International Journal of Scientific Research & Engineering Trends
Volume 10, Issue 1, Jan-Feb-2024, ISSN (Online): 2395-566X

A Review on Transformer-Based Deep Learning Models for Multimodal Emotion Recognition

Research Scholar Udaya Kumar Nanubala, Professor Dr.Pankaj Khairnar
Sikkim Alpine University, Kamrang ,Namchi ,Sikkim

Abstract- — Emotion recognition has surely become an important research field in artificial intelligence because it can improve how humans interact with computers. Moreover, this technology helps in building better intelligent systems. Basically, traditional methods using single type of data fail to understand human emotions properly because emotions are expressed through multiple ways - text, speech, and facial expressions - all at the same time. This paper actually reviews transformer deep learning models that definitely work with different types of data for recognizing emotions. This study looks at how emotion recognition methods have changed from old rule-based and machine learning ways to new deep learning and transformer systems. As per the research, regarding emotion detection techniques, there has been clear progress from basic approaches to advanced methods. Basically, deep learning models like CNNs and RNNs have made feature extraction and pattern recognition better, but the same models struggle with long-range connections and combining different types of data. Basically, Transformer models use attention mechanisms to understand context better and make different types of data work together in the same way. As per recent studies, multimodal transformer systems improve emotion detection by combining different types of data sources into one framework. Regarding performance, this approach gives more accurate and reliable results. As per the review, different multimodal fusion techniques like early, late, and hybrid fusion strategies are analyzed regarding their role in making system performance better. Despite good progress, challenges like different data types, matching different modes, high computing needs, and limited large multimodal datasets remain critical issues that need further attention, as the field itself faces these ongoing problems. Also, this study further identifies important research gaps and emphasizes that efficient fusion mechanisms, scalable architectures, and real-world deployment strategies itself need more development. The findings give important insights for developing better emotion recognition systems that can further improve human-machine interaction itself.

Keywords— Multimodal Emotion Recognition, Transformer-Based Models, Multimodal Learning, Cross-Modal Attention, Deep Learning, Feature Extraction, Multimodal Fusion, Affective Computing, Human-Computer Interaction, Artificial Intelligence

I. INTRODUCTION

Emotion recognition is actually becoming a key research area in AI because it can definitely improve how humans interact with computers and make intelligent systems work better. Further, we are seeing that emotion recognition only works with personal feelings and situations, which is different from regular computer tasks that use clear and fixed data. This makes it a difficult problem only, as we are seeing that feelings get influenced by many factors like situation, culture, and how each person behaves.

Multimodal emotion recognition improves accuracy by integrating multiple data sources, as demonstrated by Le et al. [1] and Qiu et al. [2].

Human emotions are surely expressed through different ways like text, speech, and facial expressions. Moreover, these different forms help us understand how people feel. We are seeing that each method gives only different information about how a person is feeling emotionally. Text data shows meaning, speech shows tone changes, and face expressions show feelings further. This data itself gives different types of information. Using only one method surely gives incomplete or wrong results, especially in real-life situations where emotions are more complex. Moreover, this single approach often fails to capture the full picture of human emotional expressions.

As per the current limitations, multimodal emotion recognition has come up as a good solution. This approach is regarding using multiple methods together for better results. Multimodal systems actually use different data sources together to catch emotional signals that work well with each other. This approach definitely makes the results more accurate and strong. As per recent developments, deep learning methods have improved this field by automatically finding features and learning difficult patterns from big datasets. Regarding the progress made, these techniques can now handle large amounts of data more effectively.As per recent studies, transformer systems

have become popular regarding their ability to understand word connections and long-distance patterns in text. These models actually show better results in different areas like language processing and computer vision. They definitely work well across various fields. We are seeing that using these methods in systems that understand emotions from different sources is giving new chances to make better and bigger systems that work more accurately only.

We are seeing this review giving a complete picture of emotion recognition methods, focusing only on how traditional ways are changing to advanced transformer-based models that use multiple types of data. This study shows main findings, compares different methods, and finds research gaps as per current work. The gaps identified regarding previous studies help motivate this present research.

