Paradigm Academic Press
Innovation in Science and Technology
ISSN 2788-7030
MAY. 2025 VOL.4, NO.4

Research on Multi-Modal Question Answering Emotion Recognition Method Based on User Preference

Songyu Ji[1]

[1] School of Management, Hefei University of Technology, Hefei 230009, China

Correspondence: Songyu Ji, School of Management, Hefei University of Technology, Hefei 230009, China.

doi:10.63593/IST.2788-7030.2025.05.006

Abstract

The current multimodal sentiment recognition methods fall short in handling dynamic changes in modal weights and modeling modal consistency. Specifically, when processing the MELD dataset, multiple rounds of structured processing and feature optimization were not conducted; meanwhile, the Word2Vec similarity-based sentiment lexicon expansion strategy still falls short in terms of semantic consistency and emotional accuracy. Additionally, in the original experimental setup, the model relied solely on cross-entropy loss for training, overlooking the uncertainties and inconsistencies in information fusion across modalities. Therefore, this project proposes a multimodal question-answering sentiment recognition method based on user preferences. By introducing a multimodal attention mechanism guided by sentiment preferences and sentiment prototypes in a three-dimensional sentiment representation space (Valence-Arousal-Dominance, VAD), the method enhances multimodal information fusion and modeling capabilities for modal consistency. Furthermore, an extended sentiment lexicon strategy and context-dependent modeling mechanism are designed to improve the accuracy and stability of dialogue sentiment recognition. The project conducted systematic ablation and comparative experiments on the standard multimodal dialogue sentiment dataset MELD, demonstrating that the proposed method outperforms existing representative models in accuracy, precision, and F1 score, validating its effectiveness and potential for application in multimodal question-answering sentiment recognition tasks.

1. Background of Research

In recent years, the booming development of the Internet and social media has greatly changed how people express their emotions. In online interaction scenarios such as Q&A communities and social platforms, users are no longer limited to text-based communication but increasingly use various modalities like images, videos, voice, and emojis to convey emotions and viewpoints. This multimodal expression provides richer data sources for sentiment analysis; for example, changes in speech tone, facial expressions in videos, and emotional words in text can all serve as effective clues for emotion recognition (Eyben, F., et al., 2016). However, the heterogeneity of multimodal data, information conflicts between modalities, and dynamically changing modal importance also present new challenges to emotion recognition tasks.

Despite significant progress in single-modal (especially text) sentiment analysis tasks with large-scale pre-trained language models like BERT and GPT (Sun, H., Niu, Z., Wang, H., et al., 2025), many limitations still exist when handling multimodal data. First, different modalities have distinct feature distributions (such as the discrete symbol characteristics of text and the continuous temporal characteristics of speech), making cross-modal feature alignment and fusion complex. Second, existing methods typically assume that the emotional contribution of each modality is fixed, for example, assigning static weights to text, speech, and visual modalities during training (Yang, L., 2025). However, in actual conversations, different users may prefer to express their emotions using different modalities (for instance, some users rely more on voice tone, while others prefer emojis), and the dominant modality for the same user can change across different contexts (Zhang, Y., Li,

X., & Chen, W., 2024). If this dynamic shift in modal importance is not adequately considered, it will directly impact the robustness of sentiment recognition.

In addition, there may be inconsistencies between multimodal data, meaning that the emotional information expressed in different modalities can contradict each other. Take a simple piece of text as an example: “The newly bought cup is really good quality!” If we interpret it only at the textual level, we might only perceive the speakers satisfaction with the cup. However, when this message is accompanied by an image of a cup with obvious scratches, the conveyed emotion undergoes a dramatic reversal. This complementarity between different modalities allows machines to achieve more precise and comprehensive sentiment analysis (Liu, Y., et al., 2023).

Existing methods often use simple feature stitching or weighted averaging strategies for modal fusion, which fail to effectively address such conflicts, leading to a decline in sentiment recognition accuracy. Additionally, sentiment recognition in conversational scenarios must consider the coherence of contextual emotions and the emotional influence between speakers. For example, in Q&A communities, the responders emotion may be influenced by the questioners tone, yet current research still falls short in modeling these interactive relationships (Zhang, K., et al., 2024).

