Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluation
DOI: 10.54254/2755-2721/2025.PO24691

Exploring the Absence of Modality in Multimodal Sentiment Analysis and Practical Applications

Ruohan Chen

Maynooth International Engineering College, Fuzhou University, Fuzhou, China
832304201@fzu.edu.cn

Abstract. As an important branch of sentiment analysis, multimodal sentiment analysis incorporates multiple modal information, but the problem of missing modalities is common in practical applications, which affects the accuracy and completeness of the analysis. From the perspective of improving model performance and practice, it is of great significance to solve the modal missing problem. Therefore, this paper summarizes the methods to solve the problem of missing modality in multimodal sentiment analysis, and classifies seven models into three categories according to the framework and characteristics: generative approach, joint learning approach, and the combined approach, and focuses on analyzing the advantages and disadvantages of each method. In addition, this paper mainly studies the different solutions to the missing modality in three applications: classroom teacher sentiment analysis, child sentiment analysis, and depression assessment. However, the existing models have problems such as insufficient generalization ability and limited ability to deal with complex missing modes, and this paper suggests that they can be solved by using data enhancement techniques to expand data diversity, applying adaptive model structure, and combining meta-learning and transfer learning.

Keywords: Multimodal Sentiment Analysis, Missing modality, Generative Approach, Joint Learning.

1. Introduction

One significant subfield of sentiment analysis is multimodal sentiment analysis, it can integrate text, image, speech, and other modal information, and can capture sentiment expression more comprehensively. However, in practical applications, the problem of missing modalities is prevalent due to data acquisition or environmental limitations. This modal absence can seriously affect the accuracy and completeness of sentiment analysis and lead to bias in sentiment judgment. Therefore, how to effectively solve the modal missing problem in multimodal sentiment analysis has become one of the hotspots and difficulties in current research.

From a practical point of view, solving the modal missing problem can improve the applicability and accuracy of sentiment analysis in different scenarios. For example, in the field of education, it can be used to understand students' emotional changes and formulate personalized teaching, in the field of mental health, it can provide timely treatment solutions for patients, and in the field of cinema, it can help filmmakers to recover damaged films.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.PO24691

In 2017, Poria et al. presented an overview of sentiment computing from unimodal to multimodal fusion, providing some explanations on some basic feature extraction methods and model frameworks for sentiment analysis, and suggesting some improvements in the potential performance of multimodal sentiment analysis [1]. In the same year, Soleymani et al. defined the problem of sentiment and multimodal sentiment analysis and summarized the recent progress of multimodal sentiment analysis in different fields [2]. In 2024, Guo Xu et al. summarized the algorithms of sentiment analysis for multimedia fusion and also introduced the current development of multimodal sentiment analysis [3]. However, few analytical summaries focus on solving the modal absence problem.

Therefore, this paper aims to explore the solutions to the modal absence problem in multimodal sentiment analysis, so that the technique can be better applied in practice. The main work is to organize the typical techniques and corresponding data sets of different solution methods, classify them according to their characteristics, analyze the advantages and disadvantages of each model, in addition, to analyze and summarize the models that have been put into practical application. Finally, innovative methods are proposed to solve the problems of existing models.

2. Typical technical analysis

The current methods are mainly categorized into two types: joint learning and generative methods. Joint learning learns joint representations based on the relationships between different modalities, attempting to mine the intrinsic connections between modalities to achieve effective fusion and representation learning of multimodal data. The generative approach aims to learn and synthesize novel data that statistically align with the distribution of the original dataset and generate new data that conforms to the distribution by learning from the existing data distribution to fill in the data of the missing modalities. By reviewing the literature, this paper summarizes the following models into three categories: generative methods, federated learning, and a combination of both, organized in the table in Table 1.

2.1. Joint learning approach

For models applying a joint learning approach, there are research directions such as modal coding and shared representation learning, adaptive network architecture, knowledge distillation, and transfer learning. This paper focuses on Tag-Assisted Transformer Encoder (TATE), Robust Multimodal Missing Signal Framework (RMSF), MissModal learning approach, and Unifed multimodal Missing modality self-Distillation Framework (UMDF).

The TATE module, which labels missing modalities with 4-digit numbers, covers both unimodal and multimodal missing cases and aids joint representation learning [4]. Two datasets, CMU-MOSI and IEMOCAP, are applied in the study, and the experimental results show that TATE shows significant improvement in Accuracy and macro-F1 metrics compared to multiple baseline models when dealing with unimodal and multimodal missing cases. However, the drawback is that the experiments found that text modality dominates the multimodal sentiment analysis, and the model performance decreases dramatically when text modality is missing.

