Deep Learning-Based Multimodal Sentiment Analysis

Zijian Wang
Mathematics BSc, University College London, London, U.K.

Keywords: Deep Learning, Sentiment Analysis, Multimodal, Natural Language Process.

Abstract: The area of natural language processing has a substantial amount of research that focuses on multimodal sentiment analysis. It aims at how people express their feelings through different types of speech and can be used in many areas, such as e-commerce, film and TV reviews, and more. With the advent of technologies like as machine learning, deep learning, and others, significant progress has been achieved in the area of multimodal sentiment analysis. First, this paper introduces multimodal emotion analysis, and then divides emotion analysis into narrative and interactivity according to the presence or absence of dialogue. The characteristics and distinctions of these two sentiment analysis approaches are then introduced, with respect to data, algorithm, and application, by analyzing pertinent recent domestic and international research. Lastly, this work addresses the future directions for research as well as the current drawbacks of multimodal sentiment analysis. With that said, this paper provides a reference for sentiment analysis researchers and outlines future research in this dynamic topic.

1 INTRODUCTION

Sentiment analysis, which is also called opinion or motion mining, is considered as a key area of NLP (natural language processing) that identifies and categorizes subjective information about products, services, organizations, events, and subjects (Meliville et al. 2009). This field uses methods from NLP, statistics, and machine learning in light of figuring out the meaning behind spoken or written language, which shows how people feel, what they think, and what they believe. In today's data-driven society, sentiment analysis is vital for boosting customer service and leading product development.

The rise of social media has significantly altered communication patterns, resulting in a need for multimodal sentiment analysis (Liu & Zhang 2012, S.L.C. & Sun 2017). This newer approach seeks to understand sentiments expressed not only through words but also through images, audio, and video. This comprehensive method aims to capture the complex and multifaceted nature of sentiment expression, which often includes a mix of verbal as well as non-verbal cues. The transition to multimodal sentiment analysis is a significant step in the area, which holds the promise of advancements in emotion identification technology that are both more sophisticated and accurate.

However, the move to multimodal sentiment analysis brings its own set of challenges. The main difficulty lies in effectively combining and interpreting the varied data types involved in multimodal communication. This includes the challenge of aligning and merging information from different sources, each with its unique characteristics, and dealing with the dynamic nature of interactive communications. These hurdles underscore the need for innovative solutions in areas like multimodal representation learning, alignment, and fusion to push the field forward.

Despite these challenges, the potential advantages of successful multimodal sentiment analysis are significant. It can lead to more precise and nuanced interpretations of emotions, enhancing customer insights, media experiences, and communication strategies across various platforms. By addressing the limitations of unimodal analysis and tapping into the rich potential of multimodal data, researchers aim to deepen our understanding of human sentiment. This progress is not just a leap forward in natural language processing and artificial intelligence but also marks a step towards creating more empathetic and human-centered technology (Zhang et al. 2018).

This research delves into the complexities of multimodal sentiment analysis, examining its challenges and the opportunities it presents. It adopts

Wang, Z.
Deep Learning-Based Multimodal Sentiment Analysis.
DOI: 10.5220/0012842200004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 356-363
ISBN: 978-989-758-690-3

a multidisciplinary approach, drawing from linguistics, psychology, computer science, and data science, to develop new computational models and algorithms. These models aim to better capture how humans express and perceive emotions across different contexts and cultures. Through this exploration, the study identifies crucial challenges in the field and suggests innovative solutions, shedding light on the evolution from text-only methods to multimodal approaches. It also evaluates the strengths and weaknesses of current technological methodologies, providing insights into potential applications. The dissertation concludes by summarizing the findings and looking ahead to future research directions in multimodal sentiment analysis. This includes considering the impact on natural language processing and proposing a roadmap for further research to enhance sentiment analysis technologies.

