Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/120/2025.18747

Review of Multimodal Sentiment Analysis Techniques

Zixuan Tang[1,a,*]

[1]*School of Artificial Intelligence, University of Science and Technology Beijing, Beijing, China*
a. Tangzixuan0321@hhu.edu.cn
**corresponding author*

Abstract: By integrating multimodal data such as text, image, audio and video, the multimodal emotion analysis technology has significantly improved the accuracy and robustness of emotion recognition, and has become a research hotspot in the field of artificial intelligence. This paper summarizes the research background, main research content and application of multimodal emotion analysis, and discusses the future development trend. This paper emphasizes the importance of emotion analysis in understanding user intention, improving user experience and optimizing decision support system, and points out the limitations of traditional unimodal emotion analysis, and leads to the necessity of multimodal emotion analysis. It systematically reviews the key technologies and methods of multimodal emotion analysis, including data fusion strategies, deep learning models, attention mechanisms, and cross-modal association learning, and analyzes the challenges in current research, such as data set imbalance, inter-modal inconsistency, and real-time requirements. This paper provides a comprehensive research idea for subsequent researchers, including constructing high-quality datasets, optimizing algorithms, expanding application scenarios, strengthening interdisciplinary cooperation, and emphasizing ethics and privacy protection, to promote the further development of multimodal sentiment analysis technology in the field of artificial intelligence. With the improvement of computing power and the innovation of algorithms, multimodal emotion analysis technology will be even more important in the field of artificial intelligence in the future.

Keywords: emotion analysis, artificial intelligence, multi-modal data fusion, deep learning, cross-modal correlation.

1. Introduction

With the rapid development of information technology, especially the rise of areas such as social media, online forums and customer service, people express their emotions in more and more diverse ways, no longer limited to a single text form. Multi-modal data, including text, images, audio and video, has become an integral part of people's daily communication. Traditional unimodal emotion analysis methods, which only rely on one data type (such as text only or speech only) for emotion analysis, have obvious limitations. This method does not fully capture the complex human emotional expression, especially in the face of multimodal information, a single modal information is often insufficient to provide accurate emotion analysis results. For example, a smiling expression may be accompanied by positive emotions, but may ignore this non-verbal cue if analysed from the text content alone. Therefore, the technology of multimodal emotion analysis has emerged at the historic

© 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: 10.54254/2755-2721/120/2025.18747

moment, which makes up for the deficiency of single mode analysis by combining the information of multiple modes, and effectively improves the accuracy and robustness of emotion recognition. The recent advances in deep learning and traditional feature extraction algorithms provide a technical basis for multimodal sentiment analysis. Researchers are trying to use facial expression messages, text messages and voice messages to analyze their emotions. These studies did not only drive the research. The development of emotion analysis technology has also provided the possibility to build more intelligent and natural interactive systems. The current research situation of multimodal emotion analysis technology is mainly reflected in the following aspects: on the one hand, some open source platform creation and emergence, for example, M-SENA platform mentioned in the literature [1], as an integrated open source platform, by providing flexible toolkit and reliable benchmark, promote the multimodal emotion analysis technology research and application, its modular video emotion analysis framework covers from data management to feature extraction, model training and results analysis of the whole process, provides researchers with a comprehensive technical support. On the other hand, with the development of deep learning and large model transfer learning technology, the researchers are committed to explore the deeper characteristics of fusion technology and application scenarios, especially in the field of emotional disorder detection and intervention, modal emotion analysis shows the huge potential of depression, anxiety, marks the technology in the realization of naturalization, anthropomorphic and personalized human-computer interaction has made remarkable progress. For example, literature [2] proposes a multimodal emotion analysis method based on intensive collaborative attention mechanism, which improves the accuracy of emotion analysis by integrating text, video and audio features. Research uses pre-trainingThe trained BERT model, Openface 2.0 and COVAREP extract multimodal features, and use BiLSTM and intensive collaborative attention mechanism for feature fusion. The experimental results show that this method can effectively improve the performance of emotion classification on CMU-MOSEI and CH-SIMS data sets. In conclusion, the research background of this review focuses on the development of multimodal emotion analysis technology, aiming to explore how to improve the accuracy of emotion identification by integrating and analyzing multimodal data, and the potential application of this technique in different fields.

