International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

A Comprehensive Survey on Multimodal Sentiment and Emotion Analysis in Digital Communication

Akshatha Rithesh[1], Hanumanthappa M[2]

[1]Research Scholar, Department of Computer Science and Applications, Bangalore University, Bengaluru, India
Email: *akshatharithesh27[at]gmail.com*

[2]Senior Professor, Department of Computer Science and Applications, Bangalore University, Bengaluru, India
Email: *hanu6572[at]bub.ernet.in*

Abstract: ***With the significant increase in digital communication, the need for understanding human emotions in intelligent systems has gained utmost importance. It is difficult for the traditional sentiment analysis techniques to capture the complexity and subtlety of emotions expressed in online interactions. The goal of this survey paper is to understand and compare the recent advances in multimodal sentiment and emotion analysis, focusing on text, audio, visual, and other modalities to improve emotion detection. This survey encompasses recent papers and examines various methods, including deep learning, transformer-based models, and fusion strategies, as well as popular datasets such as CMU-MOSEI, IEMOCAP, and MELD. This paper discusses the strengths and weaknesses of existing approaches, identify gaps such as modality imbalance, interpretability, and cross-cultural generalization, and suggest future directions like lightweight architectures, explainable AI, and underutilized modalities like emojis and physiological signals. This paper will give a direction for researchers and practitioners to build more robust and context-aware emotion recognition systems.***

Keywords: Multimodal Emotion Recognition, Sentiment Analysis, Affective Computing, Fusion Techniques, Transformer Models, Digital Communication

1. Introduction

In recent years, communication has shifted from a traditional means to digital. This digital communication is in the form of social media, video calls, online classes, messaging apps, etc. This shift has brought both ease and complexity. Though this has simplified the connection between people, it has increased the complexity in understanding the emotions expressed in these communications. With the increase in the use of digital communication, emotion detection plays a vital role. It is used in mental health tools that help to monitor what people post on social media. In virtual classrooms where tutors need to adjust their teaching based on student mood. In chatbots, it will help them respond kindly. The systems that are used for emotion detection must be smarter than ever, as in digital communication, there is a lack of facial cues, voice tone can get flattened, and there is no body language support. Traditionally, emotion analysis was done on the text only. But text alone will miss the tone, pitch, facial expression, and gestures. Text may lead to missed signals like Sarcasm and sadness, and hence the accuracy drops. Especially in casual, cross-cultural spaces where expression is very crucial.

Multimodal sentiment and emotion analysis combines text, voice in the audio, facial expressions, and gestures in video. A model that can detect all these, understands the feelings better, like a human. With the recent transformer models and their contrastive learning tricks, things have moved fast [1]. The patterns they catch are sharper and cleaner, comparatively better than the single-mode systems. When compared, it is visible that the multimodal is outperforming the unimodal across the board [4]. Surveys keep pointing in the same direction, too. Everyone's shifting toward using more than one stream [5][11]. These systems also hold up to problems like different languages, Noisy data, and Sparse signals, etc [7], [8], [13]. Even in real-life settings where emotions fluctuate in conversations, the reactions they're showing promise [10]. Some techniques stand out more, like joint learning, where everything trains together [13]. A combination of methods that is capable of deciding when to focus and when to let go. And models that adjust based on which modality has more weightage [11].

2. Paper Contributions and Structure

This survey aims to understand the direction in which multimodal sentiment and emotion analysis is headed. It starts by understanding what sentiment and emotion are, and what their importance is. Then it walks through the various modalities like text, audio, and visuals. Each has a role, and each one exhibits emotion better than the other. Further Techniques in Multimodal Learning are discussed and compared to see what works better. With the help of this comparison, the gaps arising were identified, like real-time issues, ethics, missing data, and lack of context, etc. Some possible ways forward for these gaps are also discussed.

The following way this paper is laid out: The Third Section looks at the background and motivation behind this paper. Fourth Section elaborates the individual modalities. Fifth Section breaks down the Techniques used in Multimodal Learning. Section Six summarizes the popular datasets used in the paper reviewed. The Seventh Section deal with the challenges and research gap associated with the sentiment analysis. In Section Eight a Comparative Study of Methodologies is done. Section nine gives us the conclusion and section Ten points to the future direction obtained by the comparative analysis.