2 DATASET

2.1 Data Requirements and Datasets

Multimodal sentiment analysis stands at the intersection of various data types, each contributing a unique perspective to the understanding of sentiments and emotions. This analysis relies heavily on the amalgamation of text, images, audio, and video data to offer a multidimensional view of sentiment expression. Textual data, ranging from concise tweets to detailed product reviews, serves as a direct articulation of thoughts and opinions, providing clear indicators of sentiment polarity. Images, whether they are standalone pictures or part of video content, convey emotions through visual elements such as colors, expressions, and symbols, offering insights into the sentiment without the need for words. Audio data adds another layer, with the tone, pace, and pitch of voice carrying subtle cues about the speaker's emotional state. Video data combines these elements, presenting a rich narrative of sentiment through dynamic interactions, facial expressions, and verbal communication, encapsulated in a temporal sequence.

To facilitate research and development in this area, several public datasets have become invaluable resources, each characterized by its multimodal content and annotations:

- CMU-MOSI: This dataset serves as a benchmark and contains video clips. Each clip is annotated with sentiment scores across multiple modalities, making it a comprehensive resource for analyzing opinion dynamics (Zadeh et al. 2016).
- IEMOCAP: The abbreviation of interactive emotional dyadic motion capture. Video and audio recordings of dialogues that have been acted out by professional actors and annotated with a variety of emotional states are included in this dataset. It is particularly useful for studies focusing on emotional expressions in conversational contexts.
- SEMAINE: A collection of audio-visual recordings from interactions with a Sensitive Artificial Listener (SAL) system, designed to elicit emotional responses. Annotations include dimensional and categorical emotion labels, facilitating research into affective computing.
- YouTube-8M: A large-scale dataset that offers a wide array of YouTube videos tagged with labels, including topics and sentiments. While it primarily serves as a resource for video understanding tasks, its extensive collection allows for sentiment analysis across diverse video content.

These datasets originate from varied sources, including social media platforms, dedicated research efforts, and public contributions, encompassing a wide range of subjects, contexts, and emotional expressions. They contain raw multimodal data and annotations that mark emotional states, giving an elementary truth for training and assessing sentiment analysis models.

The variety and depth of these datasets show how multimodal sentiment analysis is changing over time. They also give researchers and practitioners a base for making models that are sophisticated and more accurate. By leveraging these resources, the field continues to advance, enhancing our ability to decipher the complex tapestry of human emotions as expressed through the myriad channels of communication in the digital age.

2.2 Data Processing Techniques

The efficacy of multimodal sentiment analysis hinges on sophisticated data processing techniques tailored to prepare and integrate diverse data types—text, images, audio, and video—into a coherent framework that models can analyze. Each modality undergoes specific preprocessing steps to transform raw data into structured forms suitable for sentiment analysis (Soleymani et al. 2017, Poria et al. 2017).

2.2.1 Text Data Processing

Textual information, with its rich semantic and syntactic diversity, is preprocessed through tokenization and vectorization. Advanced language models like BERT or Word2Vec are employed to encode text into numerical vectors. These models capture the nuances of language, including context, sentiment, and the relationships between words, converting unstructured text into a structured form that sentiment analysis algorithms can interpret. This process involves natural language processing techniques, which include lemmatization, stemming, and the removal of stop words to refine the text data further before vectorization.

2.2.2 Image Data

Images are processed using techniques that allow for the extraction of emotional cues embedded in visual content. Convolutional Neural Networks (CNNs) then analyze these preprocessed images, extracting features that reflect visual sentiments, such as colors, textures, and facial expressions. This process enables the model to understand and interpret the sentiment conveyed through visual information directly.

2.2.3 Audio Data

Audio data requires conversion into a format that highlights features relevant to sentiment analysis, such as tone, pitch, and rhythm. Preprocessing steps include sampling, noise reduction, and the extraction of features like Mel-Frequency Cepstral Coefficients (MFCCs) or spectrograms. These features encapsulate the emotional nuances present in audio data, preparing it for analysis by models capable of processing sequential and time-series data, such as Recurrent Neural Networks (RNNs) or LSTMs.

2.2.4 Video Data

Video, as a combination of audio and visual data along with potential textual components (like subtitles), undergoes a composite preprocessing routine. Frames extracted from videos are processed in a manner similar to images, while the audio track is treated as standalone audio data. Additionally, textual information embedded in videos is extracted and processed using text data techniques. The challenge lies in effectively synchronizing these modalities to maintain the temporal coherence of sentiment expressions throughout the video.

