International Journal of Information Technologies and Systems Approach
Volume 17 • Issue 1

A Multimodal Sentiment Analysis Method Integrating Multi-Layer Attention Interaction and Multi-Feature Enhancement

Shengfeng Xie, Henan Institute of Technology, China*
Jingwei Li, Henan Institute of Technology, China

ABSTRACT

To address issues related to the insufficient representation of text semantic information and the lack of deep fusion between internal modal information and intermodal information in current multimodal sentiment analysis (MSA) methods, a new method integrating multi-layer attention interaction and multi-feature enhancement (AM-MF) is proposed. First, multimodal feature extraction (MFE) is performed based on RoBERTa, ResNet, and ViT models for text, audio, and video information, and high-level features of the three modalities are obtained through self-attention mechanisms. Then, a cross modal attention (CMA) interaction module is constructed based on transformer, achieving feature fusion between different modalities. Finally, the use of a soft attention mechanism for the deep fusion of internal and intermodal information effectively achieves multimodal sentiment classification. The experimental results CH-SIMS and CMU-MOSEI datasets show that the classification results of proposed MSA method are significantly superior to other advanced comparative methods.

KEYWORDS

Cross Modal Attention, Multi-Feature Enhancement, Multi-Layer Attention Interaction, Multimodal Sentiment Analysis, Soft Attention Mechanism

INTRODUCTION

Social media provides users with convenient channels for information dissemination and collection (Pang et al., 2021; Wu et al., 2020; Yang et al., 2022; Zhang et al., 2021). With the continuous progress and development of related fields, the majority of opinions expressed on the internet now rely on various digital media technologies, including images, voice, and video, to offer more vivid and three-dimensional information content (Dayyala et al., 2022; Wen et al., 2021; Wu et al., 2022). These contents can influence the real world through dissemination and diffusion, possessing significant research value in various fields such as society, economics, politics, and others (Ahmed et al., 2022; Basiri et al., 2021; Han et al., 2021; Lai et al., 2021; Silva et al., 2022; Su et al., 2020).

Sentiment analysis refers to the process of extracting, analyzing, inductively processing, and mining subjective data with sentiment colors. One crucial task of sentiment analysis is to classify

DOI: 10.4018/IJITSA.335940 *Corresponding Author

sentiment. Early research primarily utilized single modal data for sentiment analysis, such as image, video, or text modalities (Mahabadi et al., 2021; Wan et al., 2021; Yin et al., 2022; Zhang & Yin, 2022; Zhao et al., 2021). However, when faced with massive multimodal information, although single modal data sentiment analysis has achieved success in customer satisfaction analysis and measuring voting intentions in recent years, it cannot effectively handle multimodal data due to the diversity of information, giving rise to multimodal sentiment analysis (MSA) (Cheema et al., 2021; Yang et al., 2021; Yu et al., 2021).

MSA is a computational study of viewpoints and sentiment states based on single modal sentiment analysis, using data composed of text, images, audio, or even video data. Social media is a vast source of opinions for various products and user services. The effective combination of multiple modal information can better guide analysis (Jiang et al., 2020; Li et al., 2021; Xu et al., 2022). Sentiment analysis of videos can compensate for the shortcomings of sound and vision in text sentiment analysis, and speech and facial expressions provide important clues for better recognizing the sentiment state of opinion holders. This has significant practical implications for applications such as public opinion monitoring, product recommendations, and research on user feedback (Ortis et al., 2022; Wang et al., 2020).

With the advancement of multimodal technology, contemporary academic research on sentiment analysis tasks predominantly centers around leveraging multimodal technology to enhance the accuracy of models in these tasks. However, prevailing MSA methods based on deep learning frequently encounter challenges such as inadequate representation of text semantic information, the need to balance global and local features in image modalities, and the absence of profound fusion of internal or intermodal information.

To better address the aforementioned issues and enhance the accuracy of MSA, a novel MSA method integrating multiple feature enhancements and multi-layer attention interaction is proposed. The innovation of this method, in comparison to conventional sentiment analysis approaches, can be summarized as follows:

1) For text modality, the RoBERTa model is used to extract shallow text features in the embedding layer. A representation dictionary is constructed using the Masked Language Model (MLM) to augment knowledge and enhance the semantic features of the text modality.
2) For the image modality, the ResNet and ViT models are fused to comprehensively consider both global and local image features. In addition, the incorporation of body movements, gender, and age features in video modalities enriches the feature representation of image modalities.
3) Regarding multimodal fusion, a network structure based on the deep fusion of multi-layer attention interaction is introduced. Through the multi-level interaction of the self-attention mechanism, improved Cross Modal Attention (CMA) mechanism, and soft attention mechanism, deep fusion of internal and intermodal information is achieved.

RELATED WORK

MSA has emerged as a research hotspot in recent years. Focusing on target multimodal sentiment classification and incorporating attention modules based on text and image, An et al. (2023) constructed a model capable of achieving feature fusion and extracting information correlations. On this basis, an improved universal model for target multimodal sentiment classification was proposed based on image semantic description (ITMSC). However, this method exclusively delved into an in-depth analysis of the text, neglecting fusion analysis on user expressions of different types of information. By combining an adaptive mask memory capsule network and a self-attention mechanism, sentiment analysis and clustering were constructed. Building upon this, Zhang et al. (2023) proposed the VECapsNet model and provided a corresponding MSA model based on interactive learning. However, this approach did not consider the semantic information within the image during the analysis process and lacked