3. Background and Motivation

There has been a tremendous growth in sentiment analysis from the early 2000s[16]. During the initial stage, it was all about the text. Researchers used lexicons and a few rules,

along with some machine learning models, to find if the text is exhibiting positive, negative, or just neutral emotions. But the quantity of data increased along with the emergence of CNNs and RNNs. They made things smarter. Then came BERT and the transformers and models pushed text-based sentiment detection into new territory [4], [5]. These helped to make the results better and accurate. Texts alone don't always carry the weight of the emotions expressed [1], [4]. Hence, there was a shift from text-only to multimodal. From just reading words to listening, watching, and sensing. Researchers started building systems that could mix inputs—text, audio, video—into one understanding. Datasets like CMU-MOSI, MOSEI, and IEMOCAP made this possible, and fusion techniques caught up too [1], [3], [13].

The newer systems are even smarter as they use self-supervised learning to reduce dependence on labels [2]. They learn shared meanings across different modalities, which is called modality-invariant representations [13]. Some models are language-guided as they figure out which signal to trust more in a given moment [13]. Transformers take it all in, together [10]. This results in a more robust, more accurate, and more sensitive to context [3], [11]. Ekman gave six universal emotions: anger, joy, sadness, surprise, fear, and disgust [17]. It's shaped how emotion datasets are built, like IEMOCAP and MELD [1], [10]. Plutchik added more nuance. His wheel mapped emotions in layers, by intensity, polarity, and complexity [18]. It helped multimodal systems to handle subtle cues like things that aren't easy to name but are felt [9]. Computational models were strongly supported by these theories. Each modality plays an important part in sentiment analysis. Text brings semantics, but sometimes it leaves out the tone [4],[5], [11]. Audio data carries emotions in the form of pitch, volume, rhythm, etc. The things that one cannot read but can hear [3], [7], [9]. Visual multimodals show expressions, eye movement, and body language, which can be analysed for emotion even if there is no voice [2], [10], [13]. Emojis and symbols are used in informal communication, and they convey emotions which usually text alone often misses Zhu et al., 2023. Gestures and physiological signals, such as hand motion and heart rate, are also embedded into emotion research, particularly in real-world scenarios [10]. Models now have to rely on joint transformers to understand the layers [10], gated attention networks [14], and smart fusion guided by context [13]. It is similar to how humans retrieve meaning from sound, sight, and words, all at once. Machines are still learning to be like humans, but they're getting closer.

4. Modalities in Emotion Detection

Human emotions are very complex, and to detect these emotions correctly, systems should Emotion detection needs more than just words. That's where data modalities like text, audio, visuals, and even emojis come in handy. Each of these modalities offers a better understanding of emotion, and the combination of them gives us a clearer picture. If only text data is used as an input, though a fair level of sentiment analysis could be done, still layered sentiments like Sarcasm, Cultural context, etc, get lost. Emojis twist meanings. Abbreviations, informal grammar, and slangs used especially on social media makes it a difficult job [4],[5], [11]. That is the reason audio and visuals matter. Audio tells us about the emotions with the help of pitch, tone, and speech rate. Tools like OpenSmile and Praat help extract the emotion with the help of models like RNNs, CNNs, and transformers as they help to process them [3], [7], [9]. Visuals will have a combination of facial movements, eyes blink, and changes in expressions. Systems like FACS are used to break these down. CNNs, 3D-CNNs, even models like CLIP turn visuals into readable emotion cues. Real-world datasets like Aff-Wild2 and EmotiW make training more grounded [2], [10], [13]. And then, the emojis, GIFs, body language, and eye tracking are subtle, but they add depth in chats, posts, memes, where tone isn't spoken, it's shown [5], [10]. These newer signals are still being explored, but they're quickly becoming part of the emotional landscape.

