / 82.3
LMF	81.5 / 84.2	52.5	0.564	0.736	81.7 / 84.0
MFN	79.8 / 83.3	52.0	0.569	0.719	80.2 / 83.2
MFM	81.9 / 84.1	50.7	0.577	0.727	82.2 / 84.0
MISA	82.1 / 84.2	52.0	0.553	0.757	80.3 / 84.2
MMIM	83.6 / 83.6	50.1	0.576	0.720	83.5 / 83.2
Bert-Mag	80.6 / **85.0**	53.2	0.542	0.764	81.2 / **85.1**
Self-MM	79.1 / 84.8	**53.3**	0.540	0.761	79.9 / 84.9
MulT	79.6 / 84.1	51.7	0.570	0.729	82.3 / 83.6
TWSAFM	**84.0** / 84.5	**53.3**	**0.539**	0.759	**83.9** / 84.9

[a] Represents the model's performance is cited from the [15].

On the CMU-MOSI dataset, TWSAFM achieves a binary classification accuracy (A2) of (82.9% / 85.2%) and an F1 score of (82.8% / 85.3%), outperforming models like MISA and MulT. The low MAE of 0.718 and a Pearson correlation coefficient (Corr) of 0.786 demonstrate its precision in capturing emotional nuances. The ternary feature selection effectively prioritizes relevant features and discards noise, ensuring reliable predictions.

On the larger and more diverse CMU-MOSEI dataset, TWSAFM obtains an A2 score of (84.0% / 84.5%), an F1 score of (83.9% / 84.9%), and a seven-class accuracy (A7) of 53.3%. The cross-modal attention mechanism aligns textual, acoustic, and visual features, enabling robust generalization and outperforming advanced models like MMIM and Self-MM. TWSAFM not only excels in classification but also reduces prediction errors, reinforcing its versatility in complex scenarios.

On the CH-SIMS dataset, TWSAFM adapts well to the linguistic and cultural nuances of Chinese sentiment analysis, achieving an A2 score of 79.4%, an A3 accuracy of 66.3%, and an A5 accuracy of 41.1%. Its low MAE of 0.424 and a Corr of 0.591 highlight its robustness and precision, effectively capturing subtle emotional cues across modalities.

Across all datasets, TWSAFM consistently outperforms state-of-the-art methods. Its ternary feature selection dynamically categorizes features into acceptable, rejectable, and reconsidering, while the cross-modal attention mechanism aligns and enhances heterogeneous data. These innovations position TWSAFM as a powerful model for multimodal sentiment analysis, capable of addressing long-standing challenges and paving the way for future advancements.

In addition to the performance metrics discussed, the computational efficiency of TWSAFM was evaluated through a statistical analysis of inference times for transformer-based models, as illustrated in Fig. 5. The results demonstrate that TWSAFM achieves a balanced trade-off between accuracy and computational cost, with inference times comparable to state-of-the-art models like Bert-Mag across the CMU-MOSI and CMU-MOSEI datasets. This efficiency stems from the self-adaptive filtering module, which reduces redundant feature processing, and the optimized cross-modal attention mechanism, enabling faster modality interaction without sacrificing performance. These findings underscore TWSAFM's suitability for real-time multimodal sentiment analysis applications.

Table 4
Performance Comparison on CH-SIMS.

Model	A2(↑)	A3(↑)	A5(↑)	MAE(↓)	Corr(↑)	F1(↑)
TFN[a]	77.1	–	–	0.437	0.582	76.9
LMF[a]	77.4	–	–	0.438	0.578	77.4
MulT[a]	78.6	–	–	0.453	0.564	79.7
MISA[a]	76.5	–	–	0.447	0.563	76.6
Self-MM[a]	80.0	–	–	0.425	0.595	80.4
TFN	76.4	**66.3**	41.6	0.430	0.584	76.8
LMF	78.1	65.4	38.5	0.436	0.590	78.0
MFN	77.2	66.1	38.5	0.435	0.586	77.2
MFM	71.1	62.6	32.4	0.478	0.514	72.2
MISA	75.3	62.1	37.2	0.454	0.549	74.9
MMIM	74.6	63.0	37.1	0.472	0.532	75.1
Bert-Mag	78.1	63.7	40.7	0.467	0.478	78.2
Self-MM	77.0	63.7	40.0	0.427	0.575	77.5
MulT	77.5	65.2	39.6	0.438	0.586	77.5
TWSAFM	**79.4**	**66.3**	**41.1**	**0.424**	**0.591**	**79.4**

[a] Represents the model's performance is cited from the [27].