2. Overview of related techniques

2.1. Emotional feature extraction technology

Emotional feature extraction is the basis of multimodal emotion analysis. In the field of multimodal emotion analysis, emotion feature extraction technique is key to understanding and identifying human affective states. This process involves extracting emotionally related information from multiple modalities, including but not limited to the three most common modalities: facial expressions, textual content, and phonological features. Facial expression recognition technology (FER) identifies emotions by analyzing geometric features of the face, such as the position of the eyes, nose and mouth, as well as appearance features based on local binary patterns (LBP) and Gabor wavelets. With the development of deep learning, convolutional neural networks (CNN) and deep neural network (DNN) began to be widely used in FER, which can automatically learn complex feature representations from large amounts of data. In addition, deep learning methods also combine traditional features, such as LBP, to improve the recognition performance through feature fusion. Text sentiment analysis focuses on the extraction of sentiment tendencies from text data, which usually involves the use of a sentiment dictionary to match sentiment words in the text, or using machine learning algorithms, such as logistic regression and support vector machine (SVM), combined with feature extraction techniques such as TF-IDF to identify emotion polarity. Deep learning methods, especially recurrent neural networks (RNN) and long-and short-term memory networks (LSTM), are due to their role in processing

DOI: 10.54254/2755-2721/120/2025.18747

sequential dataAt the same time, the introduction of attentional mechanism enables the model to pay more attention to emotion-related words. Speech emotive analysis focuses on extracting affective features from speech signals, such as tone, rhythm, and volume. Traditional methods, including SVM and hidden Markov model (HMM), identify emotions by extracting acoustic features (e.g., Mel frequency inversion coefficient MFCC). However, deep learning-based CNN and deep sparse neural network (DSNN) can extract more abundant emotion features from large-scale network speech data, and improve the accuracy of emotion analysis. Combining these technologies, multi-modal emotion analysis can capture the complexity of emotion expression from different angles, and integrate information from different modes through strategies such as feature-level fusion, decision-level fusion or hybrid fusion to achieve more accurate emotion recognition. This method not only improves the performance of sentiment analysis, but also provides a technical basis for realizing more natural and intelligent human-computer interaction. With the deepening of research and technological progress, the importance of multimodal emotion analysis in the field of emotion computing is becoming increasingly prominent

2.2. Multimodal fusion technology

Mode fusion technology is a key link in multimodal emotion analysis, which involves integrating feature information from different modalities in order to more comprehensively understand and identify emotional states. Modal fusion can be achieved through different levels and ways, mainly including feature-level fusion, decision-level fusion and hybrid fusion. Feature-level fusion (early fusion) is the integration of features of different modalities immediately after feature extraction. This method is simple and straightforward, usually forming a comprehensive feature representation by joining or stitching together different modes of feature vectors. The advantage of feature-level fusion is to provide a unified feature space for subsequent model training, but the challenge is how to effectively align and weight the features from different modalities. Decision-level fusion (late fusion) is to complete the emotional analysis of each mode independently, and then the decision results of each mode are integrated to produce the final emotional judgment. The advantage of this method is that each mode can be handled independently and is missing, but may not fully utilize the complementary information between modes. Hybrid fusion combines the advantages of feature-level fusion and decision-level fusion, which integrates information in both the feature level and the decision-making level. Hybrid fusion is able to simultaneously utilize complementary information between modalities and detailed information within each modality interest, but it is more complex to implement and requires careful design of fusion strategies. With the development of deep learning technology, the modal fusion technology is also constantly improving. For example, the attention mechanism is introduced into modal fusion, allowing the model to automatically learn the importance of different modal features and assign different weights accordingly. This approach is particularly suitable for situations where there are significant differences between modalities because it can highlight the most important emotional cues. Tensor fusion network (Tensor Fusion Network) is another advanced mode fusion technology, which captures the interaction between modes through the tensor operation. The tensor fusion networks are able to explicitly model the interactions between unimodal, bimodal, and trimodal data and are trained in an end-to-end manner. In addition, deep high-order sequence fusion networks (Deep Higher Order Sequence Fusion, Deep-HOSeq) achieve deeper mode fusion by extracting the fusion of information between modes and information within modes, as well as the information of time granularity of multi-modal interactions. Overall, the development goal of modal fusion technology is to achieve the effective integration of information from different modalities to improve the accuracy and robustness of sentiment analysis. With further research, future modal fusion techniques may focus more on dynamic interactions between modes and the utilization