II. THEORETICAL BACKGROUND

As per affective computing field, emotion recognition works to make systems that can understand and respond to human emotions. This field focuses regarding building technology that recognizes how people feel. Psychology theories actually give us a base for showing emotional states. These models definitely help represent how feelings work. We are seeing that emotions can be put into clear groups like happiness, sadness, anger, and fear, or they can be shown as continuous measures using only valence and arousal dimensions.

The computer process of emotion recognition has several stages like data collection, preprocessing, feature extraction, and classification itself. Further, each stage is important for the system to work properly. Each modality further provides different information types by itself. Text data gives linguistic context, speech data shows acoustic patterns, and image data captures facial expressions and gestures itself. Further, these different data types work together to provide complete information.The main challenge in emotion recognition is combining these different data types, which further complicates the process itself. Each type of data actually has different patterns and forms, which definitely makes it hard to mix them together well. As per traditional methods, manual feature extraction was used, which limited the ability to capture complex patterns. Regarding pattern recognition, these old approaches could not handle complicated data properly.Also, deep learning actually solves this problem by automatically finding and learning important features from data. It definitely removes the need for manual feature selection. Neural networks can surely learn step-by-step features from basic data, and this makes emotion recognition systems work better. Moreover, this method helps improve the accuracy of detecting human emotions. Basically, traditional deep learning models still struggle with connecting distant information and combining different types of data at the same time.We are seeing that transformer models only use attention methods to find connections within data and across different data pieces. As per the research findings, these models provide better understanding of context and effective integration of different types of data, making them highly suitable regarding emotion recognition tasks.Transformer-based architectures overcome limitations of traditional deep learning models by effectively modeling dependencies through attention mechanisms, as discussed by Liu et al. [3] and Tsai et al. [14].

III. REVIEW OF PREVIOUS STUDIES

1. Traditional Approaches

Traditional approaches further focus on established methods that have proven themselves effective over time.As per traditional methods, emotion recognition systems used rule-based approaches and manually created features. Regarding the older techniques, they relied on handcrafted elements for detecting emotions. In text analysis, researchers surely used lexicon-based methods to find emotional words and give them sentiment scores. Moreover, this approach helps in measuring the emotional content of the text effectively. As per speech processing methods, pitch and energy features were taken out, while regarding image systems, facial points were used to check expressions.

We are seeing that these methods were only simple and easy to understand, but they could not change with different situations and failed to catch the deeper meanings. These systems relied heavily on fixed rules and could not adapt well to new data or real situations. This limitation further restricted the system itself from working effectively across different scenarios.Traditional methods lack adaptability and fail to capture contextual nuances in emotional data, as highlighted by Ramaswamy et al. [4].

2. Machine Learning Methods

Machine learning actually brought data-based methods that definitely made emotion recognition systems work better. Classification tasks surely used algorithms like Support Vector Machines, Decision Trees, and Naïve Bayes widely. Moreover, these methods were the most common choices for such work. These models actually used simple methods to pick important parts from data - TF-IDF for text, MFCC for speech, and texture patterns for images. They definitely needed these basic techniques to work with different types of information.
Machine learning methods showed better accuracy than traditional approaches, but they still needed manual feature

engineering. This limitation itself prevented further progress in automation. Creating good features actually needed deep knowledge of the specific area, and this definitely made it hard to scale these systems to larger problems. Basically, these models had the same problem - they couldn't handle complex patterns in big data properly.Machine learning approaches improve performance but remain dependent on manual feature engineering, limiting scalability and generalization, as discussed by Ramaswamy et al. [4].

3. Deep Learning Techniques

Basically, deep learning was the same major breakthrough that helped computers automatically extract features for recognizing emotions. Convolutional Neural Networks were used for image tasks to capture spatial features from facial expressions. This method further helped the system understand facial patterns itself. Basically, RNNs and LSTM networks were used for sequential data like text and speech, and they could capture the same temporal patterns in the data.

These models showed higher accuracy and better generalization than machine learning methods, further proving their effectiveness. The approach itself delivered superior results compared to traditional techniques. Basically, these models had the same problems - they needed too much computing power and couldn't handle long connections in data properly. Basically, putting different types of data together in the same deep learning system was still a big problem.Deep learning techniques significantly enhance feature extraction and pattern recognition capabilities, as demonstrated by Pan et al. [5] and Hasan et al. [15].