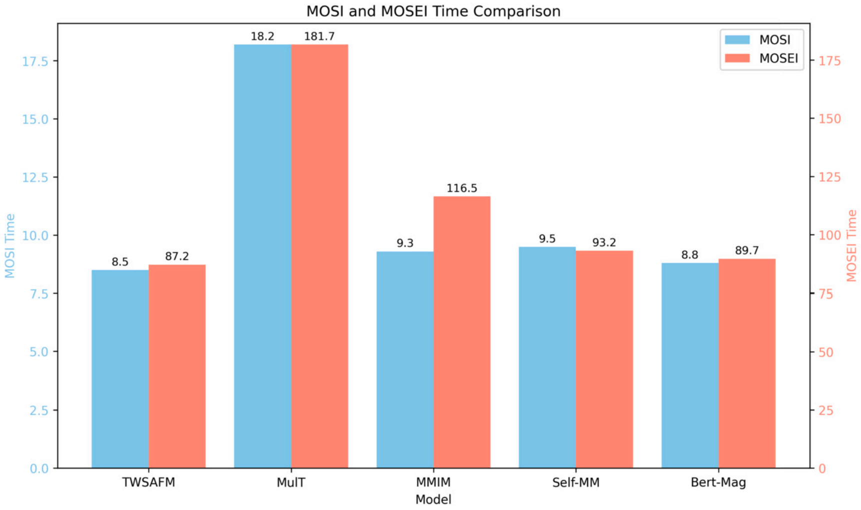


Fig. 5. Transformer-based Models' Time Comparison on Datasets.

5.2. *Parameter sensitivity and robustness analysis*

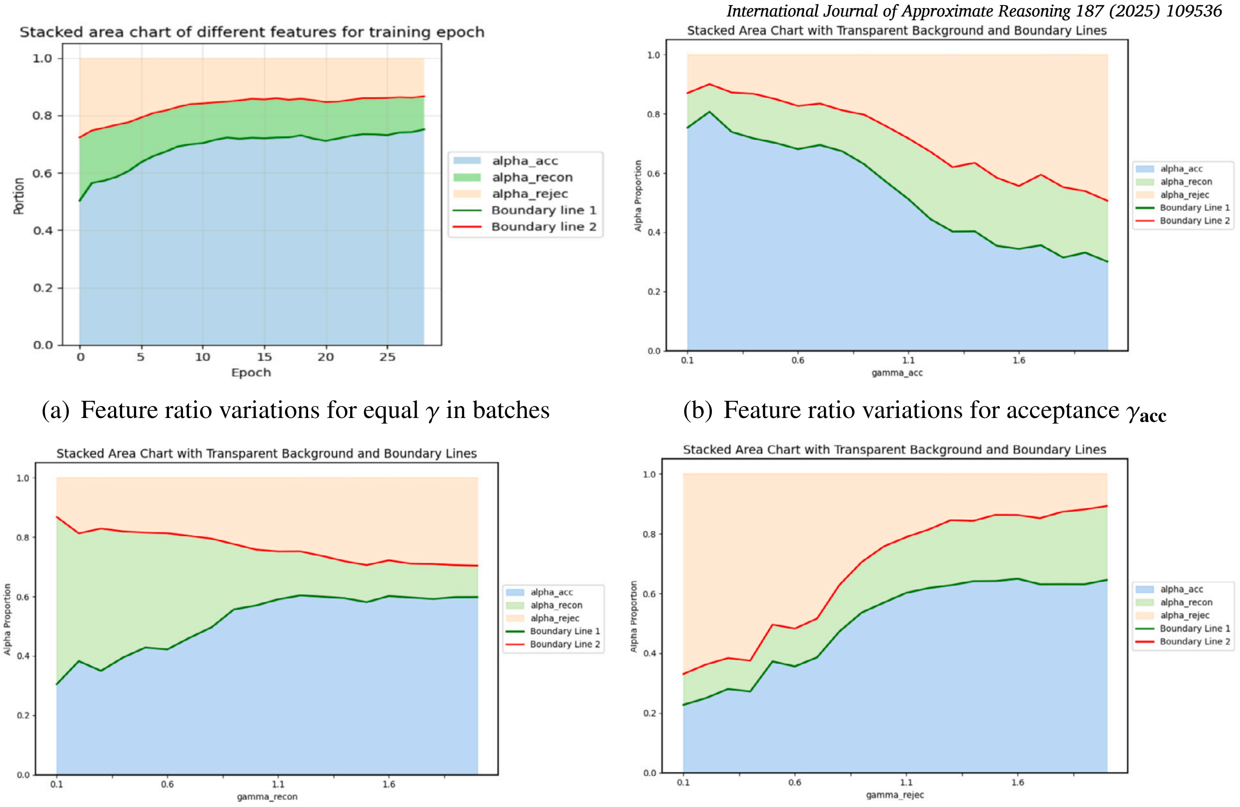
This subsection explores the influence of the hyperparameters as in Eq (11), the $\gamma = (\gamma_{acc}, \gamma_{recon}, \gamma_{dis})$ plays a crucial role in the TWSAFM model, as it governs the balance between the three types of feature selection: acceptable, rejectable, and reconsidering. By systematically varying one hyperparameter within a range (0.1 to 2.0) while keeping the others fixed at 1, we analyze its impact on performance metrics such as A7 accuracy, MAE, and correlation. The performance metrics, including A7 accuracy, MAE, and Correlation, are recorded for each configuration to assess how changes in γ influence the model's ability to filter features effectively.

The CMU-MOSI dataset is selected for analysis due to its relatively smaller size and manageable complexity compared to CMU-MOSEI and CH-SIMS. This allows for faster experimentation and more consistent observations over multiple training sessions. Additionally, CMU-MOSI provides a balanced mix of challenging multimodal interactions and clear sentiment annotations, making it ideal for evaluating the sensitivity of hyperparameters without introducing excessive noise or variability. While larger datasets like CMU-MOSEI could offer broader insights, the increased computational demands and potential overfitting risks make CMU-MOSI more suitable for controlled sensitivity analyses.