DOI: 10.54254/2755-2721/120/2025.18747

of contextual information, and how to adapt to the emotion analysis needs in different application scenarios

2.3. Multimodal emotion recognition based on large models

Large language models (LLMs) and Transformer architectures perform well in multi-modal emotion recognition. Through self-supervised learning or contrast learning, these models are able to learn richer and more expressive multimodal feature representations. Multi-modal emotion recognition technology based on large models mainly involves deep learning, natural language processing and pattern recognition. Its core lies in how to effectively extract features from different modalities, fuse features, and classify emotions. This technology uses large neural network models, such as Transformer and BERT, to capture complex data features through massive parameters, and realize the understanding and analysis of various data types, such as text, audio and image. In technical principle, large models usually undergo two stages: unsupervised pre-training and supervised fine-tuning. In the pre-training stage, the model learns the common features of the language, such as vocabulary, grammar and semantics, on a large number of unannotated text data. In the fine-tuning stage, they are trained on the annotated emotion analysis data set to accommodate the specific emotion analysis task. The core algorithm includes a Transformer model architecture that captures long-distance dependencies and complex semantic structure in text through self-attention mechanisms, which is crucial for understanding emotional tendencies. In practice, large model emotion analysis techniques can be applied to social interaction media monitoring, customer feedback analysis and advertising recommendation and other scenarios. For example, in social media monitoring, companies can use the technology to monitor user reviews and feedback in real time and analyze users emotional tendencies toward the product. In customer feedback analysis, enterprises can automatically analyze the emotional information in customer feedback and identify the positive and negative reviews of customers to improve their products and services. Large Multimodal models (LMMs)further expand the capabilities of large models to process and integrate multiple data types, including text, images, audio and video, to more fully understand and generate a variety of content. These models extend the ability of traditional large-scale language models (LLMs) through additional modes to process and generate outputs across a variety of data types, such as interpreting visual input, generating textual descriptions of images, and even processing audio data to bridge the gap between different forms of information. In conclusion, multimodal large model-based emotion recognition is a rapidly evolving field that combines multiple techniques including deep learning, attention mechanisms and cross-modal analysis to achieve more accurate emotion recognition. With the deepening of research, this field is expected to achieve more breakthroughs in the future, and play an important role in multiple industries.

3. Main experimental results

3.1. Introduction of related data sets

Multimodal emotion analysis is an important research direction in the field of artificial intelligence, which involves the analysis of emotional tendencies in many forms of data, such as text, image, audio and video. The following are some common multimodal sentiment analysis data sets and related introductions (Table 1):

DOI: 10.54254/2755-2721/120/2025.18747

Table 1: Introduction of the relevant datasets

Name	Brief Introduction
CMU-MOSEI	This is one of the largest trimodal datasets, containing two labels: emotion and emotion. There are seven categories of emotion from negative to positive, including six categories: anger, happiness, sadness, surprise, fear and disgust. The values of the labels are between [-3 and 3]. The dataset provides raw data, but often requires further processing to accommodate different research needs.
CMU-MOSI	This is a multimodal emotion dataset, including text, visual, and acoustic modalities. It comes from 93 movie review videos on Youtube. The videos were edited into 2,199 clips. Each segment is annotated with emotional intensity in the range of [-3,3]. The data set is divided into three parts, the training set (1,284 segments), the validation set (229 segments), and the test set (686 segments).
MELD	Based on video clips from the TV series Friends, containing text, audio and video messages. This dataset is used for emotion recognition, especially for sentiment analysis in conversational settings.
IEMOCAP	This is a dataset of video conversations containing multiple affective states. It provides multimodal data for emotion analysis through an interactive emotion dynamic capture database.
MUStARD	This is a multimodal satire recognition dataset containing images and text bimodal satire videos. The corresponding to each label contains multiple images, and the corresponding text is a set of dialogues.
CH-SIMS	This is a Chinese multimodal sentiment analysis data set containing 2,281 video clips from different movies, TV series and variety shows showing natural expressions, various head postures, occlusion and lighting conditions.

