International Journal on Semantic Web and Information Systems
Volume 22 • Issue 1 • January-December 2026

A Semantically-Driven Multimodal Sentiment Analysis Framework With Temporal and Synergistic Attention

Yonghong Xie
https://orcid.org/0009-0003-2507-214X
Guangzhou City University of Technology, China
Received: December 17th, 2025 | **Accepted:** February 10th, 2026

ABSTRACT

Multimodal sentiment analysis aims to identify emotional tendencies from text, audio, and visual data, but existing methods often struggle with weak temporal modeling within modalities and shallow cross-modal fusion. The proposed temporal modeling and synergistic attention–based multimodal sentiment analysis framework can address these issues. Word-level features are first extracted from all modalities, then modeled using a state-gated long short-term memory network combined with multi-head attention to capture temporal emotional dynamics while filtering noise. A hierarchical collaborative attention mechanism is further designed to enable deep, fine-grained cross-modal semantic interactions. Experiments on the Carnegie Mellon University multimodal corpus of sentiment intensity and multimodal opinion sentiment and emotion intensity datasets show that the modeling and synergistic attention-based multimodal sentiment analysis framework achieves an F1 score of 87.3% and an mean absolute error of 0.426, it achieves a 1.2–1.5% improvement while simultaneously reducing mean absolute error to its lowest value, outperforming existing state-of-the-art approaches and demonstrating its effectiveness in modeling complex multimodal emotions.

KEYWORDS

Deep Learning, Sentiment Analysis, Multi-Modal Fusion, Temporal Modeling

INTRODUCTION

With the rapid development of social media and multimodal data, sentiment analysis has become a significant research direction in natural language processing and multimodal learning. Tan et al. (2023) systematically summarized research methods, datasets, and future directions in sentiment analysis, while Zhang et al. (2016) focused on the technological evolution of sentiment analysis and opinion mining, providing a theoretical foundation for this study. Sentiment analysis aims to identify and analyze emotional tendencies in text, speech, images, and other data, finding extensive applications in public opinion monitoring, consumer behavior analysis, and human–computer interaction (Naing & Udomwong, 2024; Wang et al., 2025). However, unimodal sentiment analysis methods are often constrained by information scarcity and modality limitations, making it challenging to comprehensively capture the diverse features of emotional expression (Anna et al., 2025; Purnamasari et al., 2024). Consequently, multimodal sentiment analysis has emerged as a research hotspot, integrating

DOI: 10.4018/IJSWIS.402041

information from different modalities to more accurately uncover complex patterns in emotional expression (Yang et al., 2024).

Monomodal sentiment analysis methods have yielded rich results within their respective modalities. For instance, text sentiment analysis extracts semantic features using deep neural networks and pre-trained language models; speech sentiment analysis employs acoustic features such as pitch, energy, and rhythm variations for emotion recognition, while visual sentiment analysis focuses on extracting facial expressions and motion patterns (Ren, 2024). However, these approaches exhibit clear limitations: text models may overlook complementary emotional information from audio and visual signals, while audio and visual methods struggle to capture sentiment at the semantic level, resulting in an incomplete understanding of emotions.

In recent years, multimodal sentiment analysis methods have gradually emerged as a research hotspot. Existing studies primarily focus on two directions: intra-modal modeling and inter-modal fusion. Regarding intra-modal modeling, Zhu et al. (2022) proposed the Image–Text Interaction Network to enhance feature representation through cross-modal alignment. However, its text modeling relies solely on basic word embeddings, and image features lack high-level semantic information, resulting in limited internal representations. Regarding cross-modal fusion, Wang et al. (2023) introduced the dual perspective fusion network to model affective associations at both global and local levels while extracting fine-grained features through graph structures. However, its fusion remains at shallow alignment, failing to fully capture deep cross-modal dependencies and neglecting the temporal dynamics of affect.

Synthesizing existing work reveals three core challenges in multimodal sentiment analysis:

- **insufficient intramodal feature modeling:** Contextual dependencies and dynamic changes in sentiment within time series remain under-explored.
- **limited intermodal fusion mechanisms:** Traditional approaches often rely on concatenation or simple weighting, struggling to achieve deep cross-modal collaboration.
- **neglected temporal dynamics of emotion:** Speech and visual modalities exhibit pronounced emotional evolution over time; ignoring temporal dependencies weakens model predictive capabilities (Cheng et al., 2023; Saima et al., 2025).

Therefore, there is an urgent need for a multimodal sentiment analysis framework that simultaneously enhances intra-modal temporal modeling and intermodal deep collaborative fusion. To address the aforementioned issues, this paper proposes a temporal modeling and synergistic attention–based multimodal sentiment analysis (TMSA-AM) framework. Its primary innovations are reflected in the following three aspects. First, although recent transformer-based multimodal models (e.g., the multimodal transformer) can capture global dependencies, their ability to model local emotional transient features remains limited. TMSA-AM employs a dual-scale deep temporal network (short-term and long-term) to explicitly distinguish between local transient and overall trend-based emotional evolution, thereby more precisely depicting the dynamic changes in sentiment.

Second, unlike most transformer models that primarily employ unidirectional or shallow alignment multimodal attention, the proposed collaborative attention mechanism simultaneously models interactions and mutual dependencies across modalities. This achieves deep cross-modal coordination, significantly enhancing multimodal information fusion capabilities. Finally, TMSA-AM integrates deep temporal modeling with a collaborative attention mechanism, enabling the model to not only capture emotional evolution over time but also dynamically adjust modality weights. This achieves synergistic enhancement of temporal information and cross-modal associations, distinguishing itself from traditional transformer models where temporal and modal processing are relatively independent.

The structure of this paper is as follows: research background and motivation; relevant studies on unimodal and multimodal sentiment analysis; overall framework and module designs of the proposed TMSA-AM method; experimental evaluations using the Carnegie Mellon University multimodal

Based on The Document-Link and Time-Clue Relationships Between Blog Posts to Improve the Performance of Google Blog Search

Lin-Chih Chen (2019). *International Journal on Semantic Web and Information Systems (pp. 52-75).*

www.irma-international.org/article/based-on-the-document-link-and-time-clue-relationships-between-blog-posts-to-improve-the-performance-of-google-blog-search/217012