5.2.1. *Effect of punishment hyperparameter* $(\gamma_{acc}, \gamma_{recon}, \gamma_{dis})$

The sensitivity of the α parameters representing the proportions of acceptable, reconsidering, and rejectable features are carefully measured against variations in the γ values. As illustrated in Fig. 6, this section demonstrates the impact of a single γ value on the feature selection strategy, while other γ values remain 1 during the training process.

When all three γ values are set to 1, representing the final training configuration, the loss function incorporates only similarity loss and MAE loss. Under these conditions, the model demonstrates a preference for different feature selections, as depicted in Fig. 6(a), with the proportions of acceptance, reconsideration, and rejection being approximately 5:2.5:2.5. This distribution reflects



(a) Feature ratio variations for equal γ in batches

(b) Feature ratio variations for acceptance $\gamma_{\mathbf{acc}}$

(c) Feature ratio variations for reconsideration $\gamma_{\mathbf{recon}}$

(d) Feature ratio variations for rejection $\gamma_{\mathbf{dis}}$

Fig. 6. Feature ratio variations for different γ from 0.1 to 2.0.

the model's evaluation of feature importance, which aligns with our understanding that most real-world data contains a majority of useful information accompanied by some noise. Additionally, a small portion of high-dimensional data requires further processing to reveal its value. For instance, in determining a person's emotional state, their facial expression is relatively useful information. However, recognizing facial features can be influenced by factors like lighting and temperature, while micro-expressions containing significant information are difficult to detect.

Next, this experiment examines how variations in γ values influence the distribution of these three categories. As described above, when a penalty factor γ is applied, the model not only considers the utility of features for specific tasks but also evaluates the costs associated with different feature selection strategies. During training, the output demonstrates that when the model adopts an extremely lenient approach toward a certain selection strategy (γ set to 0.1), it classifies a larger proportion of features into this category. In such cases, the proportions of features in the acceptance and rejection strategies can exceed 70%, while the reconsideration remains limited to approximately 50%. Conversely, when the model is extremely strict with a certain strategy (γ set to 2.), the proportion of features assigned to the corresponding category decreases. Despite this, the acceptance strategy consistently retains a baseline proportion of around 30%, while the other two categories ultimately account for about 10% each.

The trends in Figs. 6(b), 6(c), and 6(d) illustrate the TWSAFM model's nuanced feature selection dynamics. In Fig. 6(b), as tolerance for the acceptance category decreases, most reduced acceptable features shift to rejection, with minimal growth in reconsideration, highlighting that excessive or insufficient initial information disrupts balance. Fig. 6(c) shows that as γ_{recon} increases, the reconsideration category grows steadily, acting as a buffer that adjusts to changes in acceptance and rejection. Similarly, in Fig. 6(d), higher penalties for rejection increase reconsideration and acceptance proportions, ensuring more high-dimensional information is retained. These patterns demonstrate the model's ability to dynamically allocate features and optimize performance by balancing the three categories.

5.2.2. *Effect of* $(\alpha_{\mathbf{acc}}, \alpha_{\mathbf{recon}}, \alpha_{\mathbf{dis}})$ *for the model's performance*

To facilitate analysis, we conducted a correlation study between the α values obtained from the aforementioned experiments and various performance metrics, such as MAE and A7. Although we understand that the magnitude of $(\alpha_{\mathbf{acc}}, \alpha_{\mathbf{recon}}, \alpha_{\mathbf{dis}})$ is influenced by $(\gamma_{acc}, \gamma_{recon}, \gamma_{dis})$, directly analyzing alpha proves to be more advantageous in deepening our understanding of the model.

As depicted in Fig. 7, it is evident that more information, including both acceptable and reconsidering information, leads to an improvement in MAE, thereby allowing the model to better fit the data. However, different strategies result in distinct fitting approaches. The acceptance strategy and reconsideration strategy resemble two modes of thinking—fast and slow [13]. A higher