The RMSF model proposes a hierarchical cross-modal interaction module that mines potential complementary semantics between missing modalities using coarse- and fine-grained cross-modal attention mechanisms [5]. When tested on two datasets, CMU-MOSI and IEMOCAP, it is found that the RMSF containing linguistic modalities performs the best in the case of bimodal deletion, while the combination of linguistic and audio modalities is comparable to the performance of the complete

DOI: 10.54254/2755-2721/2025.PO24691

modality in the case of unimodal deletion. However, it is not validated in real application scenarios. There are a lot of models and they are not lightweight enough.

MissModal learning approach significantly reduces computational complexity by simplifying the multimodal fusion network compared to generative models [6]. However, the model parameters of MissModal depend on the complexity of the multimodal fusion network and the number of modalities, which may pose the problem of increased model complexity in downstream applications of the approach.

UMDF designs a unified self-distillation mechanism, a multi-granularity cross-modal interaction module, and a dynamic feature integration module for the uncertain modal missing problem, and applies three datasets, namely MOSI, MOSEI, and IEMOCAP [7]. The experimental results show that UMDF significantly improves the performance of multimodal sentiment analysis in dealing with both modal missing and complete modal test conditions. However, the self-distillation mechanism and multi-module design increase the training complexity, the hyperparameters are difficult to adjust, and overfitting may occur on small datasets.

2.2. Generative approach

The core idea of the generative approach is to enable the model to maintain its performance in the inference phase even in the face of partial modality missing by generating or completing the data for the missing modalities. In this paper, two models, the Similar Modality Completion based-MSA model (SMCMSA) and the Modality Translation-based MSA model (MTMSA), will be explored.

SMCMSA constructs a complete modal sample database and completes the missing modes with similar modes [8]. The effectiveness and superiority of the model are demonstrated in the unimodal missing-mode experiments, multimodal missing-mode experiments, ablation experiments, and multi-type experiments. For example, in the unimodal missing experiment on the CMU-MOSI dataset, the SMCMSA model achieves the best results in accuracy (Acc) and macro-averaged F1 value (M-F1) when the missing rate is set to 0.2, 0.3, 0.4, and 0.5. However, when dealing with large-scale data, it is computationally expensive to construct the database and find similar samples, and the model performance depends on the database quality.

The MTMSA model though may lose some of the original modality-specific information in the process of converting visual and audio modalities to textual modalities [9]. However, the model is innovative in converting visual and audio modalities to textual modalities to enhance modal quality and fill in missing modalities. Ablation experiments show that textual modalities work best in unimodal analysis, bimodal combinations that include textual modalities work better, and three modalities work best when used simultaneously.

2.3. The combined approach

The Unified Multimodal Framework (UniMF) is both a generative and a joint learning approach that utilizes a translation module to deal with missing modalities, followed by a prediction module to receive the complete multimodal sequence generated by the translation module [10]. The translation module's central, Multimodal Generation Transformer (MGT), is based on the Multimodal Generation Mask (MGM) attention mechanism, which introduces [multi] tokens that are used in inference to fuse existing modal information to generate missing modalities. MGT performs layer normalization, MGM manipulation, and feed-forward network processing on the inputs, such as the input of linguistic and audio modalities, to generate data similar to the original video modalities, which compensates for the modal deficiencies. The core of the prediction module is the Multimodal

DOI: 10.54254/2755-2721/2025.PO24691

Understanding Transformer (MUT), which relies on the Multimodal Understanding Mask (MUM) attention mechanism and MultiModalSequence (MMSeq) to enhance the multimodal representation; the MUM exchanges information and fuses MMSeq with each unimodal modality through the MASKU matrix, and the MUT processes the multimodal information through layer normalization, MGM manipulation, and feed-forward network, and then outputs the sentiment analysis results from the fully connected layer. The processing power is strong and competitive or leading results are achieved on multiple datasets. But it also suffers from model complexity.

Table 1: Collation of typical methods, datasets, and shortcomings for the seven research models

Model Types	Model names	Typical methods	Data sets	Drawbacks
Joint learning approach	The TATE module	Designing a label encoding module, adopting a new projection model	CMU-MOSI、IEMOCAP	Non-optimal performance in some settings; relies on label encoding and projection mode.
	The RMSF model	Hierarchical cross-modal interaction; adaptive feature refinement; knowledge integration for self-distillation	MOSI、MOSEI	Complex model structure; high computational resource requirements
	MissModal learning approach	Multi-loss alignment representation; no change in the fusion phase	CMU-MOSI、CMU-MOSEI	Parameter dependent on network and number of modalities; limited by dataset size
	UMDF	Unified Self-Distillation; Multi-Granularity Cross-Modal Interaction; Dynamic Feature Integration	MOSI、MOSEI、IEMOCAP	Complex model structure; risk of overfitting for small datasets
Generative approach	MTMSA model	Modal translation: pre-trained model supervision	CMU-MOSI、IEMOCAP	Poor performance of some metrics at low deletion rates; reliance on pre-trained models
	SMCMSA model	Similar Modal Completion Strategies: Textual Modal Fusion	CMU-MOSI、IEMOCAP	High computational costs; reliance on database quality
The combined approach	UniMF	Contains translation and prediction modules; introduces MGM, MGT, MUT, and MMSeq	CMU-MOSI、CMU-MOSEI、MELD、UR-FUNNY	Generation modality may vary; limited ability to handle complex, unaligned sequences.

