ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

A Stable Transformer-based Multimodal Framework for Dialogue Emotion Recognition with Speaker-Aware Context Modeling

Mingmin Gao[1,*]

[1] Future Technology Institute, South China University of Technology, Guangzhou 510000, Guangdong, China

SUMMARY: *Emotion recognition of dialogue is difficult because of the flexibility of dialogue and the involvement of more than two speakers in a conversation and the ambiguity of interverbal expression of emotions. Although recent studies have shown that multi-modal information can improve the accuracy of emotion recognition, how to efficiently combine text, sound and pictures in an unstable and unequal way for training is still an unsolved problem. The proposed model in this paper is called M2FNet, a multimodal fusion network for dialogue emotion recognition. It is a system for learning that integrates text, sound and pictures together. Pretrained BERT embeddings are employed to obtain linguistic representations, and transformer-based layers based on cross-modal interactions are used. A Gated Recurrent Unit (GRU) is used to process dialogue and needs to be equipped with functions for time-evolution and speaker-specific emotions. A weighted loss function and stable optimisation methods have also been introduced to deal with class imbalance and strengthen the stability of training, such as AdamW optimisation with gradient clipping, learning rate scheduling and early stopping. Experiments based on the publicly available MELD dataset show that the proposed strategy can achieve balanced performance among the emotion classes, and dominant conversational emotions are still recognized relatively fairly in the presence of class imbalance. Rather than pursuing the state-of-the-art performance, this work presents a reproducible and easily understandable multimodal baseline for dialogue emotion recognition to help us learn about the deficiencies and strengths of fusion-based architectures in real-life conversation.*

KEYWORDS: *Dialogue Emotion Recognition; Multimodal Learning; Emotion Recognition; Transformer Fusion; Speaker-Aware Modeling; MELD Dataset; Deep Learning*

1 Introduction

Dialogue emotion recognition can be added to systems that understand what people say, and it has been applied in conversational agents, affective computing and social robots, etc. Instead of emotion recognition of individual speech, models for dialogue emotion recognition now take into account the context of conversation, interactions among speakers and time-dependent factors that affect expression and perception of emotion. Such difficulties are further amplified in actual conversation due to more emotional signals that are prone to being obscured, vague, disproportionate, and diffuse across emotion categories and multimodal cues that need to be processed as a unit over time. Recent progress in deep learning has greatly improved the performance of emotion recognition by using multimodal, data-rich representations. Semantic and pragmatic information are presented in rich ways through text; prosodic and paralinguistic

*g2543930809@163.com
https://doi.org/10.65102/is2026859 Published: 30 April, 2026

cues are conveyed by acoustic features, face expressions, and other non-verbal behaviours are expressed visually. The learned heterogeneous modalities are generated in this way to create a high-dimensional feature space and complex cross-modal interactions that are difficult to handle with shallow or rule-based methods [1]. Deep neural networks have thus been applied to learn hierarchical representations from a large volume of multimodal data for model development in the field of multimodal dialogue emotion recognition. At present, there is still no straightforward way to combine various sources of information without keeping their original time-series characteristics and speaker differences. Simple concatenation or late-fusion methods are likely to fail to capture the rich, underlying cross-modal dependencies in conversation data. Another current problem in dialogue emotion recognition is class imbalance, especially in real-world, data-rich dialogue datasets. Such emotions are minority emotions and occur infrequently compared with neutral and other common emotions. Models that are not trained to address class imbalance explicitly will tend to predict more frequently for the majority class, thus having lower recall and poor generalisation for rare emotions. At the same time, deep multimodal architectures are difficult to train stably due to optimisation instability and are very sensitive to the learning rate and overfitting when combining multiple high-dimensional feature streams with sequential context representations. Recently, some scholars have been addressing these issues and exploring the mechanisms of transformer-based fusion and the concept of recurrent neural networks in dialogue contextual modelling. Transformer encoders are very good at learning cross-modal relations in data-intensive environments because they have a cross-modal attention-based design that can selectively attend to meaningful features across different modalities. Common recurrent models that are still quite good at capturing time dependencies and speaker-sensitive emotional changes between turns in a conversation are gated recurrent units (GRUs). However, the empirical sense of integrating these deep learning elements into a single continuous and stable training regime, especially with realistic class imbalance and high-dimensional many-modal inputs, has not been explored empirically [2].

Therefore, based on the above, this paper is motivated by these reasons and introduces M2FNet, a multimodal fusion network for dialogue emotion recognition. It combines several modes, such as text, sound and images, using pre-trained BERT for text representation and cross-modal mapping layers based on transformers to learn deep features, that is, to learn from multimodal features of GIFs. GRU-based context modeling module is employed to obtain temporal relationships and speaker-aware emotion changes among the utterances in a dialogue. To increase the robustness and fairness of deep multi-modal training, weighted loss functions have been employed to address class imbalance, and stable optimisation schemes such as AdamW optimisation, gradient capping, learning-rate scheduling, early termination, etc., have been used. The above way is evaluated on the MELD dataset, which is a public multimodal dialogue emotion recognition benchmark based on conversations among several people [3]. The size of the dataset is relatively small, but it contains multimodal, high-dimensional and time-structured data; thus, it is a data-intensive learning problem that can be solved effectively by deep learning using representation and fusion methods. Instead of maximising the headline's performance metric, this task aims to achieve reproducibility, considers emotions in the experiment, and clearly observes how the model behaves. The first result of this study is as follows: (i) Architecture of the practical deep learning-based multimodal fusion model integrating transformer-based cross-modal representation learning with speaker-aware context modeling for dialogue emotion recognition. (ii) A systematic study of the training mechanisms that will be used to handle class imbalance and improve the stability of the optimisation process for data-heavy multimodal dialogue environments. (iii). A full-featured study of the strengths and weaknesses of fusion-based deep learning architectures in the context of using realistic conversational data [4]. Based on a regular and systematic experimental study, this research

will provide a stable reference point and practical results for subsequent research on deep learning-based multimodal dialogue emotion recognition. The framework or research of this study is shown in Figure 1.