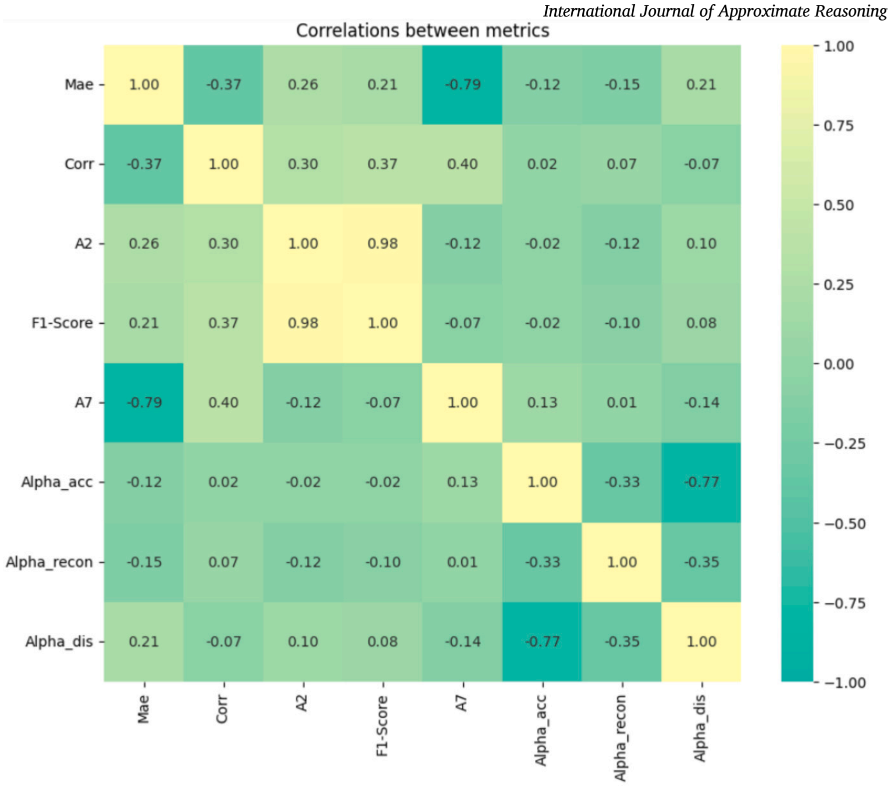


Fig. 7. Correlation of feature ratios with performance metrics.

frequency of acceptance decisions, or fast thinking, is effective in most cases as it relies more on the relevant features of individual data points. This emphasis on acceptable features enhances individual classification accuracy, as indicated by the improvement in A7. However, it may concurrently reduce the linear correlation for the dataset as a whole. On the other hand, the reconsideration strategy, which is slower and incorporates more experience, judging by the overall characteristics of the data, subsequently increases the correlation. The rejection strategy, which involves excluding certain noise before training, may increase individual classification efficiency, so the rejection strategy also plays a critical role.

Therefore, for the TWSAFM model we proposed, all three strategies are indispensable. After multiple rounds of experiments, our study adopts a gamma combination that offers stronger generalizability $\gamma = (1.2, 1.0, 1.2)$. This combination forces the model to rigorously filter out noise while preventing an excessive number of features from being classified into the acceptance category or rejection category (like the situation gammas are set to 1), simultaneously enhancing the model's slow thinking capabilities and thereby improving overall performance.

5.3. *Effect for training of TWSAFM*

5.3.1. *The role of improvement for data redundancy*

Although TWSAFM performs well on these three datasets in Table 1, the underlying mechanisms of specific structures have not yet been fully explored. Liu et al. [20] demonstrated that pre-dropout can make models more robust, but it remains uncertain whether the self-adaptive selection mechanism, with a lower uncertainty compared to dropout, outperforms dropout. To investigate this, we design a series of experiments where features are passed through either the TWSAFM or pre-dropout TWSAFM without a filtering module before proceeding to downstream tasks.

To validate the effectiveness of TWSAFM in improving data redundancy problems, we conduct validation experiments using the CMU-MOSI dataset. In the case of high data redundancy, certain features may negatively affect the model's predictions. By using dropout, the model randomly selects the participating features at each training, which helps naturally screen out the more important features, thereby reducing the interference of redundant information. Therefore, in a common machine learning practice, using dropout can reduce data redundancy to some extent and improve model performance. Based on this background, this subsection designs a set of experiments. After the multimodal data feature fusion, we apply various degrees of dropout to it and then input the processed features into subsequent models. The experiment is divided into two groups: one is TWSAFM model and the other is a pre-dropout TWSAFM model without the filtering module, 30 rounds of training are conducted for each dropout rate, and the optimal results are obtained, aiming to observe the effects of different dropout rates on the performance of these two models. The specific results are shown in the Table 5. The upper part shows the performance of pre-dropout TWSAFM without the selection mechanism, while the lower part shows the performance of TWSAFM.

Table 5
Performance for feature pre-dropout with different ratios.

Model	Dropout ratio for modalities	A2(↑)	A7(↑)	MAE(↓)	Corr(↑)	F1(↑)
Pre-dropout TWSAFM without filtering module	*0.00*	82.1 / 84.5	45.5	0.761	0.768	78.6 / 81.0
	0.05	81.9 / 83.8	43.4	0.768	0.767	78.4 / 80.1
	0.10	82.4 / 84.0	44.2	0.775	0.763	79.9 / 81.2
	0.15	79.6 / 82.0	42.1	0.768	0.761	75.4 / 77.7
	0.20	81.1 / 83.4	44.5	0.767	0.750	76.7 / 79.0
	0.25	82.1 / 84.0	42.1	0.757	0.765	78.8 / 80.5
	0.30	**83.1** / **85.4**	46.4	**0.751**	**0.775**	79.6 / **81.8**
	0.35	80.9 / 82.8	**46.7**	0.785	0.742	77.8 / 79.4
	0.40	82.1 / 83.5	42.9	0.785	0.756	**79.8** / 80.9
	0.45	80.5 / 82.5	44.0	0.776	0.761	76.3 / 78.1
	0.50	80.8 / 82.0	44.5	0.785	0.752	78.0 / 78.7
	mean(std)	81.5(1.01) / 83.4(1.05)	44.2(1.55)	0.771(0.012)	0.760(**0.009**)	**78.1(1.48)** / 79.9(1.36)
TWSAFM	*0.00*	**83.2** / **84.5**	44.9	**0.748**	**0.776**	**81.1** / **81.9**
	0.05	81.6 / 83.8	**46.1**	0.767	0.756	77.7 / 79.9
	0.10	81.3 / 83.8	45.0	0.751	0.767	77.6 / 80.2
	0.15	81.5 / 83.5	44.8	0.773	0.748	77.8 / 79.6
	0.20	82.2 / 83.9	43.4	0.765	0.765	77.9 / 80.8
	0.25	80.4 / 82.9	42.9	0.763	0.763	75.5 / 78.0
	0.30	80.8 / 81.9	42.0	0.793	0.748	78.4 / 78.9
	0.35	82.4 / 83.4	42.4	0.782	0.774	80.3 / 80.7
	0.40	82.2 / 84.4	43.7	0.749	0.769	78.9 / 81.4
	0.45	81.3 / 83.5	45.0	0.754	0.771	77.3 / 79.4
	0.50	81.3 / 83.4	45.9	0.750	0.770	78.2 / 80.1
	mean(std)	**81.7(0.79)** / **83.5(0.71)**	**44.2(1.38)**	**0.763(0.015)**	**0.764**(0.010)	78.2(1.49) / **80.1(1.11)**

