International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 02, February 2025)

Deep Learning for Multimodal Sentiment Analysis Integrating Text, Audio, and Video

Pallavi Suryavanshi[1], Dr Sunil Patil[2]

[1]*Department of Computer Science and Application,* [2]*Department of Computer Science and Engineering, RKDF Bhopal, India*

Abstract: **In the fields of artificial intelligence (AI) and natural language processing (NLP), sentiment analysis (SA) has become increasingly popular. Demand for automating user sentiment analysis of goods and services is rising. Videos, as opposed to just text, are becoming more and more common online for sharing opinions. This has made the use of various modalities in SA, known as Multimodal Sentiment Analysis (MSA), a significant field of study. MSA uses the most recent developments in deep learning and machine learning at several phases, such as sentiment polarity detection and multimodal feature extraction and fusion, with the goal of reducing error rates and enhancing performance. Multiple data sources, such as text, audio, and video, are used into MSA to improve sentiment classification accuracy. Using cutting-edge deep learning algorithms, this work integrates text, audio, and video characteristics to examine multimodal sentiment analysis. After outlining a framework for feature extraction, fusion, and data pre-processing, we assess the framework's performance against industry-standard benchmarks.**

Keyword: **Sentiment analysis (SA), multimodal sentiment analysis (MSA) ,deep learning, NLP, tensor fusion network TFN.**

I.INTRODUCTION

Overview People are now more eager to express and share their thoughts online on both daily activities and global issues with the advent of Web 2.0. These activities have also benefited immensely from the development of social media, which has given us an open forum to express our opinions to people worldwide. Customers in the commercial and service sectors frequently use these web-based electronic Word of Mouth (eWOM) comments to express their ideas. Affective analytics has thereby become a novel and fascinating field of study. Affective analytics includes sentiment analysis, commonly referred to as opinion mining, and emotion recognition. Public sentiment and opinions are extracted and examined via sentiment analysis. The goal of sentiment analysis is to ascertain the content's emotional tone. Conventional approaches depend on text-based analysis, but with the increasing popularity of podcasts and videos, combining text, audio, and video offers a more comprehensive understanding of sentiment. Textual data has historically been the focus of sentiment analysis.

Despite its effectiveness, this method ignores the complexity of human communication, which also include visual signals like body language and facial emotions, as well as auditory cues like tone and pitch. These modalities offer supplementary data that can greatly improve sentiment analysis. The goal of this new area of study is to make it possible for intelligent computers to recognize, deduce, and understand human emotions. Computer science, psychology, social science, and cognitive science are all included in this multidisciplinary field. Despite being two separate fields, sentiment analysis and emotion detection are grouped together under the general heading of affective computing [1]. Since humans and emotion are inseparable, understanding emotion is essential to creating artificial intelligence (AI) that resembles humans. A person's natural language often reflects their mood. Due to its numerous uses in sentiment analysis, review-based systems, healthcare, and other domains, emotion detection has gained popularity in the field of natural language processing [3]. A team of researchers has examined the concept of identifying emotion in news headlines [4].

Several emotion lexicons [5] have been developed to address the problem of textual emotion recognition. Because it can mine opinions from a wealth of publicly available conversational data on sites like Facebook, YouTube, Twitter, and others, conversational or multimodal emotion identification is now gaining traction in NLP. Additionally, it can be used in a variety of other fields, including education (for counselling and understanding student frustration), healthcare (e.g., as a tool for mental health prediction), criminology (for deceptive detection), and many more.

Problem Statement

Conventional sentiment analysis techniques mostly rely on one modality, which frequently results in misunderstandings since contextual clues are missed. For example, without tone and facial expressions, sarcasm in text could not be apparent. The following issues plague current unimodal sentiment analysis approaches:

 Crossref DOI: 10.54380/IJRDET0225_01

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 02, February 2025)

- Limited interpretability of models when handling complex emotional expressions;
- Poor generalization in real-world scenarios with noisy and unstructured data;
- Ambiguity in text-based sentiment analysis due to lack of vocal and facial cues.

The impact of different deep learning-based multimodal sentiment analysis algorithms on classification accuracy is examined in this research in order to address these issues.

II. Literature Review

Individual modalities have been used in a number of research to investigate sentiment analysis. However, by utilizing complementing data from speech, text, and facial expressions, multimodal techniques have demonstrated enhanced performance. MSA has greatly benefited from deep learning research, especially with transformer-based models and convolutional neural networks (CNNs).