3.2. Results of the representative experiments

Table 2: Correlation model performance

Model	Data Set	Assignment	Metric	Bear Fruit
TFN	CMU-MOSI	sentiment analysis	Acc-2	75.2%
TFN	CMU-MOSI	sentiment analysis	F1	74.0%
TFN	CMU-MOSI	sentiment analysis	MAE	0.335
LMF	CMU-MOSEI	sentiment analysis	Acc-7	39.5%
LMF	CMU-MOSEI	sentiment analysis	Acc-2	76.8%
LMF	CMU-MOSEI	sentiment analysis	F1	76.5%
MFM	IEMOCAP	Emotional recognition	Acc-2	90.2%
MFM	IEMOCAP	Emotional recognition	F1	85.8%
MFM	CMU-MOSI	sentiment analysis	Acc-2	78.1%
MFM	CMU-MOSI	sentiment analysis	F1	78.1%
MFM	CMU-MOSI	sentiment analysis	MAE	0.951

In the field of multimodal sentiment analysis, different models show different performance characteristics when processing various datasets. As shown in the table 2, tensor fusion network (TFN) [3] demonstrated high binization accuracy and F1 scores on the CMU-MOSI dataset, which may be attributed to its ability to effectively capture complex interactions between text, audio and visual modalities. The low-rank multi-modal fusion (LMF) [4] model is outstanding in the seven-fold classification task of the CMU-MOSEI dataset, indicating that its low-rank approximation method can effectively handle high-dimensional features when processing large-scale datasets. The

DOI: 10.54254/2755-2721/120/2025.18747

multimodal factorization model (MFM) [5] showed excellent performance in both IEMOCAP and CMU-MOSI datasets, especially in the binary classification accuracy and F1 score, which reflects the strong generalization ability of MFM in the multimodal emotion analysis task. These results indicate that modal fusion technology is critical to improve the accuracy and robustness of sentiment analysis, and the properties of different datasets also have significant effects on model performance. In practice, choosing the appropriate model needs to consider the characteristics and task requirements of the data set, and while combining multiple performance indicators to comprehensively evaluate the effect of the model. Future research may focus on developing more advanced modal fusion technologies, such as utilizing improvements in attention mechanisms and deep learning architectures to further enhance multimodal sentiment. The sex of sensory analysis

4. The Application of Multimodal Emotion Analysis Technology

4.1. Human-computer interaction

Multimodal emotion analysis plays an important role in the field of human-computer interaction and can accurately identify individual emotional states by analyzing text, video and voice features. This is essential to improving the user experience and improving the quality of interactions. For example, Chen et al. in 2021 proposed a multimodal emotion analysis method to analyze human communication [6] by combining text, audio and video data.

4.2. Education field

In the education field, multimodal emotion analysis techniques are used to monitor and analyze the emotional state of learners, thus to support teachers and help them better understand and support students learning. Giannakos et al in 2019 explored how multimodal data can help understanding the learning experience [7].

4.3. Health monitoring

Multimodal emotion analysis has also been applied to health monitoring, especially in monitoring status classification and emotional reassurance in patients with depression. For example, Amir et al. in 2018 proposed a multiattention recursive network for human communication understanding, which has potential applications in the field of health surveillance [8].

4.4. Research on psychological assistance in criminal investigation

In the field of criminal investigation, multi-modal emotion analysis technology can assist psychological research, and identify the emotional state of suspects by analyzing their facial expressions, voice and text information, so as to provide support for case detection. Huddar et al. discussed the progress of computational methods and challenges in their 2019 review article [9].

4.5. Retail sales of goods and public opinion

Emotion analysis is also widely used in the field of commodity retail and public opinion, which can analyze the emotional tendency of consumers towards products or events, so as to provide data support for market analysis and public relations strategies. For example, a multimodal sentiment analysis dataset for analyzing sentiment intensity and subjectivity in online opinion videos was constructed in a 2016 study by Zadeh et al [10].

DOI: 10.54254/2755-2721/120/2025.18747

5. Research Challenges

5.1. Data acquisition and privacy protection

Collecting multimodal data requires consideration of privacy protection and data security issues. Privacy protection and data security are crucial during multimodal data collection. With the wide application of big data technology, the problems of data security and privacy protection are becoming increasingly prominent. Current laws, regulations and technical means, such as data desensitization, blockchain, encryption and other technologies, are used to protect data security and privacy. Challenges include the risk of data leakage, privacy violations, and increased difficulty in security protection. To address these challenges, future trends include intelligent protection and collaborative protection, as well as the improvement of privacy protection laws and regulations. Therefore, these laws, regulations and technical means must be taken into account when conducting data acquisition to ensure that the data is collected and processed without violating personal privacy.