From the experimental results in Table 5, the TWSAFM model shows relatively stable performance in the range of 0.00 to 0.50 dropout ratio. The experiment results indicate that TWSAFM models can self-select and process input features, significantly reducing the impact of data redundancy on the model. In addition, for different dropout ratios, the performance stability of TWSAFM models is generally better than that of models without feature selection, which further indicates that feature selection can enhance the robustness of the model. On the other hand, when feature selection is omitted and input directly into the TWSAFM model, the overall trend shows that at a moderate dropout ratio, the model's performance remains good, but there is slight performance degradation in some cases. This suggests that although dropout can mitigate interference from redundant features, too high or too low a dropout ratio may inhibit the model's ability to learn important features. Choosing the appropriate dropout ratio becomes an empirical problem in the task, which also brings certain challenges to the research. TWSAFM model can effectively avoid this problem through adaptive feature selection. The experimental results emphasize the importance of rational feature selection in multimodal learning, which provides useful guidance for the subsequent model optimization.

5.3.2. The performance visualization

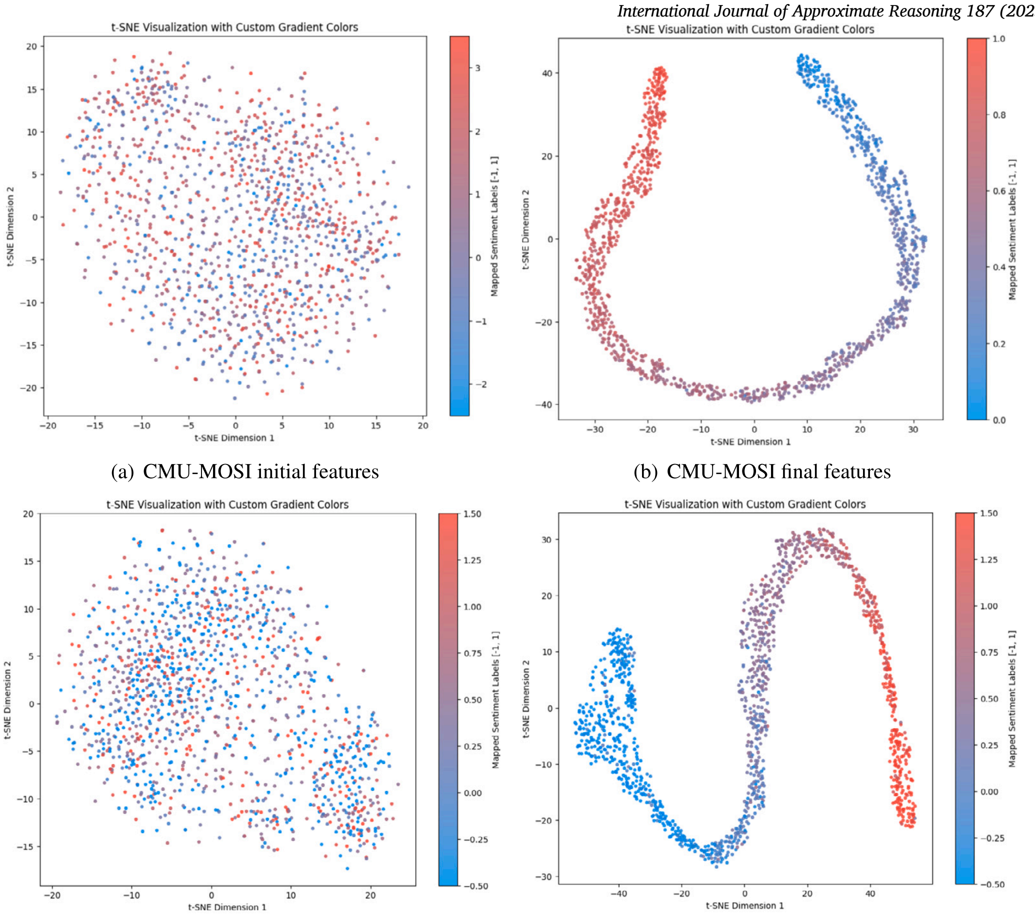
To showcase the effectiveness of TWSAFM model in sentiment prediction tasks, we employ the t-SNE technique to visualize the integrated representations generated by the model on CMU-MOSI and CH-SIMS. The resulting visualizations are presented in Fig. 8. The predicted sentiment labels, ranging from [-3, 3] (CMU-MOSI), are mapped to the interval [0, 1], with [0, 0.5] corresponding to positive sentiment and [0.5, 1] representing negative sentiment.