In view of the above challenges, this study intends to construct a multimodal question answering emotion recognition model considering users emotional preferences to improve the accuracy of emotion recognition. At the same time, the model will be applied to online question answering system to improve user experience.

2. Method Design

2.1 Structured Processing of M ELD Data Set

In order to better adapt to the multimodal emotion recognition task, this study has carried out several rounds of structured processing and feature optimization on the original MELD data set, mainly including the following aspects:

2.1.1 Original Data Repair and Structure Unification

The original MELD provides multimodal data in the form of text CSV and multimodal.pkl files, which have Utterance_ID alignment issues. To ensure that the model can accurately align multimodal features with text labels, we used additional scripts to complete and merge information such as Utterance, Speaker, and Dialogue_ID from the CSV into the original.pkl files, forming unified format data files video_train_fixed.pkl, video_dev_fixed.pkl and video_test_fixed.pkl, which are indexed and saved in dictionary form.

2.1.2 Text Feature Processing

The BERT WordPiece is used to encode the concatenated context text. To enhance the modeling capability for multi-turn conversations, a context concatenation strategy (Context Concatenation) is introduced, which dynamically concatenates the first context_size=2 instances with the same conversation content (while retaining speaker labels), forming a format such as:

[SPEAKER=Joey]: How are you? [SEP] [SPEAKER=Chandler]: I’m fine.

In this way, multiple rounds of context dependence can be captured to improve the effect of emotion understanding.

2.1.3 Modal Feature Alignment and Length Unification

The length differences of multimodal features pose a challenge to model input, so this study standardizes the maximum temporal length for all modalities: Video modality (video_features): up to 12 frames; Audio modality (audio_features): up to 12 frames; Text input: BERTs maximum length is limited to 48 tokens. Any shortfall is uniformly padded with zero padding (Zero Padding) to ensure tensor alignment during batch processing.

2.1.4 Speaker Embedding (Speaker Embedding)

In order to model the differences in emotional expression between different speakers, the speaker ID embedding mechanism is introduced. The 8 fixed roles (such as Ross, Rachel, etc.) are mapped into one-hot encoded ID and input into the embedding layer to learn their emotional style representation, which is used to assist emotion classification.

2.1.5 Tag Structure and Sorting

In the preloading data phase, all samples are sorted according to (dialogue_id, utterance_id) to ensure the temporal consistency of the conversation rounds, thus ensuring the quality of context modeling. The emotion label field maintains the original data sets 7 categories of emotion classification standards (such as joy, anger, neutral, etc.).

2.2 Upgrade the Dictionary Expansion Strategy

In the original experimental design, we adopted a Word2Vec similarity-based sentiment lexicon expansion strategy. By training a Word2Vec word vector model on the MELD dataset and finding the most similar words in the word vector space for each OOV word missing from the NRC-VAD base dictionary, we weighted and estimated their VAD sentiment values. Although this method is concise and efficient, it still falls short in terms of semantic consistency and sentiment accuracy, primarily manifested as:

1) Similar words may be semantically related but have inconsistent emotional tendencies;
2) All OOV words use fixed default values [0.5,0.5,0.5], lacking context information;
3) Single similarity source (based only on Word2Vec), limited robustness.

Based on this, we have upgraded and optimized the dictionary expansion process in the following three aspects to enhance emotional consistency and context adaptation.

1) Joint similarity measure (Multi-Source Similarity Fusion):

The original scheme is only based on Word2Vec similarity; the new scheme integrates the similarity between Word2Vec and GloVe word vectors, and can be extended to support TF-IDF, FastText, etc., to improve the accuracy of word meaning through multi-source fusion, and to alleviate the influence of semantic deviation by averaging weighted similarity words according to multiple models.

2) Adaptive default value adjustment (Adaptive Initialization):

No longer fixed allocation [0.5,0.5,0.5] for OOV words lacking similar words; combine the overall emotional distribution in the corpus or introduce slight fluctuations (such as Gaussian disturbance) to dynamically generate more natural emotional values; reduce the interference of default values on the models emotional judgment.