5.2. Data annotation and understanding

The subjectivity of emotion leads to difficulties in data annotation and requires more detailed annotation methods. The subjectivity of emotion leads to difficulties in data annotation and requires more detailed annotation methods. Data annotation is an important link in machine learning, especially in multimodal emotion recognition, which involves various types of data such as image, text and audio. For example, image annotation includes bounding box, semantic segmentation, instance segmentation, etc.; text annotation includes entity recognition, emotion analysis, part-of-speech annotation, etc. These annotation techniques are essential to improve model understanding and accuracy. Due to the complexity and subjectivity of emotion, a more refined annotation framework and tools need to be developed to better understand and analyze emotion data, thus improving the performance of the model.

5.3. Model generalization capability

How to improve the generalization ability of models on different scenarios and datasets is an important problem. Improving the generalization ability of models on different scenarios and datasets is an important problem. The improvement of model generalization ability can be achieved through a variety of methods, including cross-validation, regularization technology, etc. Cross-validation is a technique to evaluate the generalization ability of a model, in which the performance of a model is evaluated by dividing the data set into multiple subsets and alternately using one of the subsets as the test set and the rest as the training set. Regularization techniques, such as L1 and L2 regularization, improve the generalization ability of the model by adding penalty terms to the loss function, reducing the model complexity and preventing overfitting. These methods can help the model to show good performance on unseen data.

5.4. Real-time performance and efficiency

In practice, models are needed to handle emotion analysis tasks quickly and accurately. In practice, models are needed to handle emotion analysis tasks quickly and accurately. Real-time performance and efficiency are the key requirements in the practical application of emotion analysis models, especially in the fields of social media and public opinion monitoring. For real-time processing, flow processing frameworks (such as Apache Kafka, Apache Flink) can be used to process data flows, and models are deployed using the resilience and high availability of cloud computing platforms. In addition, developing the RESTful API interface is also a way to achieve rapid processing and

DOI: 10.54254/2755-2721/120/2025.18747

integration of models. These technical means can help the model to meet the requirements of real-time performance while ensuring the accuracy, and improve the efficiency and practicability of emotion analysis.

6. Development Trend and Outlook

6.1. Dataset construction

Establishing standardized multimodal emotion datasets is an important development direction in the field of multimodal emotion analysis. Future research will aim to build larger-scale, high-quality multimodal emotion databases to support the construction and exploration of more complex models of emotion analysis. These datasets will include different types of data such as images, audio, text, video, etc, as well as different types of tasks such as image classification, speech recognition, text classification, etc. High-quality data resources are crucial for researchers, and they can be fairly compared and evaluated on the same benchmarks, thus driving technological progress.

6.2. Algorithm optimization

To improve the accuracy and robustness of emotion recognition, it is crucial to develop more efficient multimodal emotion recognition algorithms. This includes improving the structure and training methods of deep learning models to improve model performance and generalization capabilities. Studying more efficient data fusion algorithms can make full use of information from different modalities, while using self-supervised and unsupervised learning methods can reduce the reliance on large amounts of annotated data. Moreover, exploring the associations between different modalities is also critical to improving the accuracy of emotion analysis.

6.3. Application and expansion

The application field of multimodal emotion analysis is expanding, including smart assistant, social media analysis, smart medical care, etc. These applications can meet the needs of emotion analysis in the different scenarios, such as intelligent customer service to improve customer service experience; analyzing user posts in social media analysis and predicting public sentiment trends; and evaluating students learning status in the education field to provide personalized teaching feedback

6.4. Strengthen interdisciplinary research

The development of multimodal emotion analysis technology requires the joint research of experts in computer science, psychology, cognitive science and other fields. Interdisciplinary collaboration can facilitate the development of multimodal emotion analysis techniques, by integrating knowledge and technology from different fields and enabling a more comprehensive understanding and analysis of human emotional expression.

