International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May-2024 SJIF Rating: 8.448 ISSN: 2582-3930

A Systematic Literature Review on Multimodal Sentiment Analysis using Deep learning techniques

Vinitha V[1], Dr.S.K. Manju bargavi[2]

[1]*Research Scholar School of Computer Science and IT Jain (Deemed to be University), Bengaluru, India*
[2]*Professor School of Computer Science and IT Jain (Deemed to be University), Bengaluru, India*

Abstract - Multimodal sentiment analysis (MSA) has emerged as an effective approach for assessing human emotions through several modalities, particularly verbal and non-verbal forms. Unimodal methods often fail to adequately capture the nuances of human affective states, whereas multimodal frameworks offer richer insight into sentiment by leveraging complementary sources of information. This study provides an overview of the MSA, fusion methods, and cutting-edge deep learning methods used in the development of MSA. It also discusses the datasets available and high-level deep learning models that have been implemented and, based on existing issues, provides future scope on developing MSA systems that are robust and scalable for real-world use cases.

Key Words: Multimodal sentiment analysis, fusion, deep learning, verbal, nonverbal.

1.INTRODUCTION

The Opinion mining or sentiment analysis (SA), is used to analyse a user's sentiment, opinions, and emotions from various modalities like video, audio, and text [1]. From the perspective of daily life, human beings often perform with different emotions in different scenarios. Sentiment also influences the action of people on choices according to psychological states. For instance, people tend to buy more online when they're feeling happy, something companies would very much like to know. Sentiment analysis, thus, can aid people to better understand a customer's sentiment by extraction of related information in multi-modal data, which should boost and enhance an individual's organization's profitability and productivity, providing this business a benefit on professional perception in the current world

Some SA applications include biomedical, and business and become well connected by healthcare, companies, researchers, education, government [3]. Emotional predictions from other domains, such as visual and speech, have been estimated as strong platforms for their high performances [4]. Therefore, multi-modal SA takes several scenarios into account and does not merely focus on the single SA of the image, video, or text. In the current times, studies have aimed to recognize sentiment expressed in multimedia across multi-modal modalities such as text, visual, and audio data. Accordingly, the web has also been evolving from a text- and hypertext-based community to a multimedia community [5]. Figure 1 depicts the fundamental aspects of deep learning algorithms corresponding to each category.

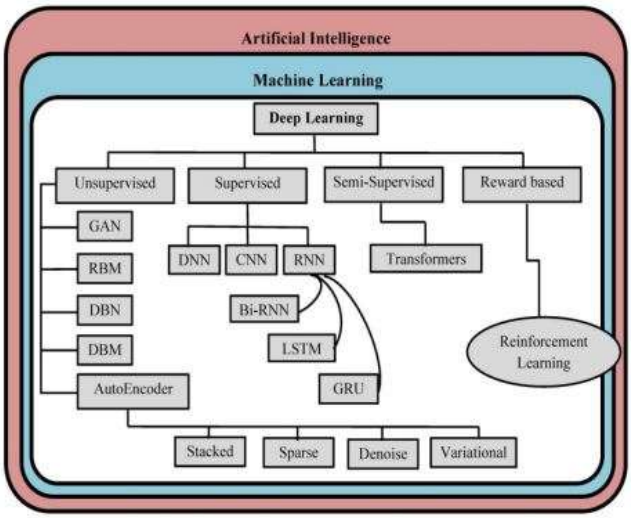

Fig -1: Taxonomy of Deep Learning Algorithms

A hierarchical framework comprising AI, ML, and DL modelling is proposed in Figure 1 for tasks such as social media analysis for MSA. Al covers machine learning (ML) at the highest level, which is concerned with enabling techniques to discover patterns in data. The machine's ability to do better at some tasks based on exposure to data and experience in ML's. In DL, which is a subfield of ML, multi-layered neural networks are employed for the solution of progressively complex problems. Four types of DL approaches are presented in the below figure: semi-supervised, supervised, reward-based, and unsupervised. This comprehensive hierarchy illustrates the application of advanced deep learning methodologies, specifically RNNs, transformers, and autoencoders, in the efficient analysis of substantial quantities of textual and audiovisual social media data for the detection of MSA.

2. LITERATURE REVIEW

Aslam et al. (2023) - Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks.