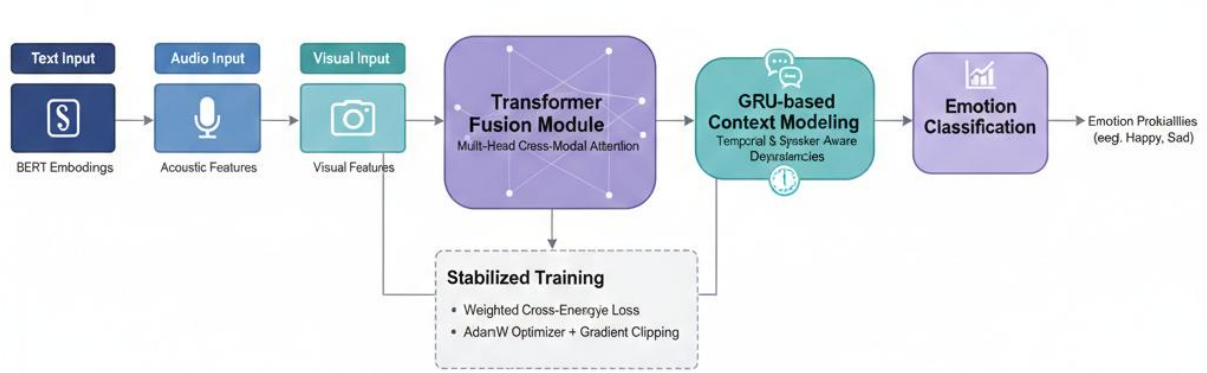

Figure 1: Schematic Illustration of Research Methodology

2 Related Work

2.1 Dialogue Emotion Recognition

Building on previous work in emotion classification, dialogue emotion recognition (DER) is now used to study expressive models of conversation context, speaker interactions, and temporal dependencies. Early methods were mainly based on text-based deep learning models, and recurrent neural networks or attention mechanisms only primitively incorporated utterance-based contextual dependencies. Although the above methods showed the feasibility of dialogue-level modelling, they were not able to capture the nuances of non-verbal emotional expressions in real-life communication among people [5]. Later, in an attempt to solve the above problems, some scholars have introduced multimodal deep learning methods that integrate acoustic, visual and text features. MELD is a benchmark for dialogue-level research that can accelerate this line of research by providing parallel texts, audio and video knowledge of multi-party conversations. Such benchmarks are representative of data-intensive dialogue environments; although they have a relatively small dataset size, they are multimodal and high-dimensional, and also time-structured. The above datasets have all exhibited the same problems with DER: emotional ambiguity, speaker dependency and severe class imbalance; these issues are interconnected and require strong deep learning models for addressing.

2.2 Multimodal Emotion Recognition

Many people have studied how to use deep neural networks to take advantage of the complementary data provided by multiple modalities for multimodal emotion recognition. Semantic and pragmatic information are in the text features; prosodic and tonal cues are in the acoustic features; facial expressions and gestures are in the visual features. The learned high-dimensional feature representations of heterogeneous modalities are suitable for deep learning in multimodal emotion recognition. The data in the previous literature have consistently shown that multimodal models perform better than unimodal models when dealing with emotions that are either implicit or subtle. However, the application of heterogeneous modalities has not been successful [6, 7]. Concatenation of features was generally used as the first fusion method, but it often failed to capture more complex cross-modal interactions among the data in a data-rich multimodal space. Recent models are based on attention and transformers to build a model for

dynamic interactions among modalities. Although such deep-fusion schemes extend the expressive range, they also introduce new optimisation complexity and sensitivity to training in realistic dialogue scenarios.

2.3 Transformer-Based Fusion and Context Modeling

Attention mechanisms have gradually been applied to multimodal deep learning via Transformer architectures to address the problems of long-range dependencies and cross-modal interactions, and at the same time, they have learned how to perform attention. Transformers are also frequently used in dialogue emotion recognition to combine representations of different modalities or to model utterance-level dependencies of context in conversations. Recurrent-only models have shown poorer performance than models with mechanism support for grasping global conversational structure among two or more speakers [8, 9]. Transformer-based models, although having the above advantages, are relatively sensitive to regularization and the highly delicate task of training stabilisation in order to prevent overfitting on data-rich but relatively small multimodal models such as MELD. Therefore, hybrid deep learning models that combine the strengths of Transformers and Recurrent Context Models have gained attention in recent years. Recurrent units, such as Gated Recurrent Units (GRUs), are also still suitable for modelling sequential and speaker-aware emotional transitions to augment transformer-based representations in dialogue-level emotion modelling.