3) Context-sensitive reasoning (Context-Aware Inference with BERT):

The pre-trained BERT model is used to extract the context vector representation of the target word in the real dialogue context; the context-aware inference strategy is introduced to simulate the method of context-based emotion prediction (such as fine-tuned VAD predictor), which greatly improves the contextual consistency and task relevance of emotion value estimation.

The upgrade of the above strategy makes the emotion dictionary expansion evolve from “static mapping of single semantic neighbors” to “multi-source fusion + context perception + distributed adaptation” dynamic reasoning process, which effectively improves the coverage, accuracy and robustness of the emotion dictionary, and provides more stable underlying feature support for multimodal emotion recognition.

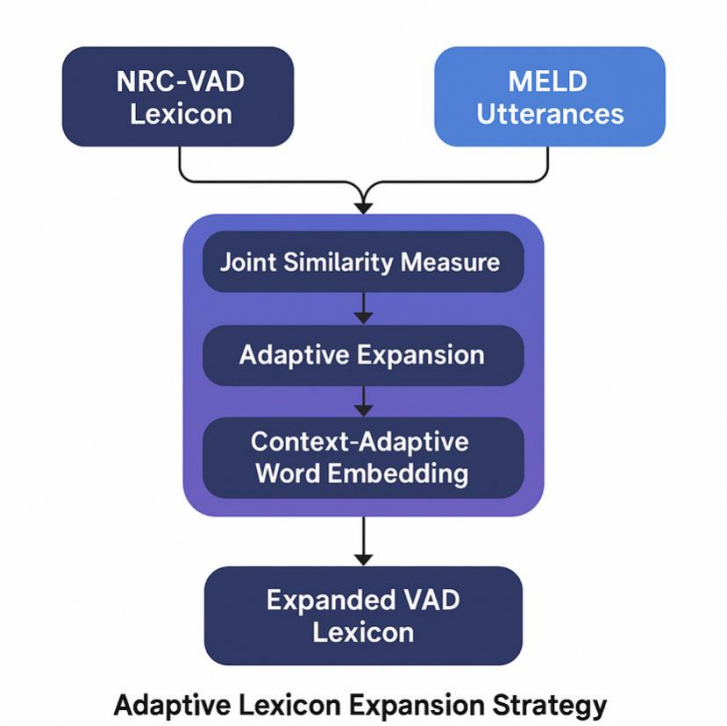

Figure 1.

2.3 Upgrade the Loss Function Design Strategy

In the original experimental setup, the model was trained solely using cross-entropy loss, ignoring the uncertainties and inconsistencies in information fusion across modalities. To enhance the robustness and credibility of multi-modal emotion recognition, we introduced a main task loss ce_total that includes Dirichlet KL divergence, and designed two additional supervision mechanisms for the emotional space: one is the VAD consistency loss between modalities (loss_vad_align), and the other is the difference supervision between text modality and dictionary VAD values (loss_vad_dict). The upgraded composite loss function loss_total effectively guides the model to achieve modal alignment and semantic enhancement within the emotional representation space, thereby significantly improving the models classification performance and generalization ability.

loss = F.cross_entropy (logits, label)

loss_total = ce_total + α * loss_vad_align + β * loss_vad_dict

1) VAD supervision loss design guided by dictionary

In multimodal emotion recognition, the text mode occupies a dominant position. In order to improve the quality of the text mode before modal fusion and make the three-mode VAD alignment (loss_vad_align) more accurate, we designed a dictionary-guided VAD supervision loss. Specifically:

(i) The text mode is embedded by BERT + VAD dictionary and then the output last_h_l represents the emotional semantics

(ii) Project the representation into a 3D space to get vad_l: the model predicts the text VAD representation

(iii) The mean value of the dictionary VAD of each sample is found from expanded_nrc_vad_lexicon.txt → vad_dict_target

(iiii) Then compare vad_l with vad_dict_target, and use MSE to measure the difference.

The existence of loss_vad_dict can not only guide the model to identify the implicit emotional tendency in semantics, but also provide prior semantic emotion supervision for low-resource samples or unpopular emotional categories.

