

A comprehensive survey on multimodal sentiment analysis: Techniques, models, and applications

Heming Zhang

Guangdong Technion-Israel Institute of Technology, Shantou, Guangdong, 515063, China

zhang09558@gtiit.edu.cn

Abstract. Multimodal sentiment analysis (MSA) is an evolving field that integrates information from multiple modalities such as text, audio, and visual data to analyze and interpret human emotions and sentiments. This review provides an extensive survey of the current state of multimodal sentiment analysis, highlighting fundamental concepts, popular datasets, techniques, models, challenges, applications, and future trends. By examining existing research and methodologies, this paper aims to present a cohesive understanding of MSA. Multimodal sentiment analysis (MSA) integrates data from text, audio, and visual sources, each contributing unique insights that enhance the overall understanding of sentiment. Textual data provides explicit content and context, audio data captures the emotional tone through speech characteristics, and visual data offers cues from facial expressions and body language. Despite these strengths, MSA faces limitations such as data integration challenges, computational complexity, and the scarcity of annotated multimodal datasets. Future directions include the development of advanced fusion techniques, real-time processing capabilities, and explainable AI models. These advancements will enable more accurate and robust sentiment analysis, improve user experiences, and enhance applications in human-computer interaction, healthcare, and social media analysis. By addressing these challenges and leveraging diverse data sources, MSA has the potential to revolutionize sentiment analysis and drive positive outcomes across various domains.

Keywords: Multimodal Sentiment Analysis, Natural Language Processing, Emotion Recognition, Data Fusion Techniques, Deep Learning Models

1. Introduction

Sentiment analysis, a crucial component of natural language processing (NLP), has become increasingly significant in understanding and interpreting human emotions conveyed through text. Traditionally, sentiment analysis focused predominantly on textual data, aiming to identify and categorize opinions expressed in written content. However, human communication is inherently multimodal, involving a blend of verbal and non-verbal cues, including speech intonations, facial expressions, gestures, and body language. This intrinsic multimodality necessitates a broader approach to sentiment analysis, leading to the development of multimodal sentiment analysis (MSA) [1].

Multimodal sentiment analysis (MSA) aims to leverage diverse sources of information to achieve a more nuanced and accurate understanding of sentiments and emotions. By integrating data from multiple modalities—text, audio, and visual—MSA provides a comprehensive view of human communication, capturing the subtle and complex ways in which emotions are expressed. This integration can significantly enhance the performance of sentiment analysis systems, offering more reliable and contextually aware interpretations.

The evolution of sentiment analysis can be traced back to the early 2000s when researchers began exploring ways to automatically classify opinions in text. Initial approaches predominantly relied on machine learning techniques applied to textual data, using features like n-grams, part-of-speech tags, and sentiment lexicons. These methods showed promise in analyzing product reviews, social media posts, and other written content, but they often fell short in capturing the full spectrum of human emotions [2].

2. Fundamentals of multimodal sentiment analysis

Multimodal sentiment analysis (MSA) focuses on understanding and interpreting human emotions through the integration of multiple data modalities. These modalities typically include text, audio, and visual data, each contributing unique and complementary information to form a holistic view of sentiment. The fundamentals of MSA encompass several core concepts and processes essential for effective sentiment analysis.

2.1. Key modalities in MSA

Text: Textual data includes written words and sentences, which provide explicit content and context. Text is analyzed using traditional NLP techniques to extract sentiment-related features such as keywords, n-grams, sentiment lexicons, and syntactic structures.

Audio: Audio data captures speech, which includes features such as tone, pitch, volume, speech rate, and intonation. These features offer insights into the speaker's emotional state that may not be evident from text alone. For example, a calm or excited tone can significantly influence the perceived sentiment [3].

Visual: Visual data encompasses facial expressions, body language, and gestures. Facial expression analysis involves detecting and interpreting movements of facial muscles to identify emotions such as happiness, sadness, anger, and surprise. Body language and gestures further provide context and enhance the understanding of sentiment.

2.2. Fusion techniques

Combining data from different modalities, known as fusion, is a critical process in MSA. The primary approaches to fusion are early fusion, late fusion, and hybrid fusion, each with its strengths and challenges:

Early Fusion: This approach involves combining raw data or features from different modalities at the initial stage of processing. The integrated data is then analyzed together to detect sentiment. Early fusion can capture interdependencies among modalities but may also introduce noise and increase computational complexity.

Late Fusion: In late fusion, each modality is processed separately to extract sentiment-related information, and the results are merged subsequently. This method allows for independent optimization of each modality but may overlook interactions between them.[4].

Hybrid Fusion: Hybrid fusion incorporates elements from both early and late fusion techniques, combining features at multiple stages to harness the strengths of both methodologies. This method aims to capture interdependencies while maintaining the robustness of separate modality processing.