This research presented a unique technique known as the Attention-based Multimodal Sentiment Analysis and Emotion Recognition (AMSAER) algorithm. Inter-modalities correlations and intra-modalities discrimination features in text, audio, and visual conditions are influenced by these frameworks. This can be followed by the development of a more complex hierarchical

algorithm, which includes intermediated fusion for the purpose of learning hierarchic correlations among the conditions at the trimodal and bimodal levels. In conclusion, this framework facilitates multimodal SA and emotion recognition by integrating four distinct methods through decision-level fusion.

Zhang et al. (2022) - Deep emotional arousal network for multimodal sentiment analysis and emotion recognition.

This study includes a Deep Emotional Arousal Network (DEAN) is introduced, which incorporates the temporal dependency into the parallel framework of the transformer and can simulate emotion coherence. The DEAN method that has been proposed comprises three mechanisms. For instance, a cross-modal transformer has been created to simulate the functions of the perceptive investigation human method. A multi-modal BiLSTM process has been enhanced to resemble the cognitive comparator, and a multi-modal gating block has been introduced to replicate the activation mechanism in the human emotion arousal technique.

Gandhi et al. (2023) - Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions.

This study discusses a variety of features related to the advancement of MSA and the various fusion techniques employed in it, including early fusion, late fusion, hybrid fusion, and attention-based fusion. They categorize new developments in MSA architectures and even specify the benefits and drawbacks of each category. This systematic review also indicates potential directions for future research, with a particular emphasis on the prospective routine settings for MSA.

Dharma et al. (2021) - A Deep Learning Using DenseNet201 to Detect Masked or Non-masked Face.

This research includes Hierarchical Attention Network (MDHAN) is recommended. User posts augmented by the incorporation of supplementary Twitter features have been considered. User submissions are encoded, semantic sequence patterns are identified, and the importance of each word and tweet is evaluated by two tiers of attention processing at both the word and tweet levels. The hierarchical attention model, owing to its structured framework, may discern patterns that produce interpretable outcomes. The advantages of combining deep learning with multi-faceted data are shown by experimental findings, which indicate that MDHAN surpasses several established and dependable baseline methods.

Das et al. (2023) - Multimodal sentiment analysis: a survey of methods, trends, and challenges.

This study encompasses research trends, methodologies, and the current state of the art in MSA. The discussion highlighted the transition from unimodal designs to multimodal ones, with a focus on transformers. This survey indicates that MSA is perceived to have numerous uses in marketing, healthcare, and political opinion polls.

Kalaiyarasi et al. (2022) - Emotion Recognition using Multimodality and Deep Learning. ACM Transactions on Asian and Low-Resource Language Information.

This study introduces a tri-modal methodology for emotion identification, incorporating face expressions, textual emotions, and vocal data. Their method, emphasizing facial expressions, was validated on 50 students in a classroom setting, demonstrating highest accuracy in emotion recognition, indicating the modality's appropriateness for the academic context.

Tejaswini et al. (2024) - Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model.

This research developed an innovative hybrid deep learning neural network architecture utilizing enhanced textual representations, termed Fasttext CNN with LSTM. This FCL method incorporates rapid text embedding for enhanced representation of out-of-vocabulary (OOV) terms with semantic information, a convolutional neural network (CNN) framework for global data elimination, and LSTM architecture for the removal of local features based on dependency.

Perti et al. (2023) - Cognitive Hybrid Deep Learning-based Multi-modal Sentiment Analysis for Online Product Reviews.

This research illustrates the application of a deep learning framework for sentiment analysis by directly eliminating characteristics from the text. This technique utilizes convolutional neural networks (CNN) with many convolutional layers. This learning experience involves LSTMs and experimenting with various ensemble models to achieve optimal results. The method uses an n-gram-based word embedding methodology to create a machine-level representation of words.

Wang et al. (2023) - TEDT: transformer-based encoding–decoding translation network for multimodal sentiment analysis.

This study presents a multimodal encoder-decoder translation network utilizing a transformer was proposed, incorporating a mixed encoder-decoder model that processes text for primary information and images and sound for supplementary information. To mitigate the adverse impact of non-natural language data on natural language data, a modalities reinforcement cross-attention unit has been proposed to convert non-natural characteristics into natural language features, thereby enhancing their quality and facilitating the integration of multi-modal features. Moreover, the dynamical filter mechanism isolates the informational error occurring in cross-modal communication to enhance the ensuing output.

