International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Multimodal Sentiment Analysis: A Systematic Review

A. Martina Betsy[1], Dr. Sumathy Kingslin[2]

[1]Research Scholar, PG & Research Department of Computer Science, Quaid-E-Millath Government College for Women, Chennai - 2
Assistant Professor, Department of Commerce (Computer Applications), Women's Christian College, Chennai - 6
Email: *martina.betsy[at]gmail.com*

Associate Professor, PG & Research Department of Computer Science, Quaid-E-Millath Government College for Women, Chennai - 2
Email: *drsumathykingslin[at]gmail.com*

Abstract: ***Multimodal aspect-based sentiment analysis has emerged as a crucial research area, integrating information from diverse modalities like text, images, audio, and video to analyze human emotions and sentiments at the aspect level. This review synthesizes recent advancements in multimodal aspect-based sentiment analysis, highlighting innovative approaches to fusing text, images, and other modalities to gain deeper insights into human emotions. We examine the strengths and limitations of various methodologies, including interactive memory networks and cross-modal attention mechanisms, and discuss the challenges of bridging the semantic gap between modalities. Furthermore, we identify key areas for future research, including the need for standardized evaluation frameworks and the potential benefits of incorporating additional modalities. By providing a comprehensive overview of this rapidly evolving field, this review aims to inform and inspire future research in multimodal aspect-based sentiment analysis.***

Keywords: Sentiment Analysis, NLP (Natural Language Processing), Multimodal Fusion, Cross-Modal Attention, Interactive Memory Networks

1. Introduction

The proliferation of multimodal user-generated content on social media, online reviews, and forums has created a pressing need for advanced sentiment analysis techniques that can effectively leverage multiple data modalities, such as text, images, and audio. Traditional aspect-based sentiment analysis (ABSA) approaches, which focus exclusively on textual data, are limited in their ability to capture the rich emotional cues present in multimodal content.

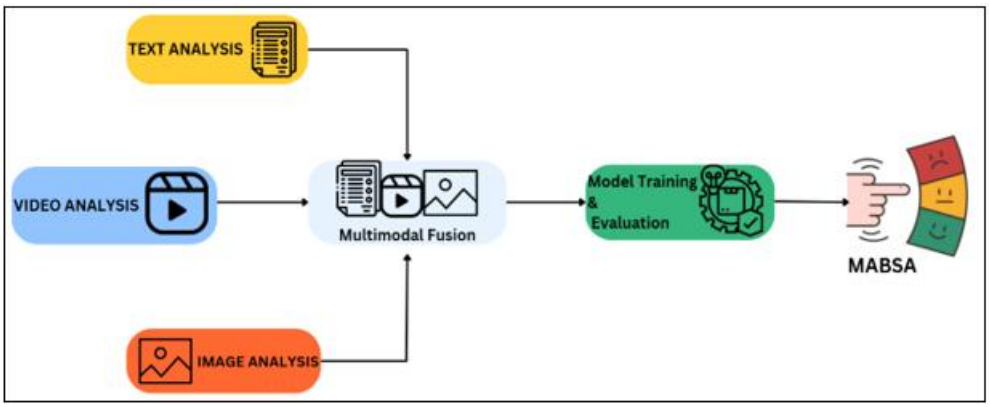

Figure 1: Diagrammatic Representation of MABSA Model

Multimodal aspect-based sentiment analysis (MABSA) has emerged as a vital research area that integrates multiple data modalities to determine sentiment polarity at the aspect level, offering more nuanced and context-aware emotion recognition than unimodal approaches. This review aims to provide a comprehensive overview of MABSA, covering its evolution, subtasks, key models, evaluation metrics, challenges, and future directions.

2. Review of Literature

Multimodal aspect-based sentiment analysis (MABSA) has emerged as a vital research area, integrating multiple data modalities to determine sentiment polarity at the aspect level. According to [1], MABSA has advanced significantly due to deep learning, with recent datasets and advanced models contributing to improved performance. However, existing research has highlighted several challenges and limitations, including the semantic gap between text and image representations [2], limited availability of annotated multimodal datasets [3], and the need for more effective cross-modal attention mechanisms [4].