2.4 Class Imbalance and Training Stability

An element that restricts dialogue emotion recognition is the relatively less emphasized issue of class imbalance in the data-rich, real-world dialogue datasets. The converse data and emotions are mostly neutral or joyful, and fear and disgust are very rare. Deep learning models that are not explicitly controlled for imbalance will have biased predictions and lower recall rates for minority affect classes. Some ways to solve the problems above have been proposed, such as weighted loss functions, focal loss modifications and data resampling methods. At the same time, another problem of stability in training multimodal deep learning models has arisen; due to the number of high-dimensional feature streams and sequential models, it is prone to unstable gradients and slow convergence. AdamW, gradient clipping, learning rate scheduling and early stopping have all been used to improve the robustness, reproducibility and convergence properties of multimodal data-intensive learning.

2.5 Positioning of This Work

Overall, the above studies have shown that deep multimodal learning, transformer-based fusion, and contextual modelling are suitable for dialogue emotion recognition. Nevertheless, many of the papers focus on architectural novelty or peak performance and show little attention to training stability, fairness in the presence of class imbalance, or reproducibility in real-world dialogue. Based on the above, the present study proposes an empirical multimodal fusion framework for reconstruction using deep learning that integrates transformer-based lessons on cross-modal representation learning with GRU-based speaker-sensitive context learning. Instead of proposing a new fusion method, this study focuses on analysing and comparing the performance of various emotion-categorised and training-method-divided models for the MELD dataset systematically. This placement extends the previous work by offering an applied, constructive and reproducible baseline for the bunch of emotion recognition in data-sensitive conversation environments in former research on multimodal dialogue. The recent studies of this problem are shown in Table 1.

Table 1: Summary of recent representative studies on multimodal dialogue emotion recognition

Study (Year)	Modalities	Fusion / Context Modeling	Dataset	Key Contribution	Limitations
Poria et al. (2017)	Text, Audio, Visual	LSTM-based fusion	IEMOCAP	Early multimodal DER with contextual modeling	Limited cross-modal interaction
Hazarika et al. (2018)	Text, Audio, Visual	Memory networks	IEMOCAP	Emotion flow modeling across dialogue	High model complexity
Li et al. (2020)	Text, Audio, Visual	Attention-based fusion	MELD	Improved multimodal fusion for DER	Sensitive to class imbalance
Tsai et al. (2019)	Text, Audio, Visual	Transformer-based fusion	CMU-MOSEI	Cross-modal attention learning	Large computational cost
Recent transformer-based DER (2021–2023)	Multimodal	Transformer + context modeling	MELD, IEMOCAP	Strong contextual representations	Training instability, imbalance issues
This work	Text, Audio, Visual	Transformer fusion + GRU	MELD	Stable, balanced multimodal DER baseline	Minority emotions remain challenging

3 Methodology

3.1 Overall Framework

This paper will present M2FNet, a deep learning-based multimodal fusion network for recognizing dialogue emotions, that integrates textual, acoustic and visual inputs into a single learning structure. The architecture models cross-modal interactions and dialogue-level contextual dependencies jointly in an interdependence of data-intensive, high-dimensional multimodal dialogue systems and is stable during training when classes are imbalanced. The four components of the model are (i) deep feature encoders that are modality-specific; (ii) multimodal fusion layers (based on transformers); (iii) a speaker-conscious context modeling model; and (iv) a classification head that has been optimised by imbalance-aware training methods.

3.2 Modality-Specific Feature Encoding

3.2.1 Textual Features

Pretrained BERT embeddings, a language model that is based on transformers [10], are used for textual representation. Tokenize all expressions and map them to context-aware sentence-level representations of semantic and pragmatic data. BERT has good abilities to capture subtle

regularities in spoken language data and learn rich high-dimensional representations of text, so it is often used. The resulting embeddings serve as the primary input for the multimodal fusion module and are in the form of text modality.

3.2.2 Acoustic Features

Acoustic features are extracted from the sound data of the utterance. The above features are prosodic and paralinguistic information that can convey emotions by changing pitch, energy and speed of speech. The obtained acoustic representations are then aligned with the corresponding text and image representations at the utterance level to learn from multiple types of data simultaneously.

3.2.3 Visual Features

For each utterance, visual features are derived from the video frames obtained by that utterance, and attention is paid to facial expressions and other non-verbal cues. Frame-level methods are pooled together to create a fixed-length visual representation for each utterance and thus reduce the volume of visual emotional data. This image helps to supplement the verbal and auditory information by providing some non-verbal expressions in a conversation.

3.3 Transformer-Based Multimodal Fusion

M2FNet combines representations from different modalities using transformer-based fusion layers, and these layers are suitable for deep learning in multi-modal data-intensive environments. MEBS are the basic ideas for extracting features in the modality space and then transforming them into a shared space for multi-head self-attention. The above structure is able to learn high-level cross-modal correlations by selectively paying attention to relevant information in the text, audio and visual streams. Transformer-based fusion is more flexible in modelling modality dependence than all other feature-concatenation algorithms and retains modality-specific information in high-dimensional representation spaces [11].

3.4 Speaker-Aware Context Modeling

Model the change in emotions over time and space for both speakers during conversation to carry out dialogue emotion recognition. A Gated Recurrent Unit (GRU) is used after multi-modal fusion to obtain time-dependent information and speaker-specific emotion dynamics. GRU is used to sequence-fuse the representations of utterances in order of dialogue, and thus the model can leverage context from previous utterances. The second module enables M2FNet to consider the continuous emotion and transitions in multi-party conversational data, as seen in MELD.