Kusal et al. (2024) - Pre-Trained Networks and Feature Fusion for Enhanced Multimodal Sentiment Analysis.

This study established a novel model for sentiment analysis that integrates pre-trained models with a feature fusion of emoticon-based images and textual data. These algorithms differentiate themselves from conventional sentiment analysis by employing emoticon data and image attributes to identify more subtle emotional distinctions, such as extreme positivity, extreme negativity, and neutrality. Feature fusion models are employed in both scenarios, integrating the extracted features from images and text to provide a more comprehensive representation.

3. Research gaps

Modality Alignment & Synchronization: Difficulties in synchronizing asynchronous modalities (e.g., speech lags after facial emotion). Most models assume well-synchronized data, which is unrealistic in noisy environments.

Modality Imbalance & Missing Modalities

Performance declines when one or more modalities (audio, video, or text) are missing or corrupted. There is still a lack of robust methodologies for imperfect multimodal data.

Scalability & Real-Time Processing

Numerous fusion architectures are computationally intensive, which restricts the development of real-time applications.

Fusion Strategies Still Limited

Although early, late, and hybrid fusion approaches exist, the dynamic weighting of modalities based on context is still in the process of evolving. however, they frequently overfit to particular datasets.

4. Various Datasets used

This section introduces several commonly used datasets in multimodal sentiment analysis. Recently, most multimodal sentiment analysis datasets are collected from the internet, which mainly include comments from different video-sharing websites. In addition, the authors sometimes also built their own dataset according to their goals. Below are few examples of MS datasets:

1. **CMU-MOSI (Multimodal Opinion Sentiment Intensity) [11]:** Among the first multimodal sentiment analysis benchmarks, is the CMU-MOSI dataset. The analysis includes 2,199 opinion videos obtained from 93 YouTube movie reviews. Each label has intensity levels ranging from –3 (very negative) to +3 (strongly positive), which enables nuanced polarity analysis instead of binary classification. The dataset contains the corresponding aligned text, features, and facial expressions. It is widely utilized to evaluate deep learning models inferring modality relations. Because of its relatively small size, CMU-MOSI is extensively employed to evaluate interpretability, feature fusion, and fine-grained sentiment regression.

2. **CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity) [12]:** The CMU-MOSEI dataset extends MOSI. It is composed of 23,454 annotated YouTube videos that represent about 5,000 speakers. Each speech has sentiment polarities (–3 to +3) and six emotions: anger, disgust, fear, happiness, sadness, and surprise. It is a challenging benchmark dataset for multimodal transformer, graph neural network, and attention-based fusion model testing and allows for training deep multimodal architecture on a large scale. With transcripts, acoustic and face visual features, MOSEI spans all modalities. Its broad and varied scale enables generalizable multimodal sentiment analysis beyond MOSI.

3. **MELD (Multimodal EmotionLines Dataset) [13]:** The MELD dataset is constituted of dialogues from Friends and developed for multi-party emotion detection. It includes 13,000 words from 1,433 conversations. Seven emotion labels (anger, disgust, fear, joy, neutral, sadness, and surprise) as well as three sentiment polarities are provided for each utterance. It provides multimodal contextual emotion modeling with audio, video, and transcripts with MELD. It captures conversational dynamics across multiple speakers, which is crucial for dialogue-based sentiment and emotion recognition. MOSI and MOSEI are monologue-based. It has been widely used in conversational AI, chatting NLP with empathy and context, and sentiment analysis.

4. **IEMOCAP (Interactive Emotional Dyadic Motion Capture Dataset) [14]:**It as a major acted conversational corpus employed in emotion and sentiment studies. It contains 12 hours of audio-visual recordings of 10 actors improvising and scripting. Emotions are labeled in about 10,000 utterances and include the categories anger, happiness, sadness, frustration, enthusiasm, and neutrality. Speech, facial motion capture, and transcriptions provide IEMOCAP with rich multimodal information. It is popularly used for studies in conversational emotion recognition, multimodal fusion, and affective computing. Its annotations are high-quality and multimodal, and even though it has been acted on, it still serves as the benchmark dataset today.