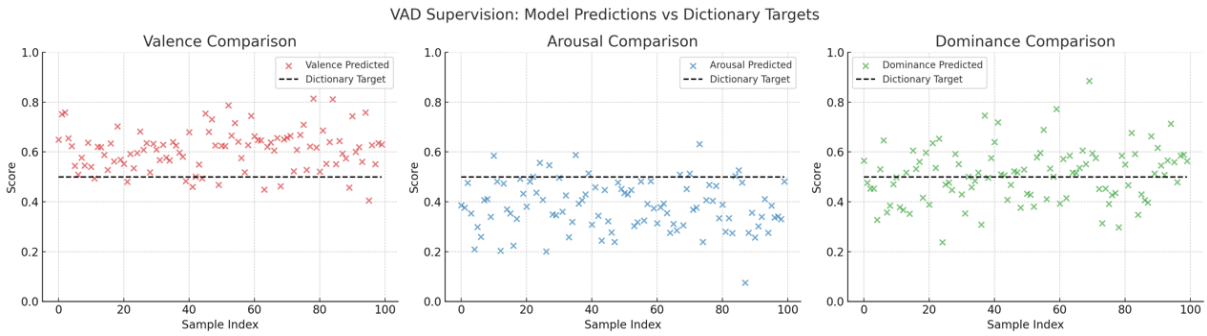

Figure 2.

The difference between the VAD value (red/blue/green dots) predicted by the model and the target value of the dictionary (black dotted line).

2) Contrastive learning of different modes in VAD space

In multimodal emotion recognition, different modalities (text, audio, video) exhibit a certain emotional consistency when expressing emotions. However, in the original model, each modality is processed separately, with only simple concatenation or attention mechanism interactions at the fusion layer, lacking clear supervisory signals to constrain the emotional space consistency across the three modalities. Therefore, we designed loss_vad_align to explicitly measure and enforce consistency between modalities in the three-dimensional VAD emotion space, enabling the model to learn more emotionally cohesive cross-modal representations.

Specifically, we map the three-mode features to the three-dimensional Valence-Arousal-Dominance space through projection network respectively, and use the mean square error to measure the difference between the three modes, so as to construct the consistency loss function:

$$\mathcal{L}_{\text{vad\_align}} = \text{MSE}(V_l, V_a) + \text{MSE}(V_l, V_v) + \text{MSE}(V_a, V_v)$$

This mechanism explicitly encourages multimodal consensus in emotional space expression, enhancing the consistency and robustness of the represented emotions. In experiments, we observed that loss_vad_align steadily decreased during training and played a positive role in the weighted fusion of total loss loss_total, contributing to improved emotional classification performance after multimodal fusion.