5. Techniques in Multimodal Learning

It is important to make all the different modalities work together by combining them in a technique to extract the emotions hidden in them. Let us review the core methods used by the researchers in the paper reviewed.

a) Early Fusion (Feature-Level)
This is a very straightforward and powerful technique. It takes features from each input, like BERT embeddings for text, MFCCs for audio, and facial vectors for visuals, and blends them all before the model even gets to work. It lets the system see how different signals interact with each other. But it has its own risks. If one of the inputs is noisy or missing, then the entire thing can break. And syncing everything perfectly together is not always easy.

b) Late Fusion (Decision-Level)
In late fusion, each modality works on its own. Each makes a prediction using its model, and the outputs generated are combined at the decision level to make a final classification. It is robust, but it misses the magic that happens when signals communicate to each other

c) Intermediate / Hybrid Fusion (Embedding-Level)
Here, each modality first builds its understanding and its embeddings. Later, it will be combined. This keeps the benefits of independence while allowing a combination of modalities output. Techniques like MISA [13] and ALMT [13] use this technique. They balance learning both what's unique and what's shared across modalities. Hybrid fusion tends to generalize better results as it interprets things more like how humans do.

d) Attention-Based Fusion
This technique comes in handy when all the signals are not equal. These models learn to understand the weightage of each input differently. If the visual feed is messy, it could lean more on the voice or the words. Models like UniMSE [1]and Gated Fusion [14] use attention to dynamically shift focus from one modality to another. It's smart, especially for real-world data, where things aren't always clean or clear.

e) Transformer-Based Multimodal Models
Transformers changed the game. There is no necessary to decide on the outcome early or later on . All that is required is to feed everything in, and let the layers sort it out.

A few examples:

- MMBT trains jointly on text and image tokens.
- VisualBERT blends image regions with sentences in the same transformer.

- FLAVA learns both contrastive and generative signals between images and text.
- UniVL goes a step further, with vision and language pretraining for all sorts of tasks.

These models are setting the standard by reading emotions better, classifying sentiment, and even understanding video.

6. Datasets in Multimodal Sentiment and Emotion Analysis

Data is a very crucial part of Multimodal emotion Analysis. Emotion detection will be a guesswork rather than proper detection if there is no synchronized text, audio, and video. The datasets most cited across the 15 reviewed papers include CMU-MOSI (Carnegie Mellon University - Multimodal Opinion Sentiment Intensity), CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity), IEMOCAP (Interactive Emotional Dyadic Motion Capture), MELD (Multimodal EmotionLines Dataset), RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), CH-SIMS (Chinese Multimodal Sentiment Dataset), UR_FUNNY, and some visual-only sets like Aff-Wild2 and EmotiW. A few even go beyond the standard, building a custom corpus to fit the specific needs of their research.

a) CMU-MOSI & CMU-MOSEI
Cited in: [1], [3], [4],[5], [11], [13], [14]. Modalities used are Text, Audio, and Visual. Both CMU-MOSI & CMU-MOSEI were built from YouTube monologue reviews. The videos were cut into opinion-level clips, and each utterance was aligned across modalities. Annotators rated them on sentiment or emotion categories using crowd-based scores [13]. It remains among the most balanced and go-to datasets for benchmarking multimodal sentiment models.

b) IEMOCAP
Cited in: [1], [3], [11], [8], [10], [14]. Modalities used are Text, Audio, and Visual. These data sets are recorded in a lab, with actors performing both scripted and improvised dialogues. Emotions are labelled (happy, sad, angry, etc.), and motion capture data is included for fine movement tracking. It is great for structured emotion studies. A bit acted, though not completely natural.

c) MELD
Cited in: [1], [14]. Modalities used are Text, Audio, and Visual. The datasets are pulled from the TV series Friends. MELD expands the Emotion Lines dataset. Dialogue is split into utterances, and emotions are labelled through crowdsourcing. It handles multi-party conversations well. It brings context and noise. Real-life-like chat. Good for emotions in group conversations.

d) RAVDESS
Cited in: [8], [9]. Modalities used are Audio and Visual. 24 actors performed emotional expressions in both speech and song in a Controlled environment. Categories include calm, sad, angry, fearful, and more. It is Reliable for clean audio-visual emotion detection. But the data is acted on, and English-only.