For the CMU-MOSI dataset, initial feature distributions are scattered with unclear class boundaries, while the TWSAFM's feature selection mechanism organizes features into tighter, distinct clusters, improving classification accuracy. Similarly, for the CH-SIMS dataset, TWSAFM reduces overlap in features, especially for neutral sentiment, and better separates samples with strong polarity. These improvements, enabled by TWSAFM's three-way decision strategies and cross-modal attention to similarity, enhance both intra-class compactness and inter-class separability, as reflected in higher accuracy and correlation scores. The t-SNE visualizations affirm TWSAFM's efficacy in multimodal sentiment analysis.

5.4. Ablation experiment

To further dissect the effectiveness of the TWSAFM model and identify the contributions of each component, a comprehensive ablation study is conducted. In this subsection, the self-adaptive filtering module, cost-sensitive module, and similarity loss function are removed respectively from TWSAFM architecture to analyze its contribution to performance.

First, the self-adaptive filtering module serves as a central innovation of the TWSAFM model, which reduces information redundancy while maintaining essential cross-modal information. To evaluate its contribution, we removed this module and used a widely used dropout module. As shown in Table 6 and Table 7, by replacing the self-adaptive filtering module, a significant drop in shown performance, particularly in classification performance. Further, the removal of this module also leads to poorer generalization across datasets, suggesting that the self-adaptive filtering module plays a vital role in controlling feature relevance and noise. Its ability to handle high-dimensional, multimodal data by categorizing features into "acceptable", "rejectable", and "reconsidering" categories is



(a) CMU-MOSI initial features

(b) CMU-MOSI final features

(c) CH-SIMS initial features

(d) CH-SIMS final features

Fig. 8. TWSAFM's performance on features fusion.

Table 6
Ablation study on CMU-MOSI and CMU-MOSEI.

Model	CMU-MOSI				CMU-MOSEI			
	A2(↑)	A7 (↑)	MAE (↓)	Corr (↑)	A2 (↑)	A7 (↑)	MAE (↓)	Corr (↑)
TWSAFM	82.9/85.2	44.9	0.718	0.786	84.0/84.5	53.3	0.539	0.759
$\text{Loss}_{sim}(-)$	81.1/83.1	44.6	0.739	0.759	82.9/83.3	52.5	0.545	0.752
Cost Sensitive(-)	82.1/84.2	43.0	0.748	0.760	82.6/83.9	52.1	0.550	0.750
Feature Selection(-)	81.9/84.0	44.2	0.742	0.776	83.2/84.0	52.9	0.548	0.753

crucial for maintaining a balance between feature richness and noise filtering. Without this module, the model struggles to differentiate between useful and redundant information, leading to higher variance in the final representations and lower performance across all key metrics.

Then we examine the impact of the cost-sensitive module, which assigns different weights to features based on their importance. By removing this module, a decline in performance is also shown, especially in scenarios involving imbalanced datasets such as CMU-MOSEI. The cost-sensitive module ensures that features categorized as "acceptable" or "reconsidering" are treated differently from those discarded, which helps the model to fine-tune its learning process. When this module is removed, the model shows a more uniform treatment of features, leading to suboptimal weight distribution across modalities. This is especially problematic when dealing with minority class examples in the dataset, causing the model to overfit certain modalities while underutilizing others. The F1 score and A7 accuracy metrics also suffered, further reinforcing the need for a cost-sensitive approach to managing feature importance.

The similarity loss function is designed to ensure that representations from different modalities are aligned in terms of sentiment, promoting a consistent cross-modal interpretation of emotional states. When we remove this component, the model's ability to integrate information from text, audio, and visual cues is compromised. A decrease in correlation metrics reflects this view, indicating that the model has difficulty learning the relationships among modalities.

Table 7
Ablation study on CH-SIMS.

	A2(↑)	A3 (↑)	A5 (↑)	MAE (↓)	Corr (↑)	F1(↑)
TWSAFM	79.4	66.3	41.1	0.424	0.591	79.4
Loss$_{sim}$(−)	78.8	65.4	41.0	0.429	0.587	78.8
Cost Sensitive(-)	76.8	64.3	39.8	0.442	0.574	76.5
Feature Selection(-)	78.5	64.5	40.5	0.431	0.585	78.4

When all modules are reintroduced in the full TWSAFM model, the performance gains are immediately apparent. The self-adaptive filtering module works synergistically with the cost-sensitive module and similarity loss function to refine the model's handling of multimodal data. In those three datasets, the full model achieves superior performance, particularly excelling in multi-class sentiment recognition tasks.

The ablation results demonstrate the critical role of each module in the TWSAFM model. By carefully balancing feature selection, cost sensitivity, and cross-modal alignment, the TWSAFM model outperforms existing methods. This ablation study validates the design choices of the TWSAFM architecture and highlights its ability to handle the complexity of multimodal sentiment analysis in real-world scenarios. In conclusion, this extended ablation experiment not only confirms the individual importance of each component within the TWSAFM model but also emphasizes how these components work together to produce a robust and efficient multimodal sentiment analysis framework, ensuring the model's adaptability and performance across diverse datasets and tasks.