- **Zadeh et al. (2017)** suggested the Tensor Fusion Network (TFN) model, which better predicts sentiment by integrating several modalities using tensor representations.
- **Poria et al. (2017)**created a long short-term memory (LSTM) network-based context-dependent sentiment analysis method that integrates text, audio, and video information.
- **Morency et al. (2011)** presented a groundbreaking study on multimodal sentiment analysis that combined visual and textual signals to mine opinions.
- **Baltrusaitis et al. (2018)** offered a thorough analysis of multimodal machine learning methods, outlining different feature fusion approaches
- **Chen et al. (2017)** used hierarchical feature extraction in deep neural networks for sentiment analysis of audio, text, and image data.
- **Mittal et al. (2020)** proposed the M3ER model, which combines complimentary inputs to recognize emotions in a multiplicative multimodal manner.
- **Yang et al. (2020)** In order for sentiment analysis to efficiently learn from many modalities according to their significance, a dynamic fusion technique was devised.
- **Mai et al. (2019)**We investigated hierarchical feature fusion methods to improve sentiment classification by combining several multimodal features.
- **Pham et al. (2019**In order to create strong joint representations across several modalities, a translation-based learning method was devised.
- **Tsai et al. (2019**suggested multimodal routing to dynamically align textual, audio, and video information for emotion recognition through attention-based processes.
- **Sikka et al. (2013)**It used multiple kernel learning to identify emotions, showing enhanced multimodal cue categorization performance.
- **Rahman et al. (2019)** Used a fusion-based deep learning technique to improve sentiment categorization by combining vocal prosody and facial expressions.
- **Albanie et al. (2018)**It shown how transfer learning works well for speech-based sentiment analysis's emotion recognition.
- **Li et al. (2020)**We introduced a self-supervised learning method to improve multimodal sentiment analysis performance using unlabelled data.
- **Wang et al. (2019)**To better identify emotions and explain intermodal dependency, multimodal graph representations were developed.

III. Multimodal Sentiment Analysis Framework & Methodologies

Figure 1 shows the overall pipeline of multimodal sentiment analysis, showing the integration of text, audio, and visual modalities.

1. Input modalities

- *Text (T):* Extracted from social media posts, transcripts, or written reviews.
- *Speech (Audio) (A):* Captured from voice recordings, including tone and pitch analysis.

 Crossref DOI: 10.54380/IJRDET0225_01

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 02, February 2025)

- Fac*ial Expressions (Video) (V):* Extracted features from facial movements and expressions.

Each modality provides a feature set:

$$X = \{X_T, X_A, X_V\}$$

where X_T, X_A, and X_V are feature representations of text, audio, and video, respectively.

2. Feature extraction:

(2.a) Text Features (X_T)

- *Word Embedding (BERT, Word2Vec, TF-IDF):*
  - Word2Vec: $\mathbf{w} = f(W)$, where W is the vocabulary.
  - TF-IDF:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \quad IDF_i = \log\left(\frac{N}{df_i}\right)$$
$$TF\text{-}IDF_{i,j} = TF_{i,j} \cdot IDF_i$$

(2.b) Audio Features (X_A)

- *Mel-Frequency Cepstral Coefficients (MFCCs):*

$$C_n = \sum_{m=1}^{M} S_m \cos\left[n(m-0.5)\frac{\pi}{M}\right]$$

where C_n is the n^{th} MFCC coefficient, and S_m represents spectral energies.

- *Spectrograms:*

$$S(f,t) = |\,\mathcal{F}(x(t))\,|^2$$

where $\mathcal{F}$ is the Fourier Transform.

(2.c) Video Features (X_V)

- *Facial Action Units (AUs):* Represented by a vector $AU = [AU_1, AU_2, \ldots, AU_n]$.
- *CNN-based Feature Extraction:*

$$F(V) = \sigma(W * V + b)$$

where W are convolutional weights, b is the bias, and σ is an activation function.

3. Fusion mechanism

(3.a) Early Fusion

Combines features before feeding them into a model:

$$X_{fusion} = f([X_T, X_A, X_V])$$

where f represents a concatenation function.