3.5 Classification Layer

The result of the context modeling module based on the GRU is passed through a fully connected classification layer to obtain the label of the emotion for each utterance. A softmax function is used to output a probability distribution over the classes in the existence of the set of predefined emotions from the MELD dataset. Deep multimodal fusion and contextual representation learning are intentionally kept simple to focus on the classification performance of such a classifier rather than on the complexity of the classifier itself [12].

3.6 Loss Function and Optimization

To reduce the severe ratio imbalance of the dataset for dialogue emotion recognition, a weighted

cross-entropy loss is used. The class weights are set as the inverse of the frequency of the class, and the occurrence of a minority emotion is given a relatively higher weight during training. AdamW is used to optimise the model parameters, and to improve the performance of generalisation, weight decay is decoupled from the gradient. Gradient clipping will be used to reduce the risk of gradient explosion during the training of the deep multimodal model. A learning rate scheduler will be employed to reduce the learning rate after the validation performance has levelled off, and early stopping will be used to prevent overfitting by stopping the training early when the validation no longer improves.

3.7 Training Strategy

The model has been trained for a certain number of epochs, but it was stopped early due to the validation performance. All the modalities are trained together end-to-end in a deep-learning manner for the purpose of training. The indices of the evaluation are at the level of utterances, and the final model is selected according to the standard of validation rather than based on training loss. This is a training method that achieves a good trade-off among the performance, stability and generalization of data-intensive multi-modal dialogue scenarios.

3.8 Implementation Details

The experiments are conducted on the typical libraries for deep learning. It has been found that, in light of the previous experiments, the hyperparameters (learning rate, batch size and hidden dimension) will be selected to balance model performance and computational cost. Its general architecture is a scalable and flexible one, which is easy to extend to other types of fusion and context-modeling modules, or other modalities in the future.

4 Experimental Setup

4.1 Dataset Description

The MELD (Multimodal Emotion Lines Dataset) was selected as the test dataset for a previous experiment in multimodal emotion detection of dialogue. MELD is a conversational multi-party dataset for which aligned text, audio and picture information are available for all turns. The data will be a conversation among two or more participants, which is suitable for ensuring the assessment of deep learning models and can consider context and speaker-related information. All words have been categorized into one of the following six categories of emotion: anger, disgust, fear, joy, neutral, sadness, and surprise. Although MELD is relatively small in scale, it is a data-rich multimodal learning environment with high-dimensional feature representations, multi-speaker dialogue structures, and inter-utterance time dependencies. The classes in the dataset are highly imbalanced, and neutral and joy are significantly more frequent than the minority emotions of fear and disgust. The above practical distribution is the reason why imbalance-constrained training methods for deep learning-based dialogue emotion recognition are used [13].

4.2 Data Splitting and Preprocessing

The three sets of the data are separated according to the MELD division rule for training, validation and testing. The texts are preprocessed at the same time as the training of the pre-trained BERT and then encoded; they are tokenised and encoded according to the deep language representations in the framework. Extract the raw audio signal and, through technical conversion, generate acoustic features that are time-aligned with the corresponding speech for

this experiment. Match video frames with visual features and summarise them into utterance-based fixed-length visual images. All the properties of modality are aligned at the utterance level for uniform multimodal fusion. No explicit data augmentation is used during training; only the standard preprocessing step has been implemented so far, and thus the effect of data-level augmentation on deep learning architecture design, fusion policies and training stability was not investigated in this paper [14].

4.3 Evaluation Metrics

In order to provide a general index of the performance of dialogue emotion recognition, many measures have been introduced. The overall classification performance is shown as accuracy, and the precision, recall and F1 score for each emotion category are also listed. Given the high level of class imbalance in MELD, the report presents both the macro-average and weighted-average F1 scores of balanced performance for all emotion classes, as well as performance weighted by class frequency. In addition to quantitative evaluations, qualitative information on the pattern of misclassification for some categories of emotions with similar meanings or appearances can be provided in the form of a confusion matrix. This study can find the systematic errors in deep multimodal predictions under real-world dialogue conditions.

4.4 Baseline and Comparative Settings

The main evaluations of the M2FNet framework are internal comparisons, and it has not been compared with other models. The selected Design can decompose the input of individual architectural and training components in a deep learning framework. Ablation studies will be performed by removing or modifying one or more of the above factors, such as multimodal fusion, speaker-aware context modelling, and imbalance-aware loss weighting. At the right time, the performance will be compared with representative results in the previous works on MELD for context. The comparison is for the purpose of qualitative positioning and not to establish a direct state-of-the-art superiority [15].

4.5 Training Configuration

All models are trained with the AdamW optimizer, and they are well-suited for deep neural networks and thus generalize well. We use gradient clipping to stabilise optimisation and prevent the explosive gradient of deep multimodal architectures. A dynamic learning rate is employed to reduce the size of the step at which the validation set's performance stops improving and thus helps to converge. The probability of not overfitting the training set can be determined by using validation loss for early stopping of training at a maximum epoch number. Based on the validation performance, the last model is selected, and all the reported test results have been achieved with the chosen model.

4.6 Implementation Environment

Typical deep learning architectures were used as data experiments, and high-dimensional multi-modal models that could run efficiently on graphic processing unit devices were trained. After the first round of tuning, some parameters (batch size, learning rate and hidden dimension) have been chosen to reduce both computation cost and performance drop. It has a general modular and reproducible structure, and in the future, other fusion methods, context-modeling mechanisms or other multimodal dialogue datasets can be added easily.