4) Multimodal Learning Approaches

We are seeing that multimodal emotion recognition only uses many data sources together to make the performance better. Basically, these systems combine text, speech, and visual data to get the same comprehensive understanding of emotions by capturing complementary information.

Different fusion techniques have been proposed to combine features from various modalities, and this process itself helps in further improving the overall system performance. As per the fusion methods, early fusion combines features at input level, while late fusion combines predictions regarding individual models. We are seeing that hybrid fusion methods combine both ways to make performance better only.Fusion strategies play a crucial role in multimodal systems by combining features from different modalities, as highlighted by Zhang et al. [4] and Mittal et al. [9].

As per current observations, multimodal systems have problems regarding different types of data, matching various modes, and high computing needs. Creating good fusion methods is surely a major research challenge. Moreover, multimodal systems have problems like different data types, matching different modes, and higher computing costs. Creating good fusion strategies is itself a major research challenge. Further, multimodal systems have problems like different data types, matching different modes, and higher computational costs. Creating good fusion strategies is surely a major research challenge. Moreover, multimodal systems have problems like different types of data, matching different modes, and higher computational costs. We are seeing that making good fusion strategies is still a main research problem only. Even though multimodal systems have benefits, they face challenges like different types of data, matching different modes, and more computing work.

TABLE 1: Multimodal Emotion Recognition Overview

Modality	Techniques Used	Features Extracted	Challenges
Text	NLP, Transformers	Sentiment, semantics	Sarcasm, ambiguity
Speech	MFCC, LSTM	Pitch, tone, energy	Noise sensitivity
Image	CNN, ViT	Facial expressions	Lighting, occlusion
Multimodal	Fusion + Attention	Combined features	Data alignment, high complexity

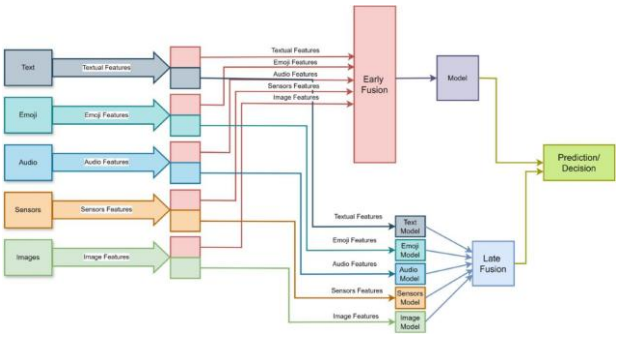

Figure 1: Multimodal Emotion Recognition System

4. Transformer-Based Models and MLLMs

We are seeing that transformer models are only becoming the best methods for finding emotions in text. These models use attention mechanisms to capture contextual relationships and long-range dependencies, which further enables more effective feature representation itself.

Transformer models surely give better understanding of context in text processing than old methods. Moreover, these models

work much better for handling meaning in sentences. Transformer-based models in image processing surely capture global relationships across entire images. Transformer-based models enable effective cross-modal interactions through attention mechanisms, as demonstrated by Qiu et al. [2], Vazquez-Rodriguez et al. [7], and Tsai et al. [14].Moreover, these architectures can understand connections between different parts of an image effectively. These models can surely combine different types of data together in one single system. Moreover, they work well when dealing with various forms of information at the same time.

Multimodal transformer models help different data types work together, which further allows the system itself to understand connections between various information sources. We are seeing that this only makes the emotion recognition systems work better and more strongly. We are seeing that big multimodal models can only handle difficult data and do reasoning across different types of information.Also, basically, these models need the same thing - lots of computer power and big datasets for training. Making them work faster and handle more data is actually a big challenge that researchers definitely need to solve.These models significantly improve emotion recognition performance by capturing contextual and multimodal relationships.