By employing these data processing techniques, multimodal sentiment analysis models are equipped to handle the complexities and subtleties of human emotions as conveyed through multiple modes of communication. This comprehensive approach to data preparation and integration is crucial for the development of accurate and effective tools to analyze sentiment.

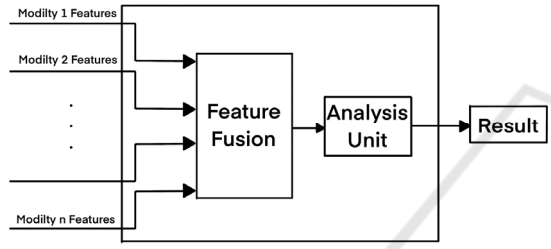
3 NARRATIVE MULTIMODAL SENTIMENT ANALYSIS

The goal of narrative multimodal sentiment analysis is to classify subjective attitudes into categories such as Positive, Negative, and Neutral. This approach stands apart from unimodal sentiment analysis by not only necessitating feature learning but also requiring the process of information fusion (Qian et al. 2019, Verma et al. 2019). This process integrates data from different modalities which include text, images, or videos, in what is referred to as multimodal interaction or multimodal fusion. The prevalent methods for multimodal fusion are categorized into three main types: feature-level fusion, decision-level fusion, and hybrid fusion, each offering unique advantages and facing distinct challenges (Atrey et al. 2010, Zhang et al. 2020), as shown in Figure 1.

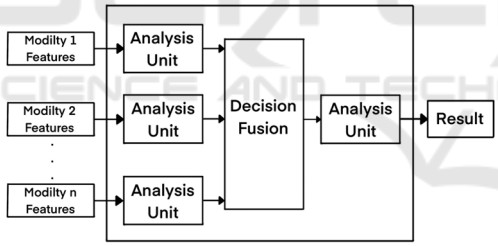
- Feature-level fusion primarily combines feature vectors from each modality, such as textual and visual feature vectors, into a singular multimodal feature vector for subsequent decision-making analysis. Its strength lies in capturing the intermodal feature correlations, facilitating a richer sentiment analysis. This method requires the early fusion of modal features, simplifying the classification process to a single classifier. However, the challenge arises from the need to map features from differing semantic spaces and dimensions into a shared space, considering their variance in time and semantic dimensions.
- Decision-level fusion, on the other hand, operates by independently extracting and classifying features from each modality to achieve local decisions, which are then merged to form the final decision vector. This method offers simplicity and flexibility, allowing each modality to utilize the most suitable feature extractors and classifiers for optimal local decisions. Despite its advantages, the necessity to learn classifiers for all modalities elevates the time cost of the analysis process.
- Hybrid fusion represents a mix of fusion of feature-level and decision-level, aiming to harness their benefits while mitigating the drawbacks of each. This approach endeavors to

provide a comprehensive and efficient strategy for multimodal sentiment analysis, optimizing the fusion process for improved sentiment classification.

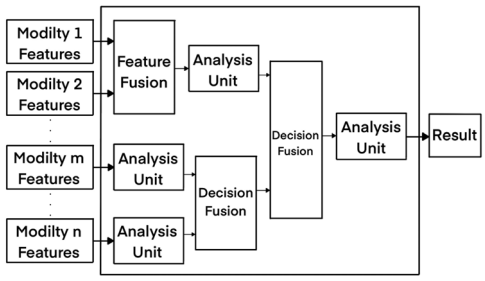
Moreover, multimodal sentiment analysis extends beyond textual analysis to include images, audio, and video, capturing the dynamic nature of speech or movement across different time frames. By classifying text and images as static documents and audio and video as dynamic documents, this paper explores both static and dynamic multimodal sentiment analysis. It emphasizes the development and current state of technology-driven static and dynamic analyses, respectively, highlighting the significance of combining textual, visual, and auditory data to discern subjective tendencies.



(a) Feature-level



(b) Decision-level



(c) Hybrid-level

Figure 1: Multimodal Fusion Strategies (Atrey et al. 2010).