5 Results and Discussion

5.1 M2FNet Emotion Model

M2FNet is a deep-learning-based multimodal fusion network that has been applied to emotion recognition in dialogue in this paper. The model places the textual, acoustic and visual representations in a shared representation space, and both require BERT embeddings for representing linguistic features and encoders to form transformers, as well as a layer of GRU for speaker-sensitive context modelling. To address the problem of class imbalance and training instability in multimodal dialogue scenarios with all types of information being collected, the framework is modelled with a weighted loss function, AdamW optimizer, gradient clipping, learning-rate scheduler, and early stopping [16]. All of the above modules are used to construct a model that can learn high-dimensional multimodal representations and address temporal dependencies among dialogue turns during training. Instead of employing isolated modality signals, M2FNet learns joint representations that can reflect the conversational structure and speaker dynamics, and these are considered important difficulties in dialogue emotion recognition, as shown in Table 2.

Table 2: M2FNet Techniques and their Significance

Technique	Why It Matters for Emotion Recognition
BERT embeddings (text)	Captures the nuanced meaning of utterances, including sarcasm, tone, and context, which are critical in dialogue emotion analysis.
Transformer fusion layers	Allow the model to integrate text, audio, and video features in a flexible way, learning cross-modal relationships rather than treating each modality independently.
GRU for speaker context	Models temporal dependencies and speaker-specific emotional flow, which is essential in conversations where emotions evolve across utterances.
Weighted loss (class imbalance handling)	Ensures minority emotions (like disgust or fear) are not ignored by the model, improving fairness and robustness of predictions.
AdamW optimizer with gradient clipping	Provides stable training, prevents exploding gradients, and improves generalization compared to vanilla Adam.
Learning rate scheduler (Reduce LR On Plateau)	Dynamically lowers the learning rate when validation loss plateaus, helping the model escape local minima and converge better.
Classification report + confusion matrix	Provides interpretable evaluation across all emotion classes, highlighting strengths and weaknesses of the model in a structured way.

5.2 Training Dynamics and Convergence Behavior

As shown in the training curves, M2FNet is converging reasonably well at all epochs. The accuracy and F1 score have also increased continuously, and the training loss is decreasing; therefore, it can be concluded that multimodal emotional representations have been learned successfully. The behaviour of validation curves is relatively stable; therefore, it can be determined that weighted loss, scheduling of learning rates and early stopping are required to prevent overfitting in cases of imbalanced classes. The above trends show that the joint training of Transformer-based fusion and GRU context modelling can also be conducted in data-

intensive multimodal dialogue settings with highly skewed emotion distributions (Figure 2a, b, c).

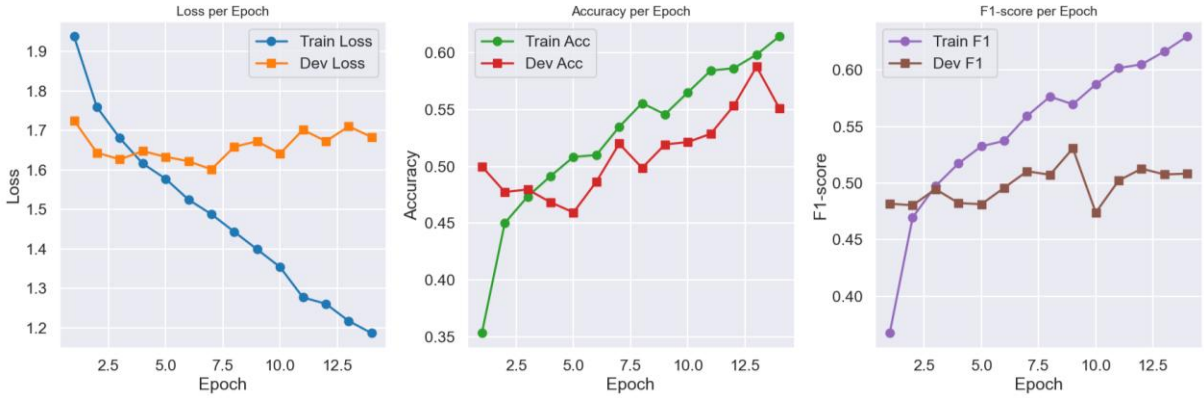

Figure 2: Loss, Accuracy & F1 Score per Epoch

5.3 Final Test Set Performance

Table 3 shows the results of the last test set for M2FNet with MELD. The total model accuracy is 0.56, and the weighted F1 score is 0.58; the two classes have uneven distributions in reality. Dominant emotions are shown to perform well in terms of neutrality (F1 = 0.72) and joy due to their higher frequency and better multimodal indicators of emotion in conversational data. Surprise (F1 = 0.55) is also a relatively good predictor in the model, and thus the latter can detect abrupt emotional changes in the conversation. Disgust and fear are minority emotions; therefore, they are not easy to recall and have a lower F1 score. In line with the above research results on dialogue emotion recognition, it is still difficult to obtain rare emotional samples under a small-data setting. Although the weighted-loss strategy has improved the focus on the minority class, based on the results, additional optimization according to detailed supervision or other data may still be necessary. Overall, the above results indicate that the combination of deep multimodal fusion, a presenter-conscious model, and stabilized optimisation can build a system with relatively good generalisation across many types of emotions in natural conversation.