6. Conclusion

In this research, a Three-Way based Self-Adaptive Filtering model (TWSAFM) is proposed, which aims at addressing two key challenges in multimodal sentiment analysis: managing information redundancy and aligning heterogeneous modality data. The proposed method leverages the TWD theory to classify modality features into acceptable, rejectable, and reconsidering domains, enabling precise filtering and prioritization of relevant information. Additionally, a cross-modal attention mechanism is incorporated to ensure an effective alignment and interaction across diverse modality features, further enhancing the robustness and adaptability of the model.

Experiments on three public datasets such as CMU-MOSI, CMU-MOSEI, and CH-SIMS demonstrate the promising performance of the TWSAFM model compared to state-of-the-art methods. In addition, parameter sensitivity and ablation studies are further conducted to validate the contribution of different modules in the TWSAFM model. In summary, the experiments demonstrate that the TWSAFM can dynamically filter redundant information and optimize cross-modal feature fusion, which highlights its potential for real-world applications in human-computer interaction, mental health assessment, and social media analytics.

In conclusion, this work provides a substantial contribution to the field of sentiment analysis by presenting a structured and adaptive solution to complex multimodal challenges. The TWSAFM model not only enhances interpretability and efficiency but also lays a robust foundation for future advancements in AI-driven sentiment analysis.

While the TWSAFM model effectively addresses the challenges of feature filtering and cross-modal alignment, several areas remain for further improvement. First, the model's feature selection mechanism would benefit from clearer theoretical explanations and better visualization to enhance its interpretability. Besides, further research may examine the model's scalability by incorporating additional modalities, such as physiological signals (e.g., heart rate, skin conductance) or dynamic social media data (e.g., user interactions, temporal sentiment shifts). For instance, integrating physiological signals like heart rate variability, galvanic skin response, or electroencephalogram (EEG) data could provide deeper insights into emotional states, as these signals often reflect subconscious reactions that complement overt expressions captured in text, audio, and visual data. Additionally, incorporating dynamic social media data, such as real-time sentiment trends from platforms like Twitter, could enhance the model's ability to adapt to evolving public opinions and emotional contexts. These modalities could provide deeper insights into emotional states and enhance the model's applicability in real-time sentiment analysis scenarios.

CRediT authorship contribution statement

Zhihui Zhang: Writing – original draft, Validation, Methodology, Formal analysis, Data curation. **Dun Liu:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Rongping Shen:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the National Science Foundation of China (Nos. 62276217 and 62406262), and the Fundamental Research Funds for the Central Universities (Nos. 2682024ZTPY021 and 2682024KJ005).

Data availability

Data will be made available on request.

References

[1] S. Al-Saqqa, H. Abdel-Nabi, A. Awajan, A survey of textual emotion detection, in: 2018 8th International Conference on Computer Science and Information Technology (CSIT), IEEE, 2018, pp. 136–142.

[2] T. Baltrusaitis, A. Zadeh, Y. Lim, L. Morency, Openface 2.0: facial behavior analysis toolkit, in: 13th IEEE Int. Conf. Automatic Face & Gesture Recognition (FG 2018), 2018, pp. 59–66.

[3] M.E. Basiri, M. Abdar, M.A. Cifci, S. Nemati, U.R. Acharya, A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques, Knowl.-Based Syst. 198 (2020) 105949.

[4] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, Covarep—a collaborative voice analysis repository for speech technologies, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 960–964.

[5] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Inf. Fusion 91 (2023) 424–444.

[6] Y. Guo, B. Sun, J. Bai, J. Ye, X. Chu, A new portfolio approach integrating three-way decision and encoder–decoder network, Expert Syst. Appl. 258 (2024) 125233.

[7] W. Han, H. Chen, S. Poria, Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 9180–9192.

[8] D. Hazarika, R. Zimmermann, S. Poria, Misa: modality-invariant and-specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1122–1131.

[9] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.

[10] J. Hou, N. Omar, S. Tiun, S. Saad, Q. He, Tchfn: multimodal sentiment analysis based on text-centric hierarchical fusion network, Knowl.-Based Syst. 300 (2024) 112220.

[11] F. Huang, X. Zhang, Z. Zhao, J. Xu, Z. Li, Image–text sentiment analysis via deep multimodal attentive fusion, Knowl.-Based Syst. 167 (2019) 26–37.

[12] X. Jia, Z. Deng, F. Min, D. Liu, Three-way decisions based feature fusion for Chinese irony detection, Int. J. Approx. Reason. 113 (2019) 324–335.

[13] D. Kahneman, Thinking, Fast and Slow, vol. 27, Farrar, Straus and Giroux, 2011, pp. 54–57.

[14] J.D.M.-W.C. Kenton, L.K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NaacL-HLT, Minneapolis, Minnesota, vol. 1, 2019, p. 2.

[15] K. Kim, S. Park, Aobert: all-modalities-in-one bert for multimodal sentiment analysis, Inf. Fusion 92 (2023) 37–45.

[16] Z. Li, Z. Huang, Y. Pan, J. Yu, W. Liu, H. Chen, Y. Luo, D. Wu, H. Wang, Hierarchical denoising representation disentanglement and dual-channel cross-modal-context interaction for multimodal sentiment analysis, Expert Syst. Appl. 252 (2024) 124236.

[17] Z. Li, P. Zhang, N. Xie, G. Zhang, C.-F. Wen, A novel three-way decision method in a hybrid information system with images and its application in medical diagnosis, Eng. Appl. Artif. Intell. 92 (2020) 103651.