3.1 Static Multimodal Emotion Analysis

Nowadays people are becoming more and more interested in digital photography as social media sites become more famous. Usage has soared, and more and more images are widely distributed on the web. These images accompany the text together to express the author's emotional information and make the connection. Using images and texts to explore public opinions, preferences, and emotions becomes a channel. In view of the user's increasing demand for emotional expression, in the highest language, the meaning level (that is, the emotional level) is the analysis of the multimodal content of the text. The more urgent it becomes, the more graphically (i.e., static) sentiment analysis attracts to the eyes of more researchers. There are many research papers in the area of graphic emotion analysis right now. Among various techniques, machine learning and deep learning have made much progress (Yuan et al. 2013).

3.2 Machine Learning and Deep Learning Approaches

With the introduction of machine learning, static multimodal sentiment analysis has been greatly advanced. This has been accomplished through the utilization of statistical algorithms such as SVM, RF, and NB (Cao et al. 2014, Wagner et al. 2011). By approaching graphic emotion analysis as a supervised classification task, these methods have made it simpler to investigate the complex link that exists between emotions expressed in written language and visual representations. They do this by using features like ANP to improve emotion inference and adding textual titles for a more complete analysis. However, despite their high recognition rates, these techniques heavily depend on the painstaking process of feature engineering, making them labor-intensive and time-consuming.

Deep learning has revolutionized static multimodal sentiment analysis, offering end-to-end solutions that circumvent the need for manual feature engineering (Devlin et al. 2019, You et al. 2016, Poria et al. 2017). Models like CNNs and LSTM networks have demonstrated superior performance across various tasks, including image processing and natural language processing. By employing techniques such as attention mechanisms and tensor fusion networks, deep learning approaches have achieved significant improvements in classifying emotions from static multimodal data. Nevertheless, these methods require extensive data and computational resources, posing

challenges in terms of training time and computational cost.

3.3 Dynamic Multimodal Emotion Analysis

Massive amounts of video data are uploaded to the Internet every day. Therefore, the study is not content to use only images and text information, but to start paying attention to other media resources related to it, such as accompanying voice and audio. It is committed to integrating a variety of media information to make the video emotion more complete, a comprehensive, accurate understanding, while illuminating psychology, philosophy, linguistics, etc., other disciplines, rich research areas (S.L.C. & Sun 2017). Compared to graphic emotion analysis, see frequency-emotion analysis often involves motion changes in different time frames, showing dynamic. Thus, dynamic emotion analysis becomes static emotion analyze the inevitable trend of development.

3.4 Machine Learning and Deep Learning Approaches

Before deep learning gained prominence, machine learning algorithms were the mainstay for analyzing video emotions, focusing on integrating textual, visual, and auditory cues for sentiment analysis (Tao 2009, Sebe et al. 2006). Techniques such as the HMM and SVM have been employed to analyze emotions through multimodal feature fusion, demonstrating the effectiveness of combining different modalities for enhanced emotion recognition. However, similar to static analysis, these methods require significant efforts in feature engineering.

Deep learning methods, like CNNs, LSTMs, and GANs, are being used more and more in dynamic multimodal sentiment analysis because they can perform end-to-end learning. These models excel in analyzing the complex interplay of textual, visual, and auditory information in videos, employing strategies like bidirectional LSTMs and feature fusion based on Gaussian kernels for emotion recognition (Zhang et al. 2009). Deep learning-based models have set new benchmarks in accuracy for video sentiment analysis, albeit with the challenges of requiring large datasets and extensive computational resources.

4 INTERACTIVE MULTIMODAL SENTIMENT ANALYSIS

Interactive multimodal sentiment analysis, which is also interpreted as multimodal conversational sentiment analysis, is used to figure out how people's feelings change during chat conversions. This field extends beyond the scope of declarative multimodal sentiment analysis by focusing on the fluidity and evolution of emotional states among participants in a dialogue. It presents unique challenges distinct from narrative multimodal sentiment analysis because of several factors:

- Interactions with other people have the potential to affect the emotional state of each person, which can result in shifts in sentiment as the conversation progresses or progresses.
- Conversations encapsulate hidden layers of information, such as cultural backgrounds, professional settings, and the nature of social relationships, which can significantly impact the emotional undertones of the dialogue.
- The thought process of speakers during a chat may not follow a linear trajectory, resulting in non-coherent discourse and sudden shifts in the topic of discussion.