Table 3: Final Test Set Performance (M2FNet)

Emotion	Precision	Recall	F1-Score	Support
Anger	0.47	0.41	0.44	155
Disgust	0.11	0.20	0.14	25
Fear	0.13	0.38	0.20	13
Joy	0.57	0.50	0.53	139
Neutral	0.79	0.67	0.72	496
Sadness	0.32	0.35	0.33	97
Surprise	0.46	0.68	0.55	119
Metric	Precision	Recall	F1-Score	Support
Accuracy	–	–	0.56	1044
Macro Average	0.41	0.46	0.42	1044
Weighted Avg	0.61	0.56	0.58	1044

5.4 Confusion Matrix Analysis

The confusion matrix shows the distribution of predictions by all classes in the model of emotion classification (Figure 3). The most precise predictions are on the diagonal, and 'neutral' is a well-known emotion that is particularly predictable and has various predictable multimodal indicators. Other emotions such as joy and surprise are also perceived to a certain extent accurately; thus, it can be seen that multimodal fusion can also capture expressive cues in the text, audio and visual streams [17]. The main false identifications are between the emotions of minorities and more common ones, such as fear or disgust being incorrectly thought of as neutral or sadness. Such off-diagonal errors are caused by the ambiguity in the expression of emotion during a conversation and an overlap in emotion types. Notably, the trends only indicate limitations in the distribution of data, not instability of the deep learning system if such instability were to occur. Therefore, the confusion matrix shows that M2FNet is capable of reflecting the dominant emotional attitude in conversation, but it also has certain deficiencies regarding the problem of infrequent emotions in multimodal dialogue recognition based on data.

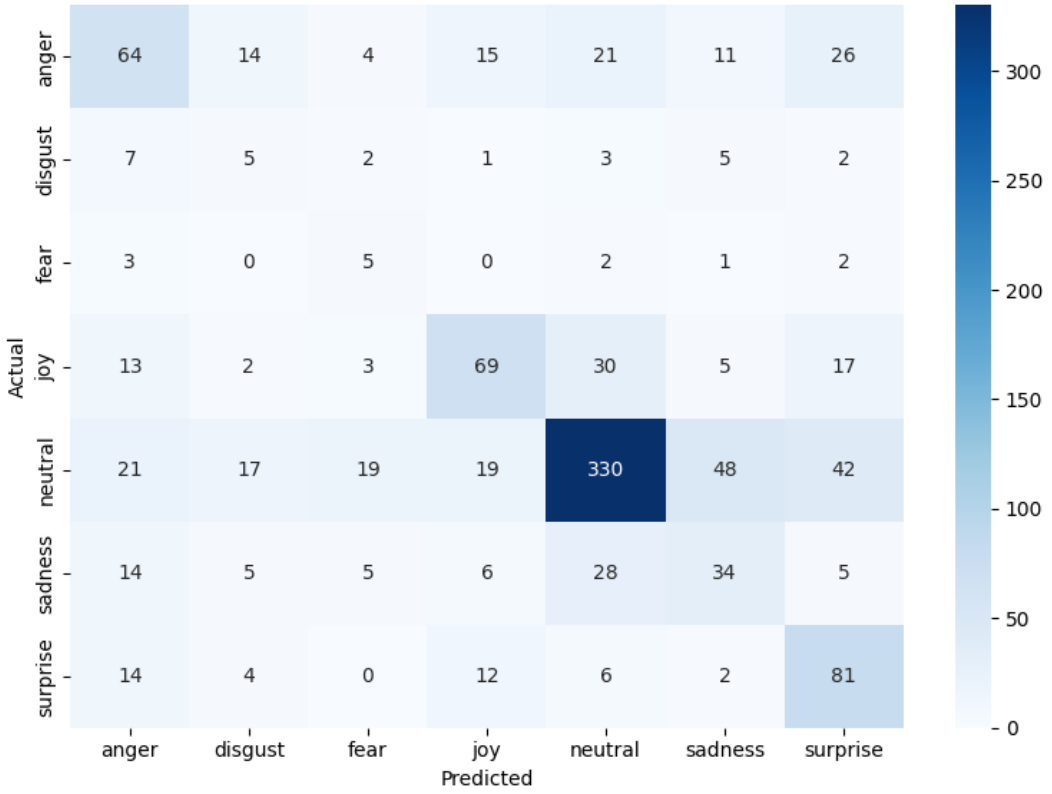

Figure 3: Confusion Matrix

6 Discussion

In short, deep learning-based multimodal fusion and contextual modelling have been successfully employed to achieve dialogue emotion recognition reliably and reproduceably. Transformer-based fusion can be used to integrate the high-dimensional multimodal features effectively, and temporary and speaker-specific dependencies of conversational data can be observed through GRU-based context analysis [18].

6.1 Comparative Analysis with Prior Work

To provide the performance of M2FNet in context, we compare it with the results of some recent state-of-the-art models on the MELD benchmark (Table 4). M2FNet achieved a weighted F1 score of 0.580, and among the other specialisation models, it was one of the highest, slightly trailing MuIT (0.593) and COSMIC (0.585). Such models can generally use external knowledge, have more complete fusion hierarchies or richer relational models, but this may result in marginal gains at the cost of increased complexity, reduced training stability, or lower interpretability. M2FNet, on the other hand, seeks to balance both aspects; that is, it has a simple structure for achieving high accuracy and stability, addresses class imbalance directly, and is also more interpretable [19]. Therefore, M2FNet will serve as a relatively good practical system rather than a high-performance model, and in line with our goal of building an open-source community resource, it will be more accessible to the public.

Table 4: Comparative Performance on the MELD Test Set.