5. **DAIC-WOZ (Distress Analysis Interview Corpus – Wizard of Oz dataset) [15]:** Mental health research uses the multimodal DAIC-WOZ dataset. In 189 semi-structured clinical interviews, participants interact with a managed virtual agent. This dataset allows multimodal behavioral analysis, including audio, facial expressions, bodily movements, and transcripts. Depression and psychological distress markers are annotated, making it useful for automatic depression identification and healthcare affective computing.

6. **MuSe Challenge Datasets [16]:** The MuSe (Multimodal Sentiment Analysis) provides annual benchmark datasets on novel affective computing topics, video speeches received continual emotive annotations like valence, arousal. Cross-cultural humor detection and social perception (like ability, dominance, and trustworthiness) were included in MuSe 2024. The datasets evaluate multimodal deep learning algorithms under tough, real-world settings using audio-visual recordings with transcripts. More than polarity detection, these standards emphasize complex affective and social qualities of multimodal data.

5.Various Baseline models

1. **Bidirectional Bridge Fusion Network (BBFN) [17]** - Design a text-biased Modality Complementation Layer for modality fusion and separation which benefits from the previous relevant modality.
2. **Text-Enhanced Transformer Fusion Network (TETFN) [18]:** Uses text as the dominating modality to guide non-textual modalities in fusion.
3. **Multi-Task Multi-Modal Distillation (MTMD) [19] :** It extracts data from multiple tasks and modalities to learn the relationship of the modality with other modalities and the task.
4. **CRNet [20]:** It exploits a multitask approach to learn multiple representation subspaces for fusion.
5. **Transformer-based Modality Binding Learning (TMBL) [21]:** A novel modality-binding network is proposed for extracting both invariant and specific modality information simultaneously.

6. Challenges and Future Scope

Deep learning has facilitated the development of multimodal sentiment analysis systems. However, multi-modal sentiment analysis continues to encounter significant challenges. This subsection examines the current state of research, problems, and future prospects in MSA.

1. **Dataset:** It is vital to multimodal sentiment analysis. A large multilingual dataset is needed. A big, diversified dataset might train a multi-modal sentiment analysis model with good generalization and wide use in different countries due to their diverse languages and races. Multimodal datasets possess inadequate annotation precision and have not attained absolute continuous values, therefore, researchers must label them more precisely. Most existing multimodal data only includes visual, voice, and text, not physiological signals like brain waves and pulses. There are few high-quality multimodal sentiment analysis data sets. Visual, audio, and text comprise most data sets, while posture and brain waves are ignored.
2. **Detection of Hidden Emotions:** The analysis of hidden emotions has always been known to be hard in multi-modal sentiment analysis tasks. Some emotions that are hidden are sarcastic emotions (like sarcastic words), emotions that need to be understood in the context they are used in, and complex emotions (like happiness and sadness). Delving into these hidden feelings is crucial. The difference between humans and computers.
3. **Modal fusion:** Modal fusion algorithms must balance the complexity of the algorithm, address the weight of different modalities in various settings, and further enhance the system's capacity for extracting the correlation between multiple pieces of information. More research is needed to determine how to allocate the weight of various modalities.
4. **Various types of video data:** Video data analysis is especially challenging in multimodal sentiment analysis. Meanwhile, the model should be insensitive to noise and suitable for low-resolution video, despite the speaker facing the camera with very high video resolution. In reality it is more complicated. The feasibility of micro-expressions or micro-gestures being elicited to generate sentiments should be studied by researchers.

Future Direction

Recent years have witnessed much interest in multimodal sentiment analysis (MSA), attributed largely to the popularity of online platforms that support text, audio, and visual content. Unlike unimodal systems, multimodal systems utilize information from multiple modalities to more effectively capture subtle emotional and affective indicators. Recent works have demonstrated the advantages of contemporary deep learning models (e.g., tensor fusion networks, transformer-based cross-modal attention, contrastive learning, etc.) in sentiment and emotion recognition performance. Nevertheless, there are still issues such as noisy modalities, incomplete observation, and real-time processing.

CONCLUSIONS

The comprehensive analysis of social media analysis for identifying sentiment in MSA highlights the growing importance of ML and DL models. The potential of these algorithms to analyze massive social media data and predict the sentiment across various modalities. This study covers the importance of modalities, the dataset and its usage, and fusion strategies, the state-of-the-art multi-modal sentiment analysis models and application. However, their effectiveness is hindered by challenges such as the lack of high-quality datasets, limited language-specific analysis, and insufficient exploration of multimodal data. To guarantee that these systems are deployed responsibly, ethical issues such as those pertaining to privacy, bias, and fairness must also be addressed. Future research should prioritize the integration of hybrid methods, the use of optimization techniques, and the incorporation of real-time detection capabilities to improve accuracy and scalability. Furthermore, investigating underrepresented regional and code-mixed languages will enhance the inclusivity and impact of these approaches.