2.3. Feature extraction

Effective feature extraction is vital for capturing relevant information from each modality in MSA. Key techniques for feature extraction include:

Text: Techniques such as bag-of-words, TF-IDF, word embeddings (e.g., Word2Vec, GloVe), and contextual embeddings (e.g., BERT, GPT) are used to convert text into numerical representations that can be analyzed for sentiment.

Audio: Feature extraction involves analyzing various speech characteristics, including Mel-frequency cepstral coefficients (MFCCs), prosody features (pitch, energy, duration), and spectral features. These features are used to assess the emotional tone of speech [5].

Visual: Visual feature extraction includes detecting facial landmarks, extracting facial action units (AUs), and analyzing body movements and gestures. Techniques such as convolutional neural networks (CNNs) and facial recognition algorithms are employed for this purpose.

2.4. Data alignment

Alignment ensures that data from different modalities are synchronized, representing the same temporal or contextual points. Accurate alignment is essential for effective fusion and analysis in MSA. Techniques for alignment include:

Temporal Alignment: Synchronizing data streams based on timestamps to ensure that corresponding moments in time are analyzed together [6].

Contextual Alignment: Ensuring that multimodal data segments correspond to the same contextual elements, such as specific sentences or conversational turns.

3. Popular data sets

Multimodal sentiment analysis relies on various datasets that incorporate text, audio, and visual information to facilitate research and development. Among the most widely used datasets is CMU-MOSI (Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity), which includes opinion videos where speakers express their thoughts on different topics. Each video segment is annotated with sentiment labels, making it a valuable resource for training and evaluating sentiment analysis models. The CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity) dataset is an extension of CMU-MOSI, offering a larger and more diverse set of annotated opinion videos. MOSEI includes data from a wider range of speakers and topics, providing a comprehensive resource for studying multimodal sentiment analysis [7].

Another notable dataset is MELD (Multimodal EmotionLines Dataset), which consists of dialogue scenes from the TV show "Friends." MELD includes textual transcripts, audio recordings, and visual clips of the dialogues, all annotated with emotion labels. This dataset is particularly useful for analyzing sentiment and emotions in conversational contexts. IEMOCAP (Interactive Emotional Dyadic Motion Capture) is another key dataset, designed for emotion recognition research. It contains recordings of scripted and improvised dialogues between actors, capturing both audio and visual data along with detailed emotion annotations. IEMOCAP is often used to study the interplay between different modalities in conveying emotions [8].

These datasets are critical for advancing the field of multimodal sentiment analysis. They provide the necessary annotated data to train and evaluate models, helping researchers understand how different modalities contribute to sentiment and emotion recognition. By leveraging these datasets, researchers can develop more accurate and robust sentiment analysis systems that better reflect the complexities of human communication. However, the availability and quality of multimodal datasets remain a challenge, underscoring the need for continued efforts in data collection and annotation.

4. Techniques and models

Techniques and models in multimodal sentiment analysis encompass a variety of approaches designed to effectively integrate and analyze data from text, audio, and visual modalities.

In contrast, late fusion processes each modality separately to extract sentiment-related information, combining the results at a later stage. This method allows for independent optimization of each modality, maintaining their individual characteristics while ultimately merging their outputs. Late fusion can be advantageous when modalities have distinct, non-overlapping information, but it might miss capturing intricate interactions between them. Hybrid fusion techniques combine aspects of both early and late fusion, integrating features at multiple stages to leverage the benefits of both approaches. Hybrid fusion aims to capture interdependencies while maintaining the robustness of separate modality processing [9].

Deep learning models have become increasingly prominent in multimodal sentiment analysis. Recurrent neural networks (RNNs), especially Long Short-Term Memory (LSTM) networks, are commonly used to handle sequential data from text and audio modalities. LSTMs are adept at capturing temporal dependencies and patterns, making them suitable for speech and language analysis. Convolutional neural networks (CNNs) are frequently employed for visual data, excelling at extracting spatial features from images and video frames. More recently, transformer models, such as BERT (Bidirectional Encoder Representations from Transformers) and multimodal adaptations like VisualBERT, have shown promise in handling multimodal data through attention mechanisms that allow for effective integration and contextual understanding across modalities [10].

Multimodal representation learning is another critical area, focusing on how to represent and integrate data from different modalities effectively. Joint representation learning seeks to learn a common representation that combines features from all modalities into a unified vector space. This approach can facilitate more direct comparisons and integrations of multimodal data. Coordinated representation learning, on the other hand, ensures that the representations of different modalities are aligned and complementary but may still reside in separate spaces. Techniques such as canonical correlation analysis (CCA) and its deep learning variants are often used to achieve this alignment, enabling models to leverage the strengths of each modality while maintaining their distinct contributions to sentiment analysis.