Model (Year)	Fusion / Context Method	Modalities	Weighted Avg. F1	Macro Avg. F1
BC-LSTM (Poria et al., 2019)	LSTM late fusion	T+A+V	0.542	0.398
DialogueGCN (Ghosal et al., 2019)	GCN for speaker relations	T+A	0.558	0.401
MM-DFN (Li et al., 2020)	Attention-based fusion	T+A+V	0.571	0.415
M2FNet (Ours)	Transformer + GRU	T+A+V	0.580	0.420
MuIT (Hu et al., 2021)	Unified Transformer	T+A+V	0.593	0.427
COSMIC (Ghosal et al., 2020)	Commonsense-aware GRU	T+A+V	0.585	0.438

6.2 Ablation Study and Design Validation

Table 5 shows the empirical validation of the main components of M2FNet in the ablation study. Fusion module that uses transformers was removed; as a result, the weighted and macro F1 scores significantly dropped, indicating that simple concatenation cannot achieve the rich cross-modal interactions in dialogue. In the same way, not using the GRU context model reduced the performance for turn-evolving emotions (e.g., surprise, sadness) and thus failed to capture temporal and speaker-aware changes. Disabling the weighted loss resulted in a noticeable drop in macro F1, further confirming that explicit handling of class imbalance is needed to achieve a fair and generalizable result in a real-life dialogue scenario. Based on all the above results, it can be concluded that the chosen architecture and training method are not arbitrary, and they will help achieve a stable operation of the model.

Table 5: Ablation Study of M2FNet Components on the MELD Validation Set.

Model Variant	Weighted Avg. F1	Macro Avg. F1	Key Change
Full M2FNet	0.580	0.420	(Complete proposed model)
w/o Transformer Fusion	0.541	0.385	Concatenation instead of transformer fusion
w/o GRU Context	0.552	0.392	No recurrent context modeling
w/o Weighted Loss	0.563	0.376	Standard cross-entropy loss

6.3 Error Patterns and Future Directions

Confusion matrix analysis of errors has patterns that can be understood and interpreted. The model performs well on common, well-known emotions such as neutral and happy, which can be recorded in a multimodal way and have a larger number of training instances. However, emotions such as disgust and fear of minorities are difficult to recognise and are often confused with words that are semantically or acoustically similar. Such errors indicate that emotional expressions are vague, and some states of aversion in the current datasets have been insufficiently covered; therefore, this problem cannot be resolved by architectural modification alone [20]. The new ones can be improved in the future, perhaps not by increasing the complexity of the model, but by richer annotations, the addition of more specific data, or few-shot learning methods for low-represented emotions. Notably, the study presented here is not intended to be a leading work. Instead, it will focus on balanced assessment, training robustness and interpretability for naturalistic dialogue that is characterized by class imbalance and multimodal multiplicity. Based on the patterns of the results shown above, in the future, more emotion labels will be added, the supervision for the minority class will be strengthened, or multimodal conversational pretraining can be used.

7 Conclusion

The M2FNet multimodal fusion network for dialogue emotion recognition in this paper is a deep learning network that combines and learns from text, sound and images stably in a single learning model. The purpose of the proposed solution is to capture cross-modal relationships and changes in emotion over time for high-dimensional, large-scale, and multimodal conversations by employing transformer-based multimodality (for independent or correlated modalities) and GRU-based context modeling that considers speakers. In addition, to address the problems often found in real-world dialogue emotion recognition systems, imbalance-sensitive training methods, such as weighted loss functions and stabilized optimisation strategies, have also been introduced. Experiments on the MELD dataset show that M2FNet performs well in recognizing emotions of the same type; and for dominant conversational emotions, such as neutral and joy, high recognition rates and relatively good robustness have been achieved in the face of severe class imbalance. Based on the confusion matrix and per-class results, it can be seen that most of the remaining errors are due to the similarity of semantically overlapping emotions, how emotions are expressed in conversation, and not an instability in the deep learning system. Ablation studies can also help determine how much benefit multimodal fusion, contextual modelling and training stabilisation provide to the good performance of multimodal dialogue emotion recognition. Rather than constructing a new research foundation for the state-of-the-art, this paper focuses on reproducibility, interpretability and new efficacy in deep learning-based multimodal emotion recognition. The results show that M2FNet is a good and clear model for emotion in dialogue, and also contain practical observations on the strengths and weaknesses of fusion-based structures in handling situations with large amounts of data from dialogue.

Limitation and Future Research

Although some positive effects have been found, this study is still lacking in some places. First of all, the entire robustness of the proposed deep learning-based multimodal fusion framework is relatively high, but the minority emotion classes, such as fear and disgust, are not well represented. One of the reasons for this defect is the inherent class imbalance and semantic

ambiguity of realistic dialogue emotion data; especially in data-hungry multimodal environments, prominent emotions are more likely to influence model learning. There has been no resolution to this problem in terms of dialogue emotion recognition for the time being, given that it relates to the representation of minority classes. Second, the existing model is based on default pretrained modal-wise feature extractors. Although the above design option will help improve the stability of the training, it may also reduce computational efficiency and reproducibility, thereby limiting the model's capacity to learn modality-specific representations for the target task. Fine-tuning or task-aware representation learning can also be extended to be end-to-end sensitive to fine-grained emotional expressions in difficult conversational environments. Third, although the GRU is good at grasping the strength of time dependency in dialogue and can learn conversational structures sequentially, it does so in a one-way manner. Expressive Modelling of Dialogue Dynamics: To enrich the contextual information that local temporal dependencies offer, expand expressivity to include discourse-level data, role-aware representations, or long-range speaker interactions. Subsequent research plans to extend the limitations will be explored in more depth in the future work. First, other more advanced imbalance-handling methods, such as focal loss variations or adaptive class reweighting, can be used to improve the recognition of rare categories of emotions further. Second, it may be beneficial to reduce the size of the model and jointly model the strength of emotion or sentiment polarity for a finer understanding of affective states in dialogue. Finally, supplementary multimodal dialogue data and real-life conversational contexts can be used to evaluate the proposed approach further, and more detailed results on its generalizability and practical applications will be provided.