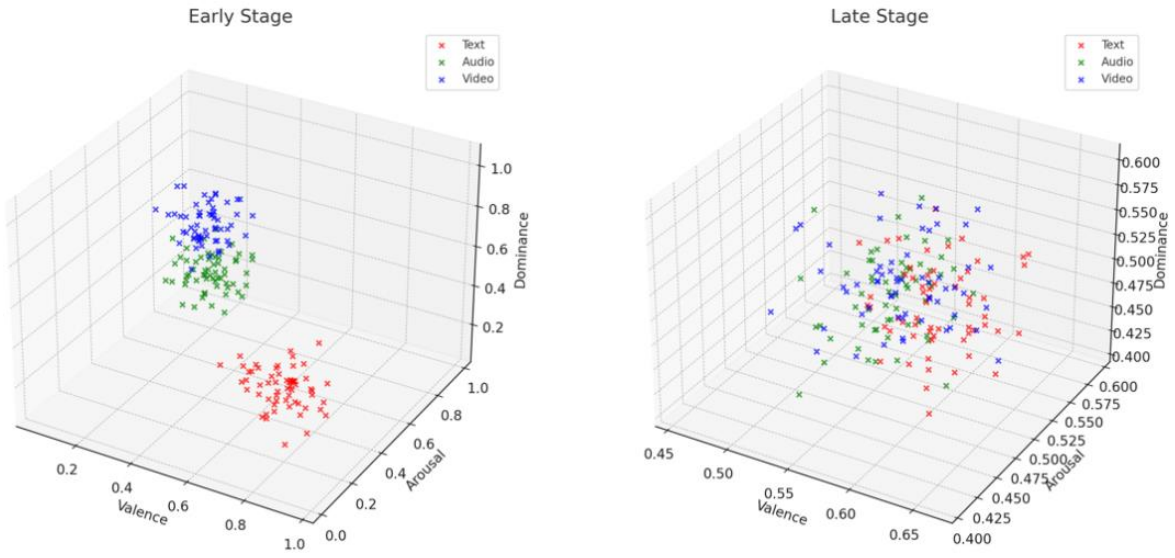

Figure 3.

The left figure shows that in the early stage of training, the emotional representation distribution of text (red), audio (green) and video (blue) modes is scattered and inconsistent, and there are conflicts between modes; the right figure shows that in the later stage of training, the three modes are gradually close to each other in VAD space through loss_vad_align, showing good emotional consistency.

3. Experimental Analysis

3.1 Experimental Settings

The original multi-turn dialogue modeling and speaker memory mechanism led to overly complex model structures and excessive parameter sizes. Despite our continuous adjustments in learning rates, regularization terms, and optimizer design during experiments, the training data on the MELD dataset exhibited severe overfitting. Therefore, we decided to simplify the model while maintaining context modeling and speaker embedding functions, thereby reducing model complexity and enhancing its generalization capability in practical applications. Specifically: we removed the bidirectional LSTM and the speaker memory mechanism.

1) Multi-round dialogue modeling: The original two-way LSTM network is removed and the [CLS] vector of each round of speech is directly used for context modeling. This simplifies the model structure and reduces the demand for computing resources.

2) Speaker memory: The speaker memory module and its dynamic update mechanism are removed, and the speaker_id is directly used to distinguish different speakers, instead of learning the dynamic memory of speakers through the model.

The simplified model retains support for multiple rounds of conversation and is optimized in the following ways:

1) Context modeling

In BatchLoaderBertData, context_size is set to 2, indicating that the current conversation and the contents of the first two rounds are input each time. The [CLS] vector of each round is extracted by BERT, and then the context information is directly concatenated to avoid the complexity and computational overhead caused by multi-layer LSTM processing.

2) Speaker embedding

speaker_id is retained, but the dynamic memory update mechanism is removed. During training, speaker_id is

mapped to an integer and concatenated with the text representation, which is processed by a fixed embedding vector.

3.1.1 Design of Ablation Experiment

Table 1.

Ablation module	**Brief description of functions**	**Experimental objective**	**Experimental setup**
Emotional dictionary embedding (VAD embedding)	The VAD 3D vector is concatenated into the text features of the BERT output (use_vad=True)	To verify the effect of emotional word dictionary information on text modeling	Set use_vad=False to disable VAD embedding
Consistency comparison learning of VAD between modes	The three modes are projected to the VAD space and then aligned by MSE loss	Verify the effect of modal consistency loss on fusion	Remove loss_vad_align from loss
Multi-round conversation context modeling	Patch the previous text Utterance and construct the context input (controlled by context_size)	Verify the impact of conversation history on emotion recognition	Setting context_size = 0 indicates no context

3.1.2 Comparative Experimental Design

In order to evaluate the performance of our proposed model in multimodal emotion recognition task, we conducted a comparative experiment with several existing models. In this experiment, we used the following benchmark models:

MUlT Model: MUlT is a multimodal learning model based on the Transformer architecture, focusing on capturing interactions between different modalities through cross-modal bidirectional attention mechanisms. It has performed well in multimodal sentiment analysis tasks, but it does not make many adaptive adjustments to information fusion between modalities. Therefore, when dealing with inconsistent modalities, it may encounter certain performance bottlenecks.

MISA Model: MISA aims to extract common and unique features across modalities by learning shared and private subspaces of each modality. When handling sentiment analysis tasks, the MISA model can effectively capture the diversity of emotions. However, its handling of inconsistent modal information is relatively fixed, failing to dynamically adjust the weights of different modalities, which may affect its robustness in complex data.