Several advanced models have been proposed to address these challenges. For example, the Hierarchical Interactive Multimodal Transformer (HIMT) model [2] uses object detection and hierarchical interaction modules to improve aspect-text, aspect-image, and text-image interactions. The Multi-Interactive Memory Network (MIMN) model [5] captures interactions between modalities, advancing aspect-level sentiment analysis beyond text-only approaches. Other

notable models include the Bidirectional Complementary Correlation-Based Multimodal Aspect-Level Sentiment Analysis (BiCCM-ABSA) model [6] and the Attention Capsule Extraction and Multi-Head Fusion Network (EF-Net) [7].

Despite significant progress, future research should focus on addressing the challenges and limitations of MABSA. Potential areas of investigation include integrating multiple modalities [8], developing more effective cross-modal attention mechanisms and fusion approaches [9], and establishing standardized task decompositions, evaluation measures, and publicly available datasets [10]

1) Evolution of Multimodal Aspect-Based Sentiment Analysis: From Text-Only to Multimodal Approaches

Traditionally, aspect-based sentiment analysis (ABSA) focused exclusively on textual data, identifying aspects (e.g., 'battery', 'screen') and estimating sentiment polarity for each aspect [3][5]. However, with the proliferation of multimodal user-generated content, researchers recognized the limitations of text-only methods, particularly their inability to capture richer emotional cues present in images and other modalities [1][5][8]. Multimodal aspect-based sentiment analysis (MABSA) has emerged as a vital research area that integrates multiple data modalities to determine sentiment polarity at the aspect level.

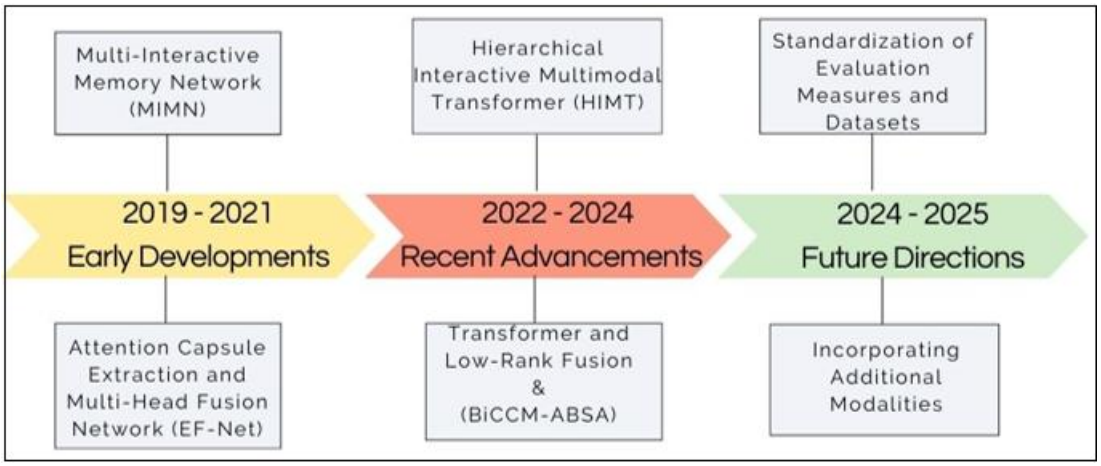

Figure 2: Evolution of MABSA

a) Early Developments

The early developments in MABSA saw the proposal of several innovative approaches. The Multi-Interactive Memory Network (MIMN) phase, proposed by N. Xu et al. in 2019 [5], focuses on capturing interactions between modalities for aspect-based multimodal sentiment analysis. This was followed by the Attention Capsule Extraction and Multi-Head Fusion Network (EF-Net) phase, introduced by D. Gu et al. in 2021 [7], which emphasizes targeted aspect-based multimodal sentiment analysis using attention mechanisms and capsule networks. Additionally, the Arabic ABSA phase, highlighted by R. Obiedat et al. in 2021 [3], addresses the challenges of limited annotated corpora and domain diversity in Arabic aspect-based sentiment analysis.