The choice of techniques and models in multimodal sentiment analysis often depends on the specific application and the nature of the data. Each approach has its strengths and limitations, and ongoing research continues to refine these methods to improve accuracy, robustness, and interpretability. By understanding and leveraging the unique characteristics of each modality, researchers can develop more sophisticated and effective sentiment analysis systems that better capture the complexities of human emotions.

5. Challenges and applications

5.1. Challenges

5.1.1. Data integration

Multimodal sentiment analysis faces several significant challenges that complicate its implementation and efficacy. One of the primary challenges is data integration, as combining data from different modalities involves aligning varying formats, sampling rates, and noise levels. Effective integration methods are necessary to handle these differences and ensure that the multimodal data is synchronized and harmonized. Noise and redundancy further complicate this process; multimodal data often contains irrelevant or repetitive information, making it crucial to devise techniques that can eliminate superfluous elements and concentrate on the most relevant features for sentiment analysis.[11].

5.1.2. Computational complexity

Another challenge is the computational complexity associated with processing multimodal data. The need to analyze and fuse large volumes of diverse data requires significant computational resources, especially for real-time applications. Developing efficient algorithms and utilizing advanced hardware, such as GPUs and TPUs, are crucial to managing this complexity. Moreover, the interpretability of complex models used in multimodal sentiment analysis remains a critical issue. These models often function as black boxes, making it difficult to understand how they arrive at their conclusions. Enhancing model transparency and developing explainable AI methods are essential to build trust and facilitate the adoption of MSA systems.

5.1.3. Datasets

The availability of large, annotated multimodal datasets is another significant challenge. Such datasets are necessary for training and evaluating models, but they are often limited, hindering the development and validation of new techniques. The creation and sharing of comprehensive, high-quality datasets are vital for the advancement of the field. Additionally, the dynamic nature of human emotions and the context-dependent nature of sentiment make it difficult to create universally applicable models. Tailoring models to specific contexts and continually updating them with new data are necessary to maintain their relevance and accuracy.

5.2. Applications

Despite these challenges, the applications of multimodal sentiment analysis are vast and varied. In human-computer interaction, MSA enhances the interaction between humans and machines by enabling more natural and responsive communication. Virtual assistants, chatbots, and interactive systems can benefit significantly from the ability to recognize and respond to user emotions, improving user experience and satisfaction. In healthcare, MSA can play a crucial role in monitoring and understanding patients' emotional states. Applications in therapy and mental health can use MSA to assess emotions more accurately and provide personalized support and interventions.

Social media analysis is another significant application area, where MSA can provide deeper insights into public sentiment on various topics. By analyzing text, audio, and visual content from social media platforms, researchers and organizations can better understand trends, opinions, and reactions, aiding in decision-making and strategic planning. Customer service can also benefit from MSA, as understanding and responding to customer emotions can lead to improved satisfaction and loyalty. In the entertainment industry, MSA can enhance user experiences in gaming and virtual reality by detecting and responding to player emotions, creating more immersive and engaging environments.

6. Future trends

The future of multimodal sentiment analysis (MSA) is poised to see several significant advancements driven by ongoing research and technological innovations. One of the key trends is the development of enhanced fusion techniques. Researchers are continuously exploring more sophisticated methods for integrating multimodal data, aiming to capture the intricate interdependencies between text, audio, and visual inputs. These advanced fusion techniques will help in creating more accurate and robust sentiment analysis models, capable of understanding the complex ways in which humans express emotions.

Another promising trend is the push toward real-time processing of multimodal data. As computational power increases and algorithms become more efficient, the ability to analyze and interpret multimodal data in real-time will become more feasible. This capability is crucial for applications that require immediate responses, such as virtual assistants, interactive systems, and real-time monitoring in healthcare. Real-time MSA will enable more dynamic and responsive interactions, significantly enhancing user experiences across various domains.

The emphasis on explainable AI is also expected to grow within the field of MSA. As models become more complex, ensuring that their decisions and processes are transparent and understandable to users is increasingly important. Developing explainable models will help build trust and facilitate wider adoption of MSA technologies. This trend involves creating techniques that can elucidate how multimodal inputs contribute to the final sentiment analysis, making the models' workings more interpretable to users and stakeholders.

Personalization is another area poised for significant advancement. Adjusting MSA models to the needs of specific users can enhance the precision and relevance of sentiment analysis. By adapting models based on user-specific data and preferences, systems can provide more personalized interactions and recommendations. This trend is particularly relevant in applications such as customer service, healthcare, and entertainment, where understanding individual users' emotional nuances can vastly improve the quality of service and engagement.

7. Conclusions

Multimodal sentiment analysis (MSA) represents a pivotal advancement in understanding and interpreting human emotions across various modalities. Throughout this review, we have explored the fundamental concepts, challenges, applications, and future trends shaping the field of MSA.