These complexities necessitate a nuanced approach to sentiment analysis that not only processes textual and multimodal data but also deciphers the intricate web of interactions among speakers. The purpose of interactive multimodal sentiment analysis is to determine whether or not a chat session contains subjective information and, if it does, to identify the emotional state of each message while also tracking how the emotions of each participant changed throughout the conversation.

Addressing these challenges requires advanced models capable of capturing the nuanced interplay of emotions in conversational contexts. This task is not only pivotal for refining sentiment analysis techniques but also contributes to broader advancements in artificial intelligence by enhancing the understanding of human-machine interactions.

4.1 Multimodal Conversational Emotion Dataset

Over the years, researchers have built various types of multimodal emotion datasets, providing experimental data for multimodal emotion analysis models. Table 1 below shows the frequently used datasets.

Table 1: Multi-Modal Sentiment Datasets.

Type	Dataset	Modality	Link
Narrative Multi-modal	T4SA	Image, Text	http://www.t4sa.it/
	CHEAVD2.0	Video, Audio	http://www.chineseldc.org/emotion.html
	Multi-ZOL	Image, Text	https://github.com/xunan0812/MIMN
	SEED	Brainwave	http://bcmi.sjtu.edu.cn/~seed/
	Yelp	Image, Text	https://www.yelp.com/dataset/challenge
	HUMAINE	Video, Audio	http://emotion-research.net/download/pilot-db
	Belfast	Video, Audio	http://belfast-naturalistic-db.sspnet.eu
	YouTube	Text, Video, Audio	E-mail request
Interactive Multi-modal	CMU-MOSI	Image, Text	https://www.amir-zadeh.com/datasets
	EmotionLines	Text	https://academiasinicanlplab.github.io/#download
	DailyDialog	Text	http://yanran.li/dailydialog
	ScenarioSA	Text	https://github.com/anonymityanonymity/
	SEMAINE	Video, Audio	http://semaine-db.eu
	IEMOCAP	Video, Audio	http://sail.usc.edu/iemocap/
	MELD	Text, Video, Audio	https://affective-meld.github.io/
	EmoContext	Text	http://humanizing-ai.com/emocontext.html

These datasets provide resources for studying human interaction related to various emotions, with each dataset designed to capture different aspects of conversational sentiment.

4.2 Multimodal Conversational Emotion Analysis Model

Interaction is a way to change the actions or thoughts of other entities that are indirect and unseen. Given the complexity and concealment of the interaction mechanism, understanding and computing interlocutor interactions has been a challenging field in social sciences. Researchers have made numerous attempts to address this issue, with early models like HMM and influence models attempting to formalize and calculate interpersonal interactions.

In the area of sentiment analysis, it has been difficult to model how discourse interacts with each other in conversation. Early versions of conversational sentiment analysis looked at each utterance on its own, without taking into account how they related to each other. However as deep learning has grown, researchers have started to pay more attention to how words interact and affect each other. They have created multimodal conversational mood analysis models that can record this data. These models are based on deep learning technology and aim to understand the complex and concealed interactions within conversations, although research in this area is still relatively sparse.

Models like the contextual LSTMs designed by Poria et al. (2017), and DialogueRNN by Majumder et al. (2019) represent efforts to trace each conversationalist's emotional state and model the evolution of emotion throughout the conversation. These and other models aim to advance the area of interactive multimodal sentiment analysis by capturing the nuanced interactions that occur within conversations.

In conclusion, interactive multimodal sentiment analysis is gaining popularity as researchers produce cutting-edge results and advance the discipline.

5 CHALLENGE FACED FOR INTERACTIVE CONSTRUCTION OF MULTI-MODAL SENTIMENT ANALYSIS

Multimodal sentiment analysis, with its convergence of disciplines like linguistics, computer science, and cognitive science, faces significant challenges in deciphering the complex interplay of sentiments across different communication modalities. As technology becomes ever more entwined with our daily communication, the ability to accurately interpret sentiments from varied data sources—text, images, and videos—becomes increasingly critical. This analysis involves not just the examination of

sentiment within individual modalities but also understanding how these different forms of expression interact and combine to convey comprehensive emotional narratives.