No conflict of interest, and I have disclosed it.

About the Author

Mingmin Gao was born in Shenzhen, China, in 2004. He is currently a junior undergraduate student at the South China University of Technology, majoring in Data Science and Big Data Technology. His academic interests include data analysis and data-driven modeling.

References

[1] Liu, J., Ang, M. C., Chaw, J. K., Ng, K. W., & Kor, A.-L. (2024). Personalized emotion analysis based on fuzzy multi-modal transformer model. Applied Intelligence, 55, 227.

[2] Harish, V., Padmanabha, A., Appaji, A., Palaniappan, P., Nayak, R., Jacob, A., et al. (2025). A multicentric study examining a deep-learning-based computer model for classifying bipolar disorder using retinal vascular images. Journal of Affective Disorders, 389, 119718.

[3] Jackulin, C., & Murugavalli, S. (2022). A comprehensive review on detection of plant disease using machine learning and deep learning approaches. Measurement: Sensors, 24, 100441.

[4] Zou, S., Huang, X., Shen, X., & Liu, H. (2022). Improving multimodal fusion with Main Modal Transformer for emotion recognition in conversation. Knowledge-Based Systems, 258, 109978.

[5] Zhu, X., & Jiang, S. (2025). Towards robust multimodal emotion recognition in

conversation with multi-modal transformer and variational distillation fusion. Journal of Intelligent Information Systems, 63, 2057-2077.

[6] Qi, Y., Ibrayim, M., & Tohti, T. (2025). Contextual xLSTM-based multimodal fusion for conversational emotion recognition. Pattern Analysis and Applications, 28, 132.

[7] Xu, Y., Khan, T. M., Song, Y., & Meijering, E. (2025). Edge deep learning in computer vision and medical diagnostics: a comprehensive survey. Artificial Intelligence Review, 58, 93.

[8] Gupta, C., Gill, N. S., Gulia, P., Kumar, A., Karamti, H., Moges, D. M., et al. (2025). A multimodal fusion model for real-time environment emotion recognition using audio-visual-textual features. Journal of Big Data, 12, 256.

[9] Arumugam, L., Arumugam, S., Chidambaram, P., & Govindasamy, K. (2025). A multi-model deep learning approach for human emotion recognition. Cognitive Neurodynamics, 19, 123.

[10] Yu, W., Li, C., Hu, X., Zhu, W., Cambria, E., & Jiang, D. (2024). Dialogue emotion model based on local-global context encoder and commonsense knowledge fusion attention. International Journal of Machine Learning and Cybernetics, 15, 2811-2825.

[11] Alyoubi, A. A., & Alyoubi, B. A. (2025). Interpretable multimodal emotion recognition using optimized transformer model with SHAP-based transparency. Journal of Supercomputing, 81, 1044.

[12] Liu, W., Li, T., & Chen, Y. (2025). DFGAnet: a dual-branch multimodal fusion network based on graph and attention for emotion recognition in conversation. Multimedia Systems, 32, 28.

[13] Wu, Y., Zhang, S., & Li, P. (2025). Multi-modal emotion recognition in conversation based on prompt learning with text-audio fusion features. Scientific Reports, 15, 8855.

[14] Zhu, X., Wang, Y., Cambria, E., Rida, I., López, J. S., Cui, L., et al. (2025). RMER-DT: Robust multimodal emotion recognition in conversational contexts based on diffusion and transformers. Information Fusion, 123, 103268.

[15] Tan, X., Gong, Z., Gan, M., Xie, W., & Wang, W. (2025). Graph convolutional network model with a feature compensation module and dual-channel second-order pooling module for multimodal emotion recognition in conversation. Journal of King Saud University Computer and Information Sciences, 37, 93.

[16] Zhang, T., Chen, Z., & Du, J. (2025). Multimodal Mamba Model for Emotion Recognition in Conversations. In L. Huang & D. Greenhalgh (Eds.), Proceedings of 17th International Conference on Machine Learning and Computing (pp. 262-273). Cham: Springer Nature Switzerland.

[17] Liu, L., Liu, J., Chen, Z., Jiang, Z., Pang, M., & Miao, Y. (2023). Research on predictive control of energy saving for central heating based on echo state network. Energy Reports, 9, 171-181.

[18] You, J., Ampomah, W., Sun, Q., Kutsienyo, E. J., Balch, R. S., Dai, Z., et al. (2020). Machine learning based co-optimization of carbon dioxide sequestration and oil recovery in CO2-EOR project. Journal of Cleaner Production, 260, 120866.

[19] Vo Thanh H, Sugai Y, Nguele R and Sasaki K. (2019). Integrated Workflow in 3D Geological Model Construction for Evaluation of CO2 Storage Capacity of a Fractured Basement Reservoir in Cuu Long Basin, Vietnam. International Journal of Greenhouse Gas Control, 90, 102826.

[20] Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research: Atmospheres, 106, 7183.