Fundamentally, MSA integrates data from text, audio, and visual sources to provide a holistic understanding of sentiment. Techniques such as fusion, feature extraction, and data alignment are essential for effectively analyzing multimodal data. However, MSA faces challenges such as data integration, computational complexity, interpretability, and the scarcity of annotated datasets. Overcoming these challenges requires ongoing research and technological innovation.

Despite these obstacles, the applications of MSA are extensive and diverse. From improving human-computer interaction and healthcare to enhancing social media analysis and customer service, MSA has the potential to revolutionize various domains. By understanding and responding to human emotions more accurately, MSA can elevate user experiences, inform decision-making, and drive innovation across industries.

Looking to the future, several key trends are expected to shape the evolution of MSA. Enhanced fusion techniques, real-time processing, explainable AI, personalization, integration with other technologies, and the development of comprehensive datasets will drive advancements in the field. These trends underscore the importance of continued research and collaboration to realize the full potential of multimodal sentiment analysis.

In summary, multimodal sentiment analysis holds great promise for understanding the complexities of human emotions and communication. By leveraging diverse sources of information and employing advanced analytical techniques, MSA can provide valuable insights into sentiment and drive positive outcomes in various applications. As the field continues to evolve, MSA will play an increasingly vital role in enhancing human-machine interaction, decision-making processes, and overall well-being.

Despite the significant advancements in multimodal sentiment analysis (MSA), several limitations persist. Firstly, the integration of diverse data modalities often leads to increased computational complexity, making real-time processing challenging. Additionally, the quality and availability of annotated multimodal datasets remain limited, hindering the development and validation of robust models. The variability in data formats, sampling rates, and noise levels across modalities further complicates effective data fusion. Moreover, many current models function as black boxes, lacking interpretability, which makes it difficult to understand how they derive their conclusions and reduces trust in their predictions.

To address these limitations, future research in MSA should focus on several key areas. Enhancing fusion techniques to better capture interdependencies between modalities while reducing noise and computational load is crucial. The development of real-time processing capabilities will be pivotal for applications requiring immediate responses. Increasing the availability and quality of annotated multimodal datasets through collaborative efforts and advanced data collection methods will significantly bolster model training and evaluation. Additionally, the emphasis on explainable AI will help improve the transparency and interpretability of MSA models, building trust and facilitating broader adoption. Integrating personalized models that adapt to individual users' emotional nuances can also enhance the accuracy and relevance of sentiment analysis in various applications.

References

- [1] Alireza, G., & Karim, M. S. (2023). A comprehensive survey on deep learning-based approaches for multimodal sentiment analysis. *Artificial Intelligence Review*, 56(Suppl 1), 1479–1512.
- [2] Diksha, S., Ganesh, C., Babita, P., et al. (2022). A comprehensive survey on sentiment analysis: Challenges and future insights. *Journal of Intelligent & Fuzzy Systems*, 43(6), 7733–7763.
- [3] Hema, K., M. S. E., & T. S. (2022). A Comprehensive Survey on Sentiment Analysis in Twitter Data. *International Journal of Distributed Systems and Technologies (IJDST)*, 13(5), 1–22.
- [4] Marouane, B., Mohammed, K., & Abderrahim, B. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226.
- [5] Rajabi, Z., & Valavi, M. (2021). A Survey on Sentiment Analysis in Persian: a Comprehensive System Perspective Covering Challenges and Advances in Resources and Methods. *Cognitive Computation*, 13(4), 1–21.

- [6] Yadav, K., Kumar, N., Maddikunta, R. K. P., et al. (2021). A comprehensive survey on aspect-based sentiment analysis. *International Journal of Engineering Systems Modelling and Simulation*, 12(4), 279–290.
- [7] Kai, J., Bin, C., & Jing, F. (2024). A Robust Framework for Multimodal Sentiment Analysis with Noisy Labels Generated from Distributed Data Annotation. School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, 310023, China. 139(3), 2965–2984.
- [8] Hongbin, W., Chun, R., & Zhengtao, Y. (2024). Multimodal sentiment analysis based on cross-instance graph neural networks. *Applied Intelligence*, 54(4), 3403–3416.
- [9] Jing, Y., & Yujie, X. (2024). Bidirectional Complementary Correlation-Based Multimodal Aspect-Level Sentiment Analysis. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 20(1), 1–16.
- [10] Chandrasekaran, G., Dhanasekaran, S., Moorthy, C., et al. (2024). Multimodal sentiment analysis leveraging the strength of deep neural networks enhanced by the XGBoost classifier. *Computer methods in biomechanics and biomedical engineering*, 21–23.
- [11] Huchao, Z. (2024). Multimodal Sentiment Analysis Method Based on Hierarchical Adaptive Feature Fusion Network. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 20(1), 1–23.