DEAN Model: DEAN combines multimodal BiLSTM and Transformer architectures to simulate the activation mechanisms of human emotions, making it particularly suitable for sentiment analysis tasks. Although DEAN can effectively capture temporal features of emotions, it does not handle conflicts between multimodal information well, leading to insufficient generalization capabilities across different data environments.

3.2 Result Analysis

3.2.1 Ablation Test Results

Table 2.

model	**Accuracy**	**Precision**	**F1_Score**
Complete body	**0.6318**	**0.6120**	**0.6159**
Remove the emotion dictionary embedding	0.5900	0.5823	0.5846
Remove contrast learning	0.5966	0.5905	0.5914
Remove multi-round conversation modeling	0.5981	0.5859	0.5870

3.2.2 Analysis of Ablation Test Results

After removing the sentiment dictionary embedding, all evaluation metrics of the model showed a significant decline, especially Accuracy dropping from 0.6318 to 0.5900, with F1_Score showing a larger decrease,

indicating that the sentiment dictionary is crucial for aligning the emotional features of the text. Precision fell to 0.5823, reflecting a decrease in model accuracy and an increase in misjudgment rates after removing the sentiment dictionary.

After removing the contrast learning, the accuracy and F1_Score of the model decreased, especially Precision slightly decreased to 0.5905, but the overall performance was still relatively stable. The impact of removing the contrast learning was small, indicating that the contribution of contrast learning in improving the performance of the model was relatively limited.

After removing the multi-turn dialogue modeling, the models accuracy and F1_Score slightly decreased. Accuracy dropped from 0.6318 to 0.5981, indicating that contextual information is less critical for sentiment recognition tasks compared to sentiment dictionary embedding and contrastive learning. Despite this, the model still maintained relatively stable performance after removing the multi-turn dialogue modeling.

3.2.3 Comparison of Experimental Results

Table 3.

method	**Acc-7 (%)**	**Precision**	**F1_Score**
MUL T	56.25	46.09	48.48
DEAN	55.56	40.78	45.44
MISA	57.13	42.16	47.83
This paper models	**0.6318**	**0.6120**	**0.6159**

3.2.4 Comparative Experimental Analysis

From the results of the comparative experiments, it can be seen that the model of this project outperforms other benchmark models in various metrics such as accuracy, precision, and F1-score. In particular, the models performance in accuracy and F1-score significantly surpasses that of models like MUlT, MISA, and DEAN, validating its effectiveness in multimodal sentiment recognition tasks. Specifically:

Accuracy: The accuracy of the model in this paper is 0.6318, significantly higher than the other three models. This indicates that our model can more accurately identify emotional types when distinguishing different emotional categories.

F1-score: The F1-score of the model in this paper is 0.6159, which is more balanced than other benchmark models, especially in multiple categories.

Accuracy: The accuracy and recall rate of the model in this paper are higher than that of other models, indicating that it has a strong ability to recognize various emotional categories when dealing with multimodal data, and can capture all categories well while avoiding misclassification.

The experimental results show that the model in this paper demonstrates excellent performance in multimodal sentiment recognition tasks by introducing the VAD sentiment lexicon and a multi-modal alignment mechanism. Compared to models such as MUlT, MISA, and DEAN, our model not only outperforms other methods in accuracy but also exhibits better adaptability in robustness and generalization. Therefore, our model can more effectively address the challenges of sentiment recognition in real-world scenarios.