e) CH-SIMS
Cited in: [13]. Modalities: Text, Audio and Visual. It is Similar to MOSI but in Chinese and features real video reviews, with utterance-level sentiment ratings. It broadens the research scope beyond English. Helps in multilingual and cross-cultural modelling.

f) UR_FUNNY
Cited in: [13]. Modalities used are Text, Audio, and Visual. It includes Stand-up comedy videos, annotated for punchlines and laughter. It's spontaneous, real-world, and tricky. With Humor, sarcasm, and timing, it's a goldmine for studying fine-grained emotional cues.

g) Aff-Wild2 & EmotiW
Cited in: [2]. Modalities used are Visual (some Audio). It is sourced from YouTube. Emotions were labelled using FACS and crowd-sourced inputs. EmotiW mixes both acted and spontaneous clips in uncontrolled settings. It is Essential for real-world visual emotion modelling, especially under noisy, unpredictable conditions.

h) Custom Datasets
Cited in: [2], [7], [10]. It is built from scratch to suit the paper's goals, cross-lingual settings, emotion in the wild, or physiological signals. These datasets often include sensor data, multilingual text, or unique annotations. When benchmarks don't fit, researchers build what they need. These datasets often push boundaries.

7. Challenges and Research Gaps

In this section, we will try to analyze the various roadblocks associated with sentiment analysis. Even with all the tech breakthroughs, emotion detection still isn't easy. Models are smarter, but there are issues, some old, some new.

a) Data Imbalance & Modality Sync
Datasets like CMU-MOSI and MELD do not treat the data fairly. Emotions like neutral show up way more than rare ones like disgust or fear [1], [4], [13]. That imbalance affects the learning. Another issue is getting the text, audio, and visual streams to line up perfectly. Not everyone speaks at the same speed, words spoken might be incomplete, and video may lag. This makes utterance-level fusion trickier than it looks [1], [3].

b) Missing or Noisy Inputs
Things are not perfect in the real world. Someone might turn off their camera in between recordings. Audio might get ruined by background noise, and when that happens, models may fail to adapt to these incomplete or noisy data. Few systems are good at adjusting dynamically when a modality drops out [8], [10], [14].

c) Cross-Cultural & Cross-Language Limits
Most research still sticks to English, but emotions aren't just an English thing. Non-English languages, dialects, and cultural gestures do not show good performance in models developed using the English language only [7], [13], [13]. What feels "angry" in one culture might not in another. Generalization is a real gap that need to be addressed.

d) Black Box Models
Transformer-based fusion models work well, but we don't always know why. They lack transparency as users, especially in sensitive domains like mental health or education, need to trust the output. The current models' explanations about the result are hard to come by [4], [5], [11].

e) Not Built for Real-Time
Most models are designed to run offline, with lots of data and lots of time. Social media, Mobile apps, live video

chats, or smart wearables need quick, efficient models that are capable of analyzing emotions in real time. That is still an underdeveloped area [11], [9], [10].

f) Dataset & Labeling Flaws
Some datasets, like RAVDESS or IEMOCAP, use acted emotions. They're clean, but not always *real* [2], [5], [13]. Others have inconsistent labels, especially for mixed or subtle feelings like sarcasm or humor. That affects the performance and hence the outcome.

g) Neglected Signals
Text, audio, and video are the modalities that get most of the attention. But other data like emojis, heart rate, Eye movement, and Hand gestures carry emotion, too. In digital spaces, they're everywhere. Yet many models still ignore them [10], [14].

8. Comparative Study of Methodologies

When we trace the evolution of multimodal sentiment and emotion analysis, from early deep learning techniques to sleek, transformer-driven architectures. It is visible that the field is shifting fast, from classic fusion methods to smarter and more flexible models. This section discusses the various methodologies used by the papers reviewed.

a) Traditional Deep Learning Approaches
With Traditional Deep Learning Approaches, sentiment and emotion analysis started getting serious. Papers like MISA [13], TransModality [12], and Gated Attention Fusion [14] relied on LSTMs, CNNs, and modality-specific encoders. These models extract the unique features in each input and then merge them using basic fusion methods like concatenation or gated units. Their performance was decent, especially on clean datasets, but they didn't scale well. Also, they could not handle Real-time data, struggled with noise, and lacked flexibility.