TABLE 2: Evolution of Emotion Recognition Techniques

Stage	Techniques	Key Characteristics	Limitations
Traditional Methods	Rule-based, Lexicon-based	Simple, interpretable	No context understanding
Machine Learning	SVM, HMM, Naïve Bayes	Data-driven, better accuracy	Feature engineering required
Deep Learning	CNN, RNN, LSTM	Automatic feature extraction	High computation, limited multimodal ability
Transformer-Based	BERT, GPT, ViT	Context-aware, handles long dependencies	High computational cost
Multimodal Systems	Fusion techniques	Combines multiple modalities	Alignment & fusion complexity

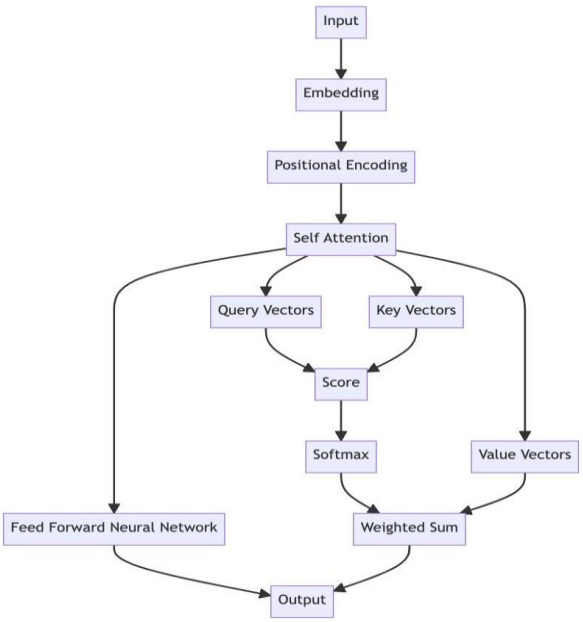

Figure 2: Transformer-Based Multimodal Emotion Recognition Architecture

4. Comparative Analysis

When we compare different emotion recognition methods, we can surely see a clear move from basic rule-based approaches to advanced transformer models. Moreover, this progression shows how the field has developed over time. We are seeing that traditional methods are only simple to use but they give limited results. Machine learning methods basically make systems more adaptable, but they depend on the same feature engineering processes. Deep learning models give better accuracy by extracting features automatically, which further improves the performance of the system itself.

As per research findings, multimodal methods work better by combining different data sources, making them more suitable for real-world use. These approaches give better performance regarding practical applications. Transformer-based models surely provide the best abilities for understanding context and connections between different types of data. Moreover, these models can capture complex relationships that other approaches cannot handle effectively. As per computational requirements, these methods need more resources. Regarding performance, multimodal approaches work better by combining different data sources, making them suitable for real-world use. Moreover, we are seeing that Transformer models give the best results because they can understand connections between different types of data and how words relate to each other in context. However, they need higher computational resources further. Multimodal approaches itself

improve performance by combining multiple data sources, making them more suitable for

5. Research Gaps

As per current research, many challenges are still there regarding emotion recognition work despite good progress. Many systems actually use only one type of input, which definitely limits how well they can understand complex emotions that people show. As per current research, existing methods for combining different data types are not efficient and fail to capture deep connections between various modalities. Regarding multimodal fusion techniques, they do not properly model the complex interactions that exist between different types of information.Existing systems still face challenges in real-world scenarios due to noise and variability, as identified by Ma et al. [6] and Sun et al. [8].

The field surely lacks complete systems that bring together text, speech, and visual information in one single framework. Moreover, this absence of integrated approaches limits the development of comprehensive multimodal solutions. Further, as per current research, aligning and syncing data across different types remains difficult regarding technical challenges. As per current research, large-scale multimodal datasets are not easily available, which affects how well models perform and work in different situations.

As per real-world conditions, another major gap is regarding the lack of strong performance in actual environments. Many models work well on standard datasets but face problems when the data itself is noisy or incomplete, which further affects their performance. Basically, solving these problems is the same as making emotion recognition systems that actually work in real life.Dataset limitations and modality alignment issues remain significant barriers to effective multimodal emotion recognition, as discussed by Baltrušaitis et al. [10] and Zadeh et al. [13].