6.5. Develop open-source tools and platforms

Developing open source multi-modal emotion analysis tools and platforms can lower the threshold for research and application, allowing more researchers and developers to participate in the research and development of this field. These tools and platforms provide detailed code and tutorials that support multiple datasets to facilitate further research and applications.

DOI: 10.54254/2755-2721/120/2025.18747

6.6. Strengthen ethics and privacy protection

Enhanced ethical and privacy protection is crucial when researching and applying multimodal sentiment analysis techniques. This includes ensuring that the technology is secure and reliable, and taking steps to protect user privacy, such as data anonymity, data encryption, etc. At the same time, it is necessary to pay attention to the changes in laws and regulations, and timely adjust the technical practices to ensure that they comply with the data protection Law, intellectual property law and other relevant laws and regulations.

7. Conclusion

This paper provides a comprehensive review of multimodal sentiment analysis techniques, from research background to applied practice to future trends, providing deep analysis and insights. The paper highlights the importance of multimodal emotion analysis in improving the accuracy and robustness of emotion recognition and highlights the key role of deep learning techniques in driving the development of this field. At the same time, the paper also puts forward future research directions, including building high-quality data sets, optimizing algorithms, expanding application scenarios, strengthening interdisciplinary cooperation, and focusing on ethics and privacy protection. Overall, multimodal sentiment analysis is moving toward a smarter and more natural interaction system, indicating a more central role in artificial intelligence. Finally, it is hoped that the comprehensive analysis of the cutting-edge technologies and challenges of multimodal emotion analysis can provide a possible research context for the academic circle and a possible technological development direction for the industry. For the in-depth discussion and future outlook of the article, it hopes to stimulate more innovative thinking, promote the further development of multi-modal emotion analysis technology in the field of artificial intelligence, and contribute to the construction of a more intelligent and natural interactive system.

References

[1] *AmirZadeh, M., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). M-SENA: An Integrated Platform for Multimodal Sentiment Analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 1103–1114)*

[2] *Zhou, S. X., Yu, K. (2024). Multimodal emotion analysis based on intensive collaborative attention. computer project. https://doi.org/10.19678/j.issn.1000-3428.0069721*

[3] *Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor Fusion Network for Multimodal Sentiment Analysis. arXiv.org. https://arxiv.org/abs/1707.07250*

[4] *Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L.-P. (2018). Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. arXiv.org. https://arxiv.org/abs/1806.00064*

[5] *Tsai, Y.-H. H., Liang, P. P., Zadeh, A., Morency, L.-P., & Salakhutdinov, R. (2019). Learning Factorized Multimodal Representations. In Proceedings of the International Conference on Learning Representations (ICLR). Retrieved fromhttps://arxiv.org/abs/1806.06176*

[6] *Chen, F. Y., Sun, Z. X., Ouyang, D. Q., Liu, X. Y. and Shao, J. (2021). Learning What and When to Drop: Adaptive Multimodal and Contextual Dynamics for Emotion Recognition in Conversation. In Proceedings of the 29th ACM International Conference on Multimedia (MM 21). Association for Computing Machinery, New York, NY, USA, 1064–1073. https://doi.org/10.1145/3474085.3475661*

[7] *Michail, N. Giannakos, K. S., Ilias, O. Pappas, V. K., Eduardo, V. (2019). Multimodal data as a means to understand the learning experience, International Journal of Information Management, Volume 48, Pages 108-119, ISSN 0268-4012, https://doi.org/10.1016/ j.ijinfomgt.2019.02.003.*

[8] *Feng, J., Li, Y., Zhang, C., Sun, F. N., Meng, F. C. Guo, A. G. and Jin, D. P. (2018). DeepMove: Predicting Human Mobility with Attentional Recurrent Networks. In WWW 2018: The 2018 Web Conference, April 23–27, Lyon, France. ACM, New York, NY, USA 10 Pages. https://doi.org/10.1145/3178876.3186058*

[9] *Huddar, M., Sannakki, S., Rajpurohit, V. (2019). A Survey of Computational Approaches and Challenges in Multimodal Sentiment Analysis. INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING. 7. 876-883. 10.26438/ijcse/v7i1.876883.*

DOI: 10.54254/2755-2721/120/2025.18747

[10] Zadeh, A., Zellers, R., Pincus, E., & Morency, L.-P. (2016). MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. arXiv.org. https://arxiv.org/abs/1606.06259