5.1 Lexical Interaction Problem in Modalities

The first challenge pertains to the lexical interactions within individual modalities, particularly text. Words in a sentence are not isolated entities but are closely interconnected, influencing each other to convey comprehensive semantic meanings. One-hot encoding, bag-of-words, and N-gram models have been successful but struggle with polysemy, when a word's meaning changes with context. Dynamic word representation techniques like ELMo and BERT, which consider the context on both sides of a word, have shown improvement in various natural language processing tasks (Peters et al. 2018). However, these methods primarily capture proximate contextual information, leaving the modeling of long-distance lexical interactions as an area ripe for exploration.

5.2 Inter-Modal Multimodal Interaction Problems

The second challenge focuses on the interactions between different modalities, aiming to integrate information from various sources like text, images, and audio to model their associations and interactions. This integration is crucial for generating richer and more accurate multimodal outputs. The main issues here involve aligning and fusing features from different modalities, each extracted from distinct semantic spaces and temporal instances. Researchers have experimented with several approaches, such as feature concatenation, deep network-based shared latent learning, tensor fusion, and attention mechanism-based fusion. Despite these efforts, achieving effective alignment and fusion within a shared space remains a significant challenge, necessitating further research to develop a unified theory of inter-modal interaction.

5.3 Discourse Flow Interaction Beyond Modalities

The third challenge emerges from the discourse flow interactions that extend beyond the modalities themselves, highlighting the importance of conversational context in sentiment analysis. The same phrase can have different meanings and emotional connotations depending on the dialogue context, making the analysis of utterance interactions crucial. Conversational sentiment analysis requires understanding the nuanced implications of statements within various dialogic backgrounds, particularly in chat scenarios where statements entail strong and repetitive interactions. Current research often treats utterances in isolation, lacking a systematic study of their interplay. Developing models that can comprehend and represent these discourse flow interactions remains a core challenge, pointing toward the need for a generalized framework for modeling discourse context interactions.

Each of these interaction challenges represents a distinct layer of complexity in multimodal sentiment analysis, from intra-modal lexical dependencies to inter-modal dynamics and beyond to the overarching conversational context. To solve these problems, new ways are required to properly record and combine the complex ways that people express their feelings across and within different modes of communication. This will help the fields of artificial intelligence and sentiment analysis make progress.

6 CONCLUSION

Multimodal sentiment analysis is well-known for its ability to figure out how people are feeling by looking at different types of communication. This paper has delved into the essentials of multimodal sentiment analysis, covering its research background, problem definitions, and current advancements. It has highlighted the challenges in this field and outlined potential future research directions.

As we look to the future, several areas are ripe for exploration. One such area is the integration of unconventional data types like touch feedback and physiological signals, which could offer new insights into emotional analysis. Even though there has been growth, there are still not enough complete datasets for interactive multimodal sentiment analysis. The datasets we have now, like MELD and IEMOCAP, are mostly based on scripted scenarios, which might not fully reflect how people behave in real life. This gap indicates a need for more authentic datasets that reflect genuine human communication.

Future research should also focus on developing lightweight models that prioritize data privacy and require less data to operate effectively. This approach addresses practical challenges like model deployment on limited-resource devices and ensures user privacy. Additionally, finding ways to enhance model performance with minimal data could solve the

problem of data scarcity and reduce reliance on extensive datasets.

Another important direction is the exploration of general interaction theories that can manage the complex interactions within discourse flows, especially in settings involving multiple speakers. Creating a systematic framework for modeling these relationships could improve sentiment analysis with multi-model.

All in all, multimodal sentiment analysis has the potential for major advances. Researchers can develop more complex, ethical, and practical technology-based solutions for understanding human emotions by addressing these problems and researching the suggested future possibilities.

REFERENCES

P. Melville, W. Gryc, R.D. Lawrence. "Sentiment analysis of blogs by combining lexical knowledge with text classification." *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1275–1284 (2009).

B. Liu, L. Zhang. "A survey of opinion mining and sentiment analysis." *Mining Text Data, Springer U.S.*, pp. 415–463 (2012).

S. L. C. & C.J. Sun. "A review of natural language processing techniques for opinion mining systems." *Inform. Fusion*, 36, 10-25 (2017).