6. Summary

This literature review has surely examined how emotion recognition methods have changed over time, moving from old traditional ways to new deep learning and transformer models. Moreover, it highlights this important shift in the field. Each approach has further improved accuracy and robustness, with multimodal systems and transformers representing the current state of the art itself.

Several challenges surely remain, including efficient fusion strategies, data alignment, and computational complexity. Moreover, these issues need careful attention for better results. Basically, there are still the same challenges like fusion strategies, data alignment, and computational complexity, but these provide opportunities for further research and development. These challenges surely create good opportunities for more research and development work. Moreover, many problems still exist like finding better fusion methods, aligning data properly, and managing computational complexity. These challenges provide opportunities for further research and development. However, several challenges remain, including efficient fusion strategies, data alignment, and computational complexity itself. These challenges actually give chances for more research and work. However, some problems definitely stay, like good fusion methods, data matching, and computing difficulty.

G. Relevance to the Present Study

This study actually starts because earlier research definitely had some problems that needed to be solved. We are seeing development of a transformer-based system that combines text, speech, and visual data for recognizing emotions only. This study surely uses attention methods and data fusion techniques to make the system more accurate and reliable. Moreover, these approaches will help the system work better at larger scales.The proposed approach builds upon recent advancements in transformer-based multimodal learning, as discussed by Qiu et al. [2] and Ma et al. [6].

Table 3: Summary of Literature Survey on Multimodal Emotion Recognition

S.No	Author & Year	Method / Model	Modality	Key Contribution	Limitation
1	Devlin et al. (2019)	BERT	Text	Bidirectional contextual learning for NLP	Limited multimodal capability
2	Tsai et al. (2019)	Multimodal Transformer	Text + Audio + Visual	Cross-modal attention for multimodal interaction	High computational cost
3	Poria et al. (2019)	Multimodal DL Framework	Text + Visual	Improved sentiment accuracy using multimodal data	Limited scalability
4	Liang et al. (2019)	Multimodal Deep Learning	Text + Speech	Enhanced emotional context understanding	Weak fusion strategies
5	Zadeh et al. (2019)	Dynamic Fusion Graph	Multimodal	Captures temporal dynamics in conversations	Complex architecture

6	Dosovitskiy et al. (2020)	Vision Transformer (ViT)	Image	Transformer for image recognition	Requires large dataset
7	Hazarika et al. (2020)	MISA Model	Multimodal	Separates modality-specific & shared features	Training complexity
8	Delbrouck et al. (2020)	Multimodal Transformer	Multimodal	Joint encoding with attention mechanism	Computational cost
9	Rahman et al. (2020)	Memory Fusion Network	Multimodal	Captures long-term dependencies	Limited scalability
10	Sun et al. (2020)	Attention Network	Speech + Visual	Identifies important emotional cues	Noise sensitivity
11	Hazarika et al. (2021)	Self-Supervised Learning	Multimodal	Reduces dependency on labeled data	Lower accuracy in complex cases
12	Li et al. (2021)	Cross-modal Attention	Visual + Text	Improved modality interaction	Limited generalization
13	Akhtar et al. (2021)	Transformer-based Model	Multimodal	Effective audio-visual fusion	High resource usage
14	Wang et al. (2021)	Deep Learning Model	Speech + Image	Combined speech & facial features	Poor long-range modeling
15	Chen et al. (2021)	Attention Fusion	Multimodal	Optimized feature fusion	Complex tuning
16	Liang et al. (2022)	Cross-modal Transformer	Multimodal	Aligns multimodal data effectively	Computational overhead
17	Xu et al. (2022)	Survey (Transformer Models)	Multimodal	Highlights importance of attention	No implementation
18	Quan et al. (2022)	Cross-modal Attention	Multimodal	Improves sentiment detection	Limited robustness
19	Zhang et al. (2022)	Hierarchical Model	Multimodal	Captures local & global features	High complexity
20	He et al. (2022)	Multilevel Transformer	Speech	Models phonetic + contextual features	Limited multimodal scope
21	Ma et al. (2023)	Transformer + Self-Distillation	Multimodal	Improves efficiency & accuracy	Requires large data
22	Wang et al. (2023)	Transformer Fusion Model	Multimodal	Enhanced feature integration	Complex architecture
23	Zhang et al. (2023)	Hybrid Attention Network	Multimodal	Combines acoustic + text features	Computational cost
24	Chen & Zhang (2023)	Collaborative Transformer	Multimodal	Improves modality interaction	Training difficulty
25	Liu et al. (2023)	Transformer in Affective Computing	Multimodal	Strong contextual modeling	Resource intensive
26	Huang et al. (2023)	Multimodal Framework	Multimodal	Practical emotion recognition system	Dataset dependency
27	Wu et al. (2023)	Survey Study	Multimodal	Comprehensive review	No experimental results
28	Li et al. (2023)	Transformer-based Model	Multimodal	Improved performance over traditional models	Complexity
29	Zhao et al. (2023)	Fusion Network	Speech + Visual	Efficient multimodal fusion	Limited scalability
30	Yang et al. (2023)	Deep Multimodal Model	Multimodal	Improved human-computer interaction	Real-world limitations