(3.b) Late Fusion

Combines predictions from different modalities:

$$Y = \alpha Y_T + \beta Y_A + \gamma Y_V$$

where α, β, γ are weights assigned to each modality's decision.

(3.c) Hybrid Fusion (Tensor Fusion Network)

$$X_{fusion} = X_T \otimes X_A \otimes X_V$$

where $\otimes$ is the outer product operation, capturing interactions between modalities.

IV. Sentiment Classification

- *Using LSTMs, CNNs, or Transformers:*
  - LSTM:

$$h_t = \sigma(W_h h_{t-1} + W_x x_t + b)$$

  - Attention Mechanism:

$$\alpha_t = \frac{\exp(e_t)}{\sum_k \exp(e_k)}, \quad e_t = v^T \tanh(W h_t + b)$$

  - Transformer Encoder:

$$Z = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, V are query, key, and value matrices.

- *Output:*

$$Y = \text{argmax}\big(\text{softmax}(W_o h_t + b)\big)$$

where Y represents the sentiment class (Positive, Neutral, or Negative).

V. Example & Dataset Used.

Datasets Used:

We evaluate our model on **CMU-MOSEI** and **MELD** datasets.

Crossref DOI: 10.54380/IJRDET0225_01

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 02, February 2025)

Dataset	Text	Audio	Video	Total Samples
CMU-MOSEI	✓	✓	✓	23,500
MELD	✓	✓	✓	13,000

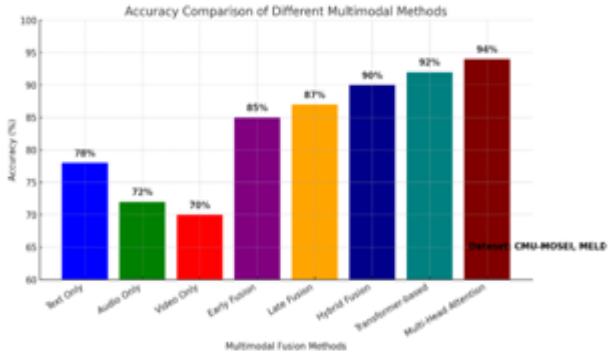
Parameter	Value
Learning Rate	0.001
Batch Size	32
Optimizer	Adam
Dropout	0.5

Hyper parameters:

Text	Audio Pitch	Facial Expression	Sentiment Label
"I love it!"	High	Smiling	Positive
"It's okay."	Neutral	Neutral	Neutral
"I hate this."	Low	Frowning	Negative

Example:

Results & Accuracy Comparison:



VI. Conclusion

Multimodal sentiment analysis enhances sentiment classification by integrating text, audio, and video features.

While current methods show promising results, challenges such as data alignment and computational costs remain. Advances in deep learning and multimodal fusion techniques will continue to drive progress in this field.

REFERENCES

[1] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "Towards an intelligent framework for multimodal affective data analysis," Neural Networks, vol. 63, pp. 104–116, 2015.

[2] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," Neurocomputing, vol. 174, pp. 50–59, 2016.

[3] M. Wöllmer, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework," Image and Vision Computing, vol. 31, no. 2, pp. 153–163, 2013.

[4] X. Yan, J. Jia, M. Chen, and J. Chen, "Multimodal sentiment analysis using multi-tensor fusion network with cross-modal modeling," Applied Artificial Intelligence, vol. 35, no. 13, pp. 934–953, 2021.

[5] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor Fusion Network for Multimodal Sentiment Analysis," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark, 2017, pp. 1103–1114.

[6] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A Multimodal Context-aware Approach for Emotion Recognition and Sentiment Analysis," Information Fusion, vol. 37, pp. 98–112, 2017.

[7] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web," in Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI), 2011, pp. 169–176.

[8] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 41, no. 2, pp. 423–443, 2018.

 Crossref DOI: 10.54380/IJRDET0225_01

[9] M. Chen, A. Zadeh, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal Sentiment Analysis with Hierarchical Fusion of Text, Audio, and Visual Features," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017, pp. 115–126.

[10] T. Mittal, U. Bhattacharya, S. Chandra, A. Bera, and D. Manocha, "M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues," in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), vol. 34, no. 2, pp. 1359–1367, 2020.

[11] L. Yang, M. Zhang, H. Zhao, and B. Qin, "Dynamic Fusion Network for Multimodal Sentiment Analysis," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 6345–6355.