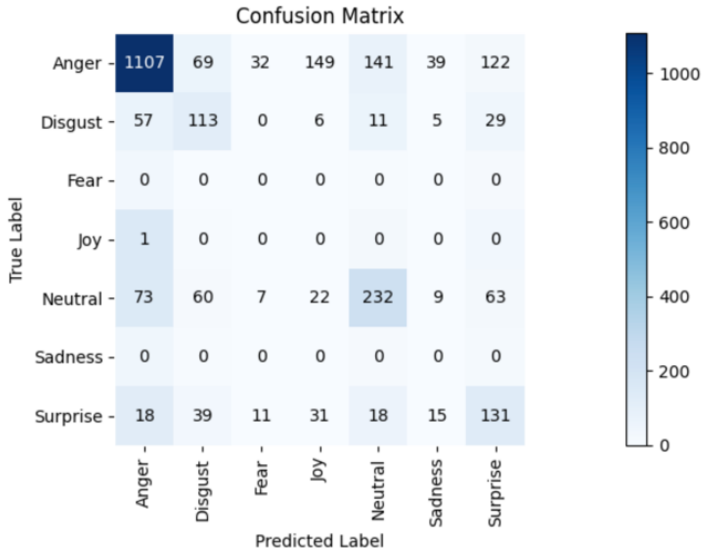

Figure 4. Confusion matrix for the first epoch

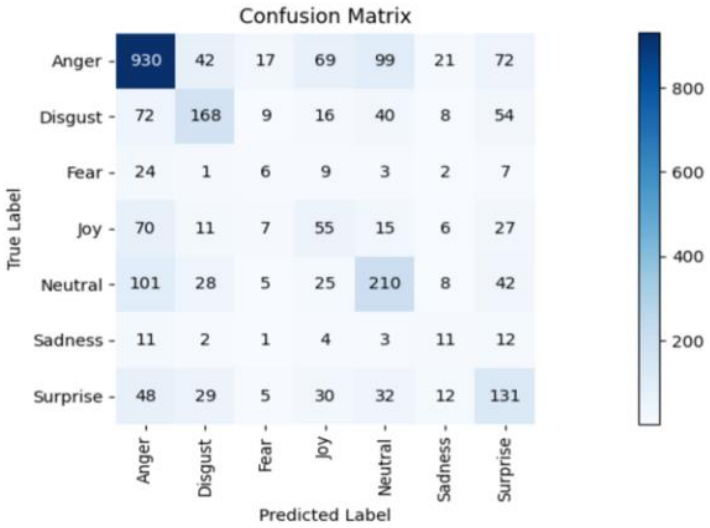

Figure 5. Confusion matrix at the end of the epoch after training

4. Conclusion and Outlook

This paper systematically studies and experiments on “a multimodal question-answering emotion recognition method based on user preferences.” By introducing domain-specific extended VAD sentiment dictionaries, sentiment preference-guided multimodal attention mechanisms, inter-modal VAD consistency contrast learning, and context-dependent modeling techniques, the models ability to understand and integrate multimodal emotional information is significantly enhanced. The ablation experiment results on the MELD dataset show that the sentiment dictionary embedding, inter-modal consistency contrast learning, and context modeling modules all contribute positively to overall performance improvement; comparative experiments demonstrate that this method outperforms advanced baseline models such as MULT, MISA, and DEAN in accuracy (63.18%), precision (61.20%), and F1 score (61.59%). These achievements fully validate the innovation and effectiveness of our projects method in the field of multimodal emotion recognition, while providing a theoretical foundation and practical reference for user emotion understanding in multimodal sentiment computing and online question-answering systems.

In our view, future work can start from the multi-task learning framework, combining emotion recognition and intention understanding tasks to build a more comprehensive dialogue understanding system.

References

Eyben, F., et al., (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*.

Liu, Y., et al., (2023). Conflict-Aware Cross-Modal Contrastive Learning for Multimodal Sentiment Analysis. *IEEE Transactions on Affective Computing*, *14*(2), 567-582.

Sun, H., Niu, Z., Wang, H., et al., (2025). Multimodal sentiment analysis with mutual information-based disentangled representation learning. *IEEE Transactions on Affective Computing*.

Yang, L., (2025). A dynamic weighted fusion model for multi-modal sentiment analysis. *IEEE Transactions on Affective Computing*, *16*(3), 1234-1248.

Zhang, K., et al., (2024). DialogueGNN: Context-Aware Emotion Propagation for Conversational Sentiment Analysis. *ACL*, 789-805.

Zhang, Y., Li, X., & Chen, W., (2024). Text-dominant strategy for multistage optimized modality fusion in multimodal sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *32*(1), 45-60.