REFERENCES

1. A. Aslam, A.B. Sargano, Z. Habib: Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks. Appl. Soft Comput. 144, 110494 (2023)
2. F. Zhang, X.C. Li, C.P. Lim, Q. Hua, C.R. Dong, J.H. Zhai: Deep emotional arousal network for multimodal sentiment analysis and emotion recognition. Inf. Fusion 88, 296–304 (2022)
3. A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain: Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Inf. Fusion 91, 424–444 (2023)
4. A.F. Dharma: A deep learning using DenseNet201 to detect masked or non-masked face (2021)

5. R. Das, T.D. Singh: Multimodal sentiment analysis: a survey of methods, trends, and challenges. ACM Comput. Surv. 55(13s), 1–38 (2023)
6. M. Kalaiyarasi, B.V.V. Siva Prasad, J.V.N. Ramesh, R.K. Kushwaha, R. Patel: Student's emotion recognition using multimodality and deep learning. ACM Trans. Asian Low-Resour. Lang. Inf. Process. (2023).
7. V. Tejaswini, K. Sathya Babu, B. Sahoo: Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 23(1), 1–20 (2024)
8. A. Perti, A. Sinha, A. Vidyarthi: Cognitive hybrid deep learning-based multimodal sentiment analysis for online product reviews. ACM Trans. Asian Low-Resour. Lang. Inf. Process. (2023)
9. S. Kusal, P. Panchal, S. Patil: Pre-trained networks and feature fusion for enhanced multimodal sentiment analysis. In: 2024 MIT Art, Design and Technology School of Computing Int. Conf. (MITADTSoCiCon), pp. 1–7. IEEE, April 2024
10. F. Wang, S. Tian, L. Yu, J. Liu, J. Wang, K. Li, Y. Wang: TEDT: transformer-based encoding–decoding translation network for multimodal sentiment analysis. Cogn. Comput. 15(1), 289–303 (2023)
11. A. Zadeh, R. Zellers, E. Pincus, L.P. Morency: MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv:1606.06259 (2016)
12. A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.P. Morency: Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proc. 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers), pp. 2236–2246 (2018)
13. S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea: MELD: A multimodal multi-party dataset for emotion recognition in conversations. arXiv:1810.02508 (2018)
14. C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, et al.: IEMOCAP: Interactive emotional dyadic motion capture database. Lang. Resour. Eval. 42(4), 335–359 (2008)
15. J. Gratch, R. Artstein, G.M. Lucas, G. Stratou, S. Scherer, A. Nazarian, et al.: The Distress Analysis Interview Corpus of human and computer interviews. In: LREC, vol. 14, pp. 3123–3128 (2014)
16. S. Amiriparian, L. Christ, A. Kathan, M. Gerczuk, N. Müller, S. Klug, et al.: The MuSe 2024 multimodal sentiment analysis challenge: social perception and humor recognition. In: Proc. 5th Multimodal Sentiment Analysis Challenge and Workshop: Social Perception and Humor, pp. 1–9 (2024)
17. W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.P. Morency, S. Poria: Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In: Proc. 2021 Int. Conf. on Multimodal Interaction, pp. 6–15 (2021)
18. D. Wang, X. Guo, Y. Tian, J. Liu, L. He, X. Luo: TETFN: a text-enhanced transformer fusion network for multimodal sentiment analysis. Pattern Recognit. 136, 109259 (2023)
19. R. Lin, H. Hu: Multi-task momentum distillation for multimodal sentiment analysis. IEEE Trans. Affect. Comput. (2023)
20. Z. Zhao, J. Huang, D. Zhou, D. Xu, J. Cao: Co-space representation interaction network for multimodal sentiment analysis. Knowl.-Based Syst. 283, 111149 (2024)
21. J. Huang, J. Zhou, Z. Tang, J. Lin, C.Y.-C. Chen: TMBL: Transformer-based multimodal binding learning model for multimodal sentiment analysis. Knowl.-Based Syst. 285, 111346 (2024)