3. Application scenario study

The solution to the missing modality problem is ultimately to be applied to real life, this paper focuses on three applications: classroom teacher sentiment analysis, children's sentiment analysis, and depression prognosis.

3.1. Classroom teacher sentiment analysis

Teachers are crucial in education. Their emotional input can promote the cognitive and emotional co-development of students and improve the quality of teaching. Under the trend of intelligent education, teacher sentiment analysis with the help of multimodal sentiment analysis has become an important way of teaching intelligent evaluation. However, the practical application also has the problem that audio and visual modalities are susceptible to the noise interference and lighting

DOI: 10.54254/2755-2721/2025.PO24691

conditions of the environment in sentiment analysis, etc. To solve this problem in the future, Wu Zheng introduces textual modalities and designs a fusion method based on the cross-attention mechanism [11]. The accuracy on the CMU-MOSI dataset reaches 85.67%, and the accuracy on the self-constructed dataset is improved by 2.44% compared with the model using only audio-visual modalities. The results show that the model performance is relatively stable when the modal weights change in a certain range, and the model achieves excellent results on multiple datasets.

3.2. Children's emotional analysis

As children are a vulnerable group in society, even a slight shock or mood swings may affect their mental health and hinder their normal growth and development. According to the Autism Population Report released by the Centers for Disease Control (CDC), the incidence of autism is rapidly increasing year by year [12]. Wenhao Wu researched this and proposed the development and application of algorithms for children's emotion analysis from facial expression and multimodal information fusion [13]. Due to children's weak language ability, they are unable to accurately express their physical and emotional states in words as adults do, the researcher proposed a facial expression recognition algorithm based on the attention mechanism and explicit recognition to assist multimodal fusion, and in the pain assessment of neonatal scenarios, the average accuracy of the explicit recognition method and the attention mechanism method in the two- and four-classification tests were 96.7% and 86.1%, which indicates that these two algorithms perform well in handling the task of recognizing children's facial expressions. As for the problem that children's emotional features are not obvious and the complexity of the scene leads to the interference or even failure of some modal features in the actual scene, the thesis adopts a hybrid fusion strategy, and the accuracy of the threshold-based decision-level fusion method is as high as 97.9% in the evaluation of the overall monitoring results.

3.3. Depression prediction

Depression is a widespread mental illness that affects people across the globe, spanning all ages and primarily affecting adults. This bipolar disorder is characterized by symptoms such as pessimism, hopelessness, lack of pleasure, and sadness, which significantly affect life and leads to depression. Therefore, the prediction and estimation of depression is an important area of research. Sharma et al. proposed a multimodal analysis approach for depression detection using multimodal fusion that can complement information to improve detection accuracy while an intelligent chatbot can guide the user to provide more information to compensate for missing modalities [14]. The results show that compared with the traditional unimodal method, the whole multimodal system achieves an accuracy of 94.22%, a precision of 90.41%, a recall of 93.20%, and an F1 score of 91.87%, which verifies its effectiveness in depression prediction.

4. Challenges and solutions

By analyzing these seven models and application scenario studies, Approaches to solving the problem of missing modality are better understood, and at the same time better understand the future challenges and research directions, for example, most of the existing methods to cope with missing modules are trained and tested in specific datasets and experimental conditions, and the models, when faced with different scenarios and types of data with missing modules, have a generalization ability is insufficient. This paper argues that two methods, which utilize multiple data augmentation

DOI: 10.54254/2755-2721/2025.PO24691

techniques to expand the diversity of training data, simulate different real-world scenarios and integrate multiple datasets from different sources for training, can effectively address this challenge.

Another example is that module missing patterns in real-life scenarios are often very complex, and there may be multiple types of missing at the same time or multiple modalities with different degrees of missing at the same time. To address this challenge, the approach suggested in this paper is to widely apply an adaptive model structure so that it can automatically recognize different missing patterns. Optimize the complexity of their models. Alternatively, a combination of meta-learning and transfer learning is used to allow the model to be trained on several different datasets to learn the processing strategies under different missing patterns. Researchers can cycle through the process to analyze the problem and solve it, to get closer to the ideal multimodal sentiment analysis effect, and to gradually overcome the difficulties posed by modal absence.