b) Recent Advancements

Recent advancements in MABSA have led to the development of more sophisticated models. The Hierarchical Interactive Multimodal Transformer (HIMT) phase, proposed by J. Yu et al. in 2023 [2], uses object detection and hierarchical interaction modules to improve aspect-text, aspect-image, and text-image interactions. Furthermore, the Transformer and Low-Rank Fusion phase, introduced by M. Jin et al. in 2024 [4], focuses on aspect-based sentiment analysis on multimodal data using transformer and low-rank fusion approaches. The Bidirectional Complementary Correlation-Based Multimodal Aspect-Level Sentiment Analysis (BiCCM-ABSA) phase, proposed by J. Yang and Y. Xiong in 2024 [6], employs cross-modal attention mechanisms and gating strategies to align text-image features.

c) Current Challenges and Future Directions

Despite the progress made in MABSA, there are still several challenges that need to be addressed. The Standardization phase, highlighted by H. Zhao et al. in 2024 [10], emphasizes the need for standardized task decompositions, evaluation measures, and publicly available datasets in MABSA. Moreover, the Incorporating Additional Modalities phase, discussed by T. Chen in 2025 [8], explores the potential benefits of incorporating additional modalities and the need for further research in MABSA.

2) Defining MABSA and Its Subtasks

MABSA extends ABSA by incorporating multiple modalities to assess aspect-level sentiments. The subtasks in MABSA typically include:

- Aspect Extraction: Identifying aspect terms or categories from multimodal content.
- Aspect Sentiment Classification: Determining sentiment polarity for each aspect using multimodal inputs.
- Aspect Sentiment Pair Extraction: Extracting pairs of aspects and their corresponding sentiments from multimodal data. [2][10]

Recent surveys have highlighted the lack of established task decompositions and evaluation measures for these subtasks in the multimodal setting, underscoring the need for standardized benchmarks and comprehensive corpora [3][5][10]

3. Key Models and Methodological Advances

a) Interactive and Attention-Based Architectures

Researchers have developed several advanced architectures to address the unique challenges of MABSA:

- Multi-Interactive Memory Network (MIMN): This model supervises both textual and visual information with respect to the given aspect, learning interactive influences across modalities as well as within each modality [5]
- Hierarchical Interactive Multimodal Transformer (HIMT): HIMT extracts salient semantic features from images using object detection, models hierarchical interactions among aspect-text and aspect-image pairs, and introduces auxiliary reconstruction modules to bridge the semantic gap between text and image representations [2]
- BiCCM-ABSA: Leveraging bidirectional complementary correlation, this transformer-based model employs cross-modal attention mechanisms and gating strategies to align text-image features for more accurate sentiment classification [6]
- ABSA-TLRF: This model utilizes cross-modal alignment via attention mechanisms and low-rank fusion to integrate global and local information between modalities, leading to improved emotion fusion results [4]
- EF-Net for TABMSA: Employs multi-head attention for textual data and ResNet-152 for image processing, integrating capsule networks to capture interactions among multimodal inputs for targeted aspect-based analysis [7]

b) Comparative Overview of Recent Models

The Table I below summarizes key multimodal sentiment analysis models, highlighting their modalities, techniques, and notable contributions:

Table I: Comparative Overview of Recent Models

Model	Modalities Used	Key Techniques	Notable Contributions
MIMN	Text, Image	Interactive Memory Networks	Supervises cross-modality influences
HIMT	Text, Image	Hierarchical Transformer	Object-level semantics; semantic gap
BiCCM-ABSA	Text, Image	Cross-modal Attention	Bidirectional feature alignment
ABSA-TLRF	Text, Image	Low-Rank Fusion, Attention	Global-local information integration
EF-Net (TABMSA)	Text, Image	Capsule Networks, MHA	Targeted aspect-based analysis

These models demonstrate the effectiveness of multimodal approaches in sentiment analysis, leveraging techniques like interactive memory networks, hierarchical transformers, and cross-modal attention to improve performance. By integrating text and image modalities, these models can capture richer emotional cues and provide more accurate sentiment analysis results.