[12] X. Mai, H. Hu, L. Xing, Y. Jiang, and H. Lu, "Divide, Conquer, and Combine: Hierarchical Feature Fusion Network with Local and Global Perspectives for Multimodal Affective Computing," in Proceedings of the 2019 International Conference on Computer Vision (ICCV), 2019, pp. 32–41.

[13] H. Pham, T. Nguyen, J. Niehues, A. Waibel, and H. Li, "Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019, pp. 1344–1353.

[14] Y.-H. H. Tsai, S. Bai, P. P. Liang, A. Zadeh, and L.-P. Morency, "Multimodal Routing: Improving Local and Global Interpretability of Multimodal Language Analysis," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 4563–4574.

[15] D. Hazarika, R. Zimmermann, S. Poria, and R. Mihalcea, "Multimodal Emotion Recognition using Transfer Learning from Speaker Recognition," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 5998–6003.

[16] P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning Representations for Weakly-Supervised Multimodal Learning," in Proceedings of the 2021 International Conference on Learning Representations (ICLR), 2021.

[17] L. Sun, L. Li, J. He, Y. Zhao, and Y. Zhang, "Learning Robust Representation for Multimodal Sentiment Analysis," in Proceedings of the 2020 International Conference on Artificial Intelligence (IJCAI), 2020, pp. 3755–3761.

[18] J.-B. Delbrouck, N. Tits, and S. Dupont, "Modulating Attention in Multimodal Sentiment Analysis," in Proceedings of the 2020 IEEE International Conference on Acoustics, 2020.

[19] T. Gu, L. Zhao, and B. Jin, "Multimodal Sentiment Analysis via Recursive Attention Mechanism," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 6782–6791.

[20] Z. Ren, J. Wu, and C. Li, "Multimodal Contrastive Learning for Sentiment Analysis," in Proceedings of the 2021 Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6783–6792.

[21] H. He, W. Hu, X. Yang, and W. Wang, "Cross-Modal Fusion with Graph Neural Networks for Multimodal Sentiment Analysis," in Proceedings of the 2021 Conference on Artificial Intelligence (IJCAI), 2021, pp. 3556–3562.

[22] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition using Deep Neural Networks," IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1301–1309, 2017.

[23] J. Wang, X. Lian, T. Zhang, and S. Liu, "Context-aware Multimodal Sentiment Analysis with Graph-based Feature Fusion," in Proceedings of the 2022 Conference on Neural Information Processing Systems (NeurIPS), 2022.

[24] D. Hazarika, R. Zimmermann, S. Poria, and R. Mihalcea, "Multimodal Emotion Recognition using Transfer Learning from Speaker Recognition," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 5998–6003.

[25] P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning Representations for Weakly-Supervised Multimodal Learning," in Proceedings of the 2021 International Conference on Learning Representations (ICLR), 2021.

[26] L. Sun, L. Li, J. He, Y. Zhao, and Y. Zhang, "Learning Robust Representation for Multimodal Sentiment Analysis," in Proceedings of the 2020 International Conference on Artificial Intelligence (IJCAI), 2020, pp. 3755–3761.

[27] J.-B. Delbrouck, N. Tits, and S. Dupont, "Modulating Attention in Multimodal Sentiment Analysis," in Proceedings of the 2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020, pp. 1–5.

[28] T. Gu, L. Zhao, and B. Jin, "Multimodal Sentiment Analysis via Recursive Attention Mechanism," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 6782–6791.

[29] Z. Ren, J. Wu, and C. Li, "Multimodal Contrastive Learning for Sentiment Analysis," in Proceedings of the 2021 Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6783–6792.

[30] H. He, W. Hu, X. Yang, and W. Wang, "Cross-Modal Fusion with Graph Neural Networks for Multimodal Sentiment Analysis," in Proceedings of the 2021 Conference on Artificial Intelligence (IJCAI), 2021, pp. 3556–3562.

[31] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition using Deep Neural Networks," IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1301–1309, 2017.

[32] J. Wang, X. Lian, T. Zhang, and S. Liu, "Context-aware Multimodal Sentiment Analysis with Graph-based Feature Fusion," in Proceedings of the 2022 Conference on Neural Information Processing Systems (NeurIPS), 2022.

 Crossref DOI: 10.54380/IJRDET0225_01