5. Conclusion

This paper reviews seven approaches to deal with the problem of missing modalities in multimodal sentiment analysis, focusing on analyzing the frameworks and characteristics based on different approaches, which are classified into three categories: generative approaches, joint learning approaches, and approaches combining the two. The core is also explained, and the advantages and disadvantages of each modality are analyzed. In addition, the solutions for modality loss in three application scenarios of multimodal sentiment analysis in classroom teacher sentiment analysis, child sentiment analysis, and depression prognosis are presented, and finally, the current challenges are presented and suggested approaches are proposed based on the corresponding problems. The review shows that modal absence causes a big problem for sentiment analysis, but for different types of absence, such as unimodal absence, multimodal absence, and stochastic absence, each method has different advantages, and the solution to the problem of modal absence needs to be based on the specific problem to choose the method.

In future work, it is necessary to further explore in depth the theoretical basis and practical application effects of these innovative methods, and improve the performance and reliability of multimodal sentiment analysis models in module missing scenarios by continuously optimizing the model architecture, algorithm design, and experimental validation.

References

[1] PORIA S, CAMBRIA E, BAJPAI R, et al. A review of affective computing: from unimodal analysis to multimodal fusion [J]. Information Fusion, 2017, 37: 98-125.
[2] SOLEYMANI M, GARCIA D, JOU B, et al. A survey of multimodal sentiment analysis [J]. Image and Vision Computing, 2017, 65: 3-14.
[3] GUO Xu, Mairidan Wushouer, Gulanbaier Tuerhong. Survey of Sentiment Analysis Algorithms Based on Multimodal Fusion 2024 60(2).
[4] Zeng, J., Liu, T., & Zhou, J. (2022). Tag-assisted Multimodal Sentiment Analysis under Uncertain Missing Modalities. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1545–1554. https: //doi.org/10.1145/3477495.3532064
[5] Li, M., Yang, D., & Zhang, L. (2023). Towards robust multimodal sentiment analysis under uncertain signal missing. IEEE Signal Processing Letters, 30, 1497–1501. https: //doi.org/10.1109/lsp.2023.3324552
[6] Lin, R., & Hu, H. (2023). MissModal: Increasing Robustness to missing Modality in Multimodal sentiment analysis. Transactions of the Association for Computational Linguistics, 11, 1686–1702. https: //doi.org/10.1162/tacl_a_00628
[7] Li, M , Yang , D, Lei, Y , et al. A Unifed Self-Distillation Framework for Multimodal Sentiment Analysis with Uncertain Missing Modalities.

DOI: 10.54254/2755-2721/2025.PO24691

[8] Sun, Y., Liu, Z., Sheng, Q. Z., Chu, D., Yu, J., & Sun, H. (2024). Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities. Information Fusion, 110, 102454. https: //doi.org/10.1016/j.inffus.2024.102454

[9] Liu, Z., Zhou, B., Chu, D., Sun, Y., & Meng, L. (2023). Modality translation-based multimodal sentiment analysis under uncertain missing modalities. Information Fusion, 101, 101973. https: //doi.org/10.1016/j.inffus.2023.101973

[10] Huan, R, Zhong, G, Chen, P. UniMF: A Unified Multimodal Framework for Multimodal Sentiment Analysis in Missing Modalities and Unaligned Multimodal Sequences

[11] Wu Zheng. (2024). Research on the Analysis and Application of Teachers 'Emotions in Classroom Based on Multi-modal Fusion (Master's Thesis, Yunnan Normal University). Master's degree. https: //link.cnki.net/doi/10.27459/d.cnki.gynfc.2024.001087doi: 10.27459/d.cnki.gynfc.2024.001087.

[12] CENTERS FOR DISEASE CONTROL AND PREVENTION. Key findings from the ADDM network: A snapshot of autism spectrum disorder [J]. Community Report on Autism, 2016: 6.

[13] Wu Wenhao. (2024). A Multi-modal Information-based Algorithm for Children's Emotional Analysis and Its Application (Master's Thesis, Harbin Institute of Technology). Master's degree. https: //kns.cnki.net/kcms2/article/abstract?v=Jy1bRTva-nXnsV15tHFMYcnNXnpKS5GYbpaF6jr-40IeihCZGmnr9mYpumNmdVD7D8pUGfuEVSphsLtjSKkyCwU5EyCtFR9sOYapAYklTlo3aQImErchY2RnkfmNuw1SjZmBondPWCawHxoJ2Ikpzwz71IjN6adiEHw9qSzehbmw-0w7NZQAvry4TOWBb_l1& uniplatform=NZKPT& language=CHS

[14] Sharma, A., Saxena, A., Kumar, A., & Singh, D. (2024). Depression Detection Using Multimodal Analysis with Chatbot Support. 2024 2nd International Conference on Disruptive Technologies (ICDT), 328–334. https: //doi.org/10.1109/icdt61202.2024.10489080