4. Discussion of Findings

The analysis of multimodal aspect-based sentiment analysis (MABSA) reveals several key challenges that impact its effectiveness. These challenges can be broadly categorized into five areas:

Figure 3: Key findings of MABSA

a) Semantic Gap between Modalities

The semantic gap between text and image representations is a persistent challenge in MABSA. This gap arises from the differences in how text and images convey meaning, making it difficult for models to align semantic concepts across modalities. To address this challenge, existing models employ auxiliary reconstruction modules and hierarchical interaction mechanisms. For instance, models such as the Hierarchical Interactive Multimodal Transformer (HIMT) use object detection and hierarchical interaction modules to improve aspect-text, aspect-image, and text-image interactions [2]. Similarly, the Bidirectional Complementary Correlation-Based Multimodal Aspect-Level Sentiment Analysis (BiCCM-ABSA) model leverages bidirectional complementary correlation and cross-modal attention mechanisms to align text-image features [6].

b) Object-Level Semantics and Contextual Integration

Another challenge in MABSA is the need for object-level semantics and contextual integration. Many early approaches overlooked object-level semantics in images or focused narrowly on aspect-text and aspect-image interactions. Recent models, however, incorporate object detection and context-aware fusion to capture more nuanced relationships between aspects and multimodal content. For example, the HIMT model uses object detection to extract salient semantic features from images, while the Transformer and Low-Rank Fusion approach introduced by M. Jin et al. focuses on aspect-based sentiment analysis on multimodal data using transformer and low-rank fusion approaches [2][4].

c) Lack of Standardized Datasets

The limited availability of annotated multimodal datasets, especially in languages other than English, hampers progress in MABSA research. For instance, Arabic ABSA faces challenges due to a lack of annotated corpora and domain

diversity [3]. This highlights the need for more diverse and annotated datasets to support research in MABSA.

d) Expansion to New Modalities

While most current work in MABSA focuses on text and images, there is growing interest in incorporating additional modalities such as audio. This could further improve sentiment detection precision and enable more comprehensive analysis of multimodal content [8].

e) Standardization and Benchmarking

Finally, there is a clear need for standardized task decompositions, evaluation measures, and publicly available datasets to facilitate fair comparisons and accelerate progress in MABSA research. Standardization would enable researchers to compare the performance of different models more effectively and identify areas for improvement [3][5][10].

5. Conclusion and Future Work

Multimodal aspect-based sentiment analysis (MABSA) has emerged as a vital research area, integrating multiple data modalities to determine sentiment polarity at the aspect level. This comprehensive review has highlighted recent advancements in MABSA, including innovative approaches to fusing text, images, and other modalities. Despite significant progress, MABSA still faces several challenges, including the semantic gap between modalities, limited availability of annotated datasets, and the need for standardized evaluation frameworks. By addressing these challenges and exploring new research directions, researchers can develop more effective MABSA systems that can capture richer emotional cues and provide more accurate sentiment analysis results. This review aims to inform and inspire future research in MABSA, promoting further advancements in this rapidly evolving field.

References

[1] S. Lai et al., "Multimodal sentiment analysis: A survey," ArXiv, 2023.
[2] J. Yu et al., "Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis," IEEE Transactions on Affective Computing, 2023.
[3] R. Obiedat et al., "Arabic aspect-based sentiment analysis: A systematic literature review," IEEE Access, 2021.
[4] M. Jin et al., "Aspect based sentiment analysis on multimodal data: A transformer and low-rank fusion approach," 2024 4th International Conference on Computer Communication and Artificial Intelligence (CCAI), 2024.
[5] N. Xu et al., "Multi-interactive memory network for aspect based multimodal sentiment analysis," 2019.
[6] J. Yang and Y. Xiong, "Bidirectional complementary correlation-based multimodal aspect-level sentiment analysis," Int. J. Semantic Web Inf. Syst., 2024.
[7] D. Gu et al., "Targeted aspect-based multimodal sentiment analysis: An attention capsule extraction and multi-head fusion network," IEEE Access, 2021.
[8] T. Chen, "A review of multimodal aspect-based sentiment analysis," Advances in Engineering Innovation, 2025.
[9] M. Jin et al., "Aspect based sentiment analysis on multimodal data: A transformer and low-rank fusion approach," 2024.
[10] H. Zhao et al., "A survey on multimodal aspect-based sentiment analysis," IEEE Access, 2024.