[18] D. Liang, B. Yi, W. Cao, Q. Zheng, Exploring ensemble oversampling method for imbalanced keyword extraction learning in policy text based on three-way decisions and smote, Expert Syst. Appl. 188 (2022) 116051.

[19] Z. Liu, Y. Shen, V. Bharadhwaj Lakshminarasimhan, P.P. Liang, A. Zadeh, L.-P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, arXiv:1806.00064 [cs.AI], 2018.

[20] Z. Liu, Z. Xu, J. Jin, Z. Shen, T. Darrell, Dropout reduces underfitting, in: International Conference on Machine Learning, in: PMLR, 2023, pp. 22233–22248.

[21] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, S. Poria, Multimodal sentiment analysis using hierarchical fusion with context modeling, Knowl.-Based Syst. 161 (2018) 124–133.

[22] B. McFee, C. Raffel, D. Liang, D.P. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: audio and music signal analysis in python, in: SciPy, 2015, pp. 18–24.

[23] F. Min, Z.-H. Zhang, W.-J. Zhai, R.-P. Shen, Frequent pattern discovery with tri-partition alphabets, Inf. Sci. 507 (2020) 715–732.

[24] W. Rahman, M.K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, E. Hoque, Integrating multimodal information in large pretrained transformers, in: Proceedings of the Conference, Association for Computational Linguistics, Meeting, in: NIH Public Access, vol. 2020, 2020, p. 2359.

[25] F. Shen, X. Zhao, G. Kou, Three-stage reject inference learning framework for credit scoring using unsupervised transfer learning and three-way decision theory, Decis. Support Syst. 137 (2020) 113366.

[26] R.-P. Shen, D. Liu, X. Wei, M. Zhang, Your posts betray you: detecting influencer-generated sponsored posts by finding the right clues, Inf. Manag. 59 (2022) 103719.

[27] L. Sun, Z. Lian, B. Liu, J. Tao, Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis, IEEE Trans. Affect. Comput. 15 (2023) 309–325.

[28] A.-H. Tan, B. Subagdja, D. Wang, L. Meng, Self-organizing neural networks for universal learning and multimodal memory encoding, Neural Netw. 120 (2019) 58–73.

[29] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the Conference, Association for Computational Linguistics, Meeting, in: NIH Public Access, vol. 2019, 2019, p. 6558.

[30] Y.-H.H. Tsai, P.P. Liang, A. Zadeh, L.-P. Morency, R. Salakhutdinov, Learning factorized multimodal representations, arXiv:1806.06176 [cs.LG], 2018.

[31] X. Wang, M. Wang, H. Cui, Y. Zhang, A dual-channel multimodal sentiment analysis framework based on three-way decision, Eng. Appl. Artif. Intell. 137 (2024) 109174.

[32] Z. Wang, Z. Wan, X. Wan, Transmodality: an end2end fusion method with transformer for multimodal sentiment analysis, in: Proceedings of the Web Conference 2020, 2020, pp. 2514–2520.

[33] L. Xiao, X. Wu, S. Yang, J. Xu, J. Zhou, L. He, Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis, Inf. Process. Manag. 60 (2023) 103508.

[34] A. Yadollahi, A.G. Shahraki, O.R. Zaiane, Current state of text sentiment analysis from opinion to emotion mining, ACM Comput. Surv. 50 (2017) 1–33.

[35] X. Yang, Y. Li, Q. Li, D. Liu, T. Li, Temporal-spatial three-way granular computing for dynamic text sentiment classification, Inf. Sci. 596 (2022) 551–566.

[36] X. Yang, M.A. Loua, M. Wu, L. Huang, Q. Gao, Multi-granularity stock prediction with sequential three-way decisions, Inf. Sci. 621 (2023) 524–544.

[37] Y. Yao, Three-way decision: an interpretation of rules in rough set theory, in: Rough Sets and Knowledge Technology: 4th International Conference, RSKT 2009, Gold Coast, Australia, July 14–16, 2009, Proceedings, vol. 4, Springer, 2009, pp. 642–649.

[38] Y. Yao, Three-way decisions with probabilistic rough sets, Inf. Sci. 180 (2010) 341–353.

[39] Y. Yao, Three-way decision and granular computing, Int. J. Approx. Reason. 103 (2018) 107–123.

[40] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, K. Yang, Ch-sims: a Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3718–3727.

[41] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 10790–10797.

[42] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1103–1114.

[43] A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, L.-P. Morency, Memory fusion network for multi-view sequential learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.

[44] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, arXiv:1606.06259 [cs.CL], 2016.

[45] A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, in: Long Papers, vol. 1, 2018, pp. 2236–2246.

[46] Y. Zeng, W. Yan, S. Mai, H. Hu, Disentanglement translation network for multimodal sentiment analysis, Inf. Fusion 102 (2024) 102031.

[47] C. Zhang, Z. Yang, X. He, L. Deng, Multimodal intelligence: representation learning, information fusion, and applications, IEEE J. Sel. Top. Signal Process. 14 (2020) 478–493.

[48] D. Zong, C. Ding, B. Li, J. Li, K. Zheng, Q. Zhou, Acformer: an aligned and compact transformer for multimodal sentiment analysis, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 833–842.