Y. Z. Zhang, D. W. Song, P. Zhang, et al. "A Quantum-Inspired Multimodal Sentiment Analysis Framework." *Theoretical Computer Science*, 752, 21-40 (2018).

A. Zadeh, R. Zellers, E. Pincus, L.P. Morency. "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos." arXiv preprint arXiv:1606.06259 (2016).

M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.F. Chang, M. Pantic. "A survey of multimodal sentiment analysis." *Image Vis. Comput.*, 65, 3-14 (2017).

S. Poria, E. Cambria, R. Bajpai, A. Hussain. "A review of affective computing: from unimodal analysis to multimodal fusion." *Inform. Fusion*, 37, 98-125 (2017).

Y. F. Qian, Y. Zhang, X. Ma, et al. "EARS: Emotion-Aware Recommender System Based on Hybrid Information Fusion." *Information Fusion*, 46, 141-146 (2019).

S. Verma, C. Wang, L. M. Zhu, et al. "DeepCU: Integrating Both Common and Unique Latent Information for Multimodal Sentiment Analysis." *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, New York, USA: ACM, 3627-3634 (2019).

P. K. Atrey, M. A. Hossain, A. El Saddik, et al. "Multimodal Fusion for Multimedia Analysis: A Survey." *Multimedia Systems*, 16(6), 345-379 (2010).

Y. Zhang, L. Rong, D. Song, P. Zhang. "A Survey on Multimodal Sentiment Analysis." Pattern Recognition and Artificial Intelligence, 33(5), 426-438 (2020).

J. B. Yuan, S. McDonough, Q. Z. You, et al. "Sentribute: Image Sentiment Analysis from a Mid-level Perspective." *Proceedings of the 2nd International Workshop on Issues of Sentiment Discovery and Opinion Mining*, pp. 10-12 (2013).

Cao, D. L., Ji, R. R., Li, "Visual Sentiment Topic Model Based Microblog Image Sentiment Analysis." *Multimedia Tools and Applications*, 75(15), 8955-8968.

J. Wagner, E. Andre, F. Lingenfelser, et al. "Exploring Fusion Methods for Multimodal Emotion Recognition with Missing Data." *IEEE Transactions on Affective Computing*, 2(4), 206-218 (2011).

J. Devlin, M. W. Chang, K. Lee, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171-4186 (2019).

Q. Z. You, L. L. Cao, H. L. Jin, et al. "Robust Visual-Textual Sentiment Analysis: When Attention Meets Tree-Structured Recursive Neural Networks." *Proceedings of the 24th ACM International Conference on Multimedia*, pp. 1008-1017 (2016).

S. Poria, E. Cambria, D. Hazarika, et al. "Multi-level Multiple Attentions for Contextual Multimodal Sentiment Analysis." *Proceedings of the IEEE International Conference on Data Mining*, pp. 1033-1038 (2017).

S. L. C. & C.J. Sun. "A review of natural language processing techniques for opinion mining systems." *Inform. Fusion*, 36, 10-25 (2017).

J. H. Tao. "A Novel Prosody Adaptation Method for Mandarin Concatenation-Based Text-to-Speech System." *Acoustical Science and Technology*, 30(1), 33-41 (2009).

N. Sebe, I. Cohen, T. Gevers, et al. "Emotion Recognition Based on Joint Visual and Audio Cues." *Proceedings of the 18th International Conference on Pattern Recognition*, pp. 1136-1139 (2006).

X. Y. Zhang, C. S. Xu, J. Cheng, et al. "Effective Annotation and Search for Video Blogs with Integration of Context and Content Analysis." *IEEE Transactions on Multimedia*, 11(2), 272-285 (2009).

S. Poria, E. Cambria, D. Hazarika, et al. "Context-Dependent Sentiment Analysis in User-Generated Videos." *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Stroudsburg*, USA: ACL, 873-883 (2017).

N. Majumder, S. Poria, D. Hazarika, et al. "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations." *Proceedings of the AAAI Conference on Artificial Intelligence*, Palo Alto, USA: AAAI Press, 6818-6825 (2019).

M. E. Peters, M. Neumann, M. Iyyer, et al. "Deep Contextualized Word Representations." *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, Stroudsburg, USA: ACL, 2227-2237 (2018).