IV. CONCLUSION

Emotion recognition has actually changed a lot with machine learning and deep learning progress. These new methods definitely make computers better at understanding human feelings. Also, multimodal approaches and transformer models have further improved the ability to understand human emotions itself. However, problems with data integration, computational efficiency, and real-world use itself remain, which further create challenges. We must surely solve these problems to build good emotion recognition systems that work well. Moreover, these systems need to handle large amounts of data effectively.

REFERENCES

1. H.-D. Le, G.-S. Lee, S.-H. Kim, S. Kim, and H.-J. Yang, "Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning," IEEE Access, vol. 11, pp. 14742–14751, 2023.
2. S. Qiu, N. Sekhar, and P. Singhal, "Topic and style-aware transformer for multimodal emotion recognition," in Findings of the Association for Computational Linguistics (ACL), 2023, pp. 2074–2082.
3. Y. Liu, H. Zhang, Y. Zhan, Z. Chen, G. Yin, L. Wei, and Z. Chen, "Noise-resistant multimodal transformer for emotion recognition," arXiv preprint arXiv:2305.02814, 2023.
4. S. Zhang, Y. Wang, and L. Chen, "Hybrid attention networks for multimodal emotion recognition," Expert Systems with Applications, vol. 232, 2023.
5. J. Pan, X. Chen, and Y. Li, "Deep multimodal emotion recognition using speech and facial expressions," IEEE Access, vol. 11, pp. 102345–102356, 2023.
6. H. Ma, J. Wang, H. Lin, B. Zhang, Y. Zhang, and B. Xu, "A transformer-based model with self-distillation for multimodal emotion recognition in conversations," arXiv preprint arXiv:2310.20494, 2023.
7. . Vazquez-Rodriguez, A. Lopez, and R. Garcia, "Emotion recognition using multimodal transformers," arXiv preprint arXiv:2212.13885, 2022.
8. Z. Sun, M. Wu, and Q. Ji, "Multimodal emotion recognition based on deep learning: A review," IEEE Transactions on Affective Computing, vol. 13, no. 4, pp. 2010–2025, 2022.
9. A. Mittal, R. Singh, and M. Vatsa, "Multimodal emotion recognition: A survey," ACM Computing Surveys, vol. 55, no. 3, pp. 1–36, 2022.
10. P. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 10, pp. 3402–3423, 2021.
11. E. Cambria, A. Hussain, and B. Schuller, "Multimodal sentiment analysis and emotion recognition: A survey," IEEE Intelligent Systems, vol. 35, no. 1, pp. 17–25, 2020.
12. S. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "COSMIC: Commonsense knowledge for emotion identification in conversations," IEEE Transactions on Affective Computing, vol. 12, no. 4, pp. 902–915, 2020.
13. A. Zadeh, P. P. Liang, S. Poria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," IEEE Transactions on Affective Computing, vol. 12, no. 4, pp. 998–1009, 2020.
14. Y.-H. H. Tsai, S. Bai, M. Yamada, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 3, pp. 1456–1470, 2022.
15. M. K. Hasan, M. S. Rahman, and M. A. Hossain, "Deep learning-based multimodal emotion recognition: A survey," Information Fusion, vol. 76, pp. 121–142, 2021.