b) Transformer-Based Architectures
Next comes Transformers. Models like those in [1], [3], [11], and [10] started using them not just for text, but for all the types of modalities, audio, visuals, and fusion too. Cross-modal attention, dual-stream transformers, multi-head self-attention, everything was in there. These methods handle long-term dependencies way better than the traditional methods. And they allow joint training across modalities. It is no surprise that they are much better than older models on datasets like CMU-MOSI and MOSEI.

c) Contrastive and Self-Supervised Learning
More recent papers started pushing toward efficiency and alignment. Systems like UniMSE [1]and ALMT [13] brought in contrastive learning. These models don't just learn from labels, but they learn from the structure of the data. They align different modalities into a shared space. This results in better generalization, even with less supervision. They also handle missing or mismatched modalities way better.

d) Adaptive and Attention-Based Fusion
Some models got smarter with the help of fusion. Instead of merging all signals equally, they started weighing them. Gated Attention [14], Dual Attention Transformers [7], and Hybrid Transformer Fusion [9] use attention to figure out which modality to trust more in a given moment. So if the audio is noisy or a face is blurry, the model adapts. This makes them way more reliable for real-world use.

e) Survey and Taxonomy-Oriented Works
A few papers didn't build new models, but they mapped the field. Studies like [4], [5] and [11] categorize multimodal techniques into early, late, hybrid, and attention-based fusion. They give the big picture, like trends, gaps, and evolution paths. These surveys help frame where things are headed.

Table 1: Comparative study of Multimodal Sentiment and Emotion Analysis Methodologies

#	Title	Year	Modalities	Datasets	Approach / Contribution	Performance Highlight	Research Gap
1	TransModality: Transformer End-to-End Multimodal Fusion [12]	2020	T, A, V	MOSI, MELD, IEMOCAP	Modality translation transformers with aligned attention	F1 ~78% on MOSI	Lacks real-time adaptability and robustness to noise
2	MISA: Modality-Invariant & Specific Representations [13]	2020	T, A, V	MOSI, MOSEI, UR_FUNNY	Learns shared and unique features per modality	SOTA on MOSI & MOSEI	Weak on domain transfer and low-resource domains
3	Adaptive Language-Guided Multimodal Transformer (ALMT) [11]	2023	T, A, V	MOSI, MOSEI, CH-SIMS	Text-guided filtering of other modalities	+3–6% over strong baselines	High data demand and black-box behavior
4	UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition [1]	2022	T, A, V	MOSI, MOSEI, MELD, IEMOCAP	Contrastive fusion of syntactic & semantic layers	Achieves SOTA on MOSI & MELD	Real-time application and cross-domain generalization are needed
5	Facial Emotion Recognition with Inter-Modality Attention Transformer [2]	2023	A, V	Custom video datasets	Self-supervised learning with attention fusion	86.4% on benchmark facial emotion tasks	Lacks cross-cultural evaluation
6	Survey on Deep Learning-Based Multimodal Emotion Recognition [4]	2023	T, A, V	Multiple	Taxonomy of deep learning pipelines	n/a (survey)	Lack of unified benchmarks and interpretability
7	Survey of Multimodal Sentiment Fusion Methods [5]	2023	T, A, V	Multiple	Overview of early, late, hybrid fusion methods	n/a (survey)	Minimal real-world deployment & robustness studies

8	Using Transformers for Multimodal Emotion Recognition [6]	2023	T, A, V	Multiple	Review of transformer-based architectures	n/a (survey)	Scarcity of explainable, lightweight transformer models
9	Cross-Language Speech Emotion Recognition with Dual Attention Transformers [7]	2023	A	Multiple speech datasets	Dual attention for cross-lingual emotion recognition	Improved multilingual generalization	Requires deeper linguistic and cultural adaptation
10	Noise-Resistant Multimodal Transformer (NORM-TR) [8]	2023	T, A, V	RAVDESS, IEMOCAP	Transformer with noise-aware training	Resilient under noisy conditions	Domain transfer and generalizability underexplored
11	Large Language Models Meet Text-Centric Multimodal Sentiment Analysis [15]	2023	T, V	MOSI, MOSEI	Explores LLMs' integration in multimodal sentiment	Conceptual; reviews performance trade-offs	LLM adaptation for multimodality still early-stage
12	Hybrid Transformer Fusion for Speech Emotion Recognition [9]	2024	A	RAVDESS	Transformer + dual cross-entropy loss	Outperforms traditional fusion on RAVDESS	Limited to audio-only data
13	Joint Multimodal Transformer for Emotion Recognition in the Wild [10]	2024	T, A, V, P	Wild multimodal sets incl. physiological signals	Joint attention + prompt tuning	Effective on real-world multimodal data	Causal reasoning and emotion source detection underdeveloped
14	Multi-Modal Emotion Recognition by Text, Speech and Video Using Transformers[3]	2024	T, A, V	IEMOCAP	Transformer-based early & late fusion	Improved over baseline multimodal fusion	Latency and real-world usability were not discussed
15	Gated Attention-Based Multimodal Sentiment Analysis [14]	2024	T, A, V	MOSI, MOSEI	Cross-modal attention with a gating mechanism	~83.9% accuracy on MOSI	Poor coverage of non-English and short text scenarios

T = Text, **A** = Audio, **V** = Visual, **P** = Physiological

9. Conclusion

This survey provided a comprehensive overview of recent advances in multimodal sentiment and emotion analysis, emphasizing the integration of text, audio, visual, and auxiliary modalities in digital communication settings. Through the comparative study of 15 peer-reviewed papers from the last five years, we observed a clear evolution from early feature-level fusion and RNN-based architectures to transformer-based multimodal models that excel in capturing deep cross-modal dependencies.

We highlighted how different modalities contribute uniquely to affect detection, text offers semantic richness, audio reflects tone and prosody, visuals convey facial expressions and gestures, and emojis or symbols add emotional cues in informal settings. While transformer architectures like UniMSE, MISA, and VisualBERT deliver state-of-the-art performance, issues like modality imbalance, cross-lingual generalization, real-time deployment, and model interpretability remain significant hurdles. Moreover, while benchmarks like CMU-MOSEI and IEMOCAP have advanced the field, there is still a lack of large-scale, naturalistic datasets covering diverse languages, cultures, and contexts.

10. Future Directions

Emotion detection is a fast-evolving field, yet it has its loopholes. The gaps we have discussed here have given a few directions for future work. Some of the things which can be worked on in the future are.

1) Robustness to Missing or Noisy Modalities:
Real-world data, which is used as input in models, is not perfect. It can have various problems like audio cuts, camera failure, and people going out of the screen, etc. Future models need to adapt dynamically to any kind of issues faced in the data. They need to adjust based on what's available or missing. This will be an important factor, especially for deployment in unpredictable environments.
2) Cross-Lingual and Cross-Cultural Learning:
Most datasets are still English-based, and that has created a problem for datasets from another language or culture. Emotions don't rely on any language, gestures, or tone. Future models should be trained on diverse languages and cultures to truly scale globally and avoid bias in interpretation.
3) Explainability in Multimodal Models:
Multimodal models are powerful, but often not transparent. We need to know the workings behind models. We need to understand that Why does a system thinks someone is sad or angry. Visualizing attention, showing based on modality, the decision was taken, or attributing predictions to specific features can help in understanding the working. Especially in domains like mental health or education, where trust matters more.
4) Lightweight Models for Real-Time Use:
Heavy transformer stacks won't fit on mobile or edge devices. What we need are Smaller, smarter models that still perform well. That way, emotion-aware systems can run in real-time—in classrooms, video calls, wearable tech, or on phones.
5) Integrating Underexplored Modalities:
There's more to emotion than text, voice, and face. Emojis, GIFs, eye gaze, heart rate, even touch? They all

count. Incorporating these types of subtle signals can help add more texture to models, particularly where virtual reality and gaming and social media are concerned.

6) Multimodal Pretraining & Transfer Learning:
Just like BERT revolutionized text, we need a transition to multimodal learning. Task-agnostic pretrained general emotion models. Tools like FLAVA, UniVL, and Multimodal GPTs are just starting here.

The future of emotion AI, in short, is flexible, broad, interpretable and fast. We're closer. But there is work to do.

References

[1] Hu, G., Lin, T. E., Zhao, Y., Lu, G., Wu, Y., & Li, Y. (2022). Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*

[2] Chaudhari, A., Bhatt, C., Krishna, A., & Travieso-González, C. M. (2023). Facial emotion recognition with inter-modality-attention-transformer-based self-supervised learning. *Electronics*, *12*(2), 288.

[3] Shayaninasab, M., & Babaali, B. (2024). *Multi-modal emotion recognition using pretrained transformers*. arXiv preprint arXiv:2402.11599. https://doi.org/10.48550/arXiv.2402.11599

[4] Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, *25*(10), 1440

[5] Zhu et al., 2023 Zhu, L., Zhu, Z., Zhang, C., Xu, Y., & Kong, X. (2023). Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusi5]n*, *95*, 306-325.

[6] Hazmoune, S., & Bougamouza, F. (2024). Using transformers for multimodal emotion recognition: Taxonomies and state of the art review. *Engineering Applications of Artificial Intelligence*, *133*, 108339.

[7] Zaidi, S. A. M., Latif, S., & Qadir, J. (2023). Cross-language speech emotion recognition using multimodal dual attention transformers. *arXiv preprint arXiv:2306.13804*.

[8] Liu et al., 2024 Liu, Y., Zhang, H., Zhan, Y., Chen, Z., Yin, G., Wei, L., & Chen, Z. (2024). Noise-resistant multimodal transformer for emotion recognition. *International Journal of Computer Vision*, 1-21.

[9] Huang, J., Tao, J., Liu, B., Lian, Z., & Niu, M. (2020, May). Multimodal transformer fusion for continuous emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3507-3511). IEEE.

[10] Waligora, P., Aslam, M. H., Zeeshan, M. O., Belharbi, S., Koerich, A. L., Pedersoli, M., ... & Granger, E. (2024). Joint multimodal transformer for emotion recognition in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4625-4635).

[11] Zhang, H., Wang, Y., Yin, G., Liu, K., Liu, Y., & Yu, T. (2023). Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. *arXiv preprint arXiv:2310.05804*.

[12] Wang, Z., Wan, Z., & Wan, X. (2020). TransModality: An End2End Fusion Method with Transformer for Multimodal Sentiment Analysis. *WWW 2020*. https://doi.org/10.1145/3366423.3380000

[13] Zadeh, A., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2020). *MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. ACM Transactions on Multimedia Computing, Communications, and Applications, 16*(3), 1–23.

[14] Kumar, A., & Vepa, J. (2020). *Gated mechanism for attention based multimodal sentiment analysis*. arXiv preprint arXiv:2003.01043

[15] Yang, H., Zhao, Y., Wu, Y., Wang, S., Zheng, T., Zhang, H., Ma, Z., Che, W., & Qin, B. (2024). *Large language models meet text-centric multimodal sentiment analysis: A survey*. arXiv preprint arXiv:2406.08068

[16] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up? Sentiment classification using machine learning techniques." *arXiv preprint cs/0205070* (2002).

[17] Ekman, Paul. *An Argument for Basic Emotions*. Cognition and Emotion, vol. 6, no. 3–4, 1992, pp. 169–200. Taylor & Francis, https://doi.org/10.1080/02699939208411068.

[18] Plutchik, Robert. "The Nature of Emotions." *American Scientist*, vol. 89, no. 4, 2001, pp. 344–350. JSTOR

Author Profile

Ms. Akshatha Rithesh is a Ph.D. scholar in Computer Science at Bangalore university and also working as an Assistant Professor at S-VYASA School of Advanced studies. Her specialization is in Natural Language Processing and Multimodal Sentiment Analysis. Her research focuses on integrating textual, visual, and symbolic modalities (emojis, hashtags) to enhance emotion detection in digital communication.

Dr. Hanumanthappa M is working as a Professor and Chairman, Department of Computer Science and Applications, Bangalore University, Bangalore. He has vast teaching and industry experience spreading over two decades at postgraduate level and various IT industries. He has authored more than 100 research papers in reputed International Journals and peer reviewed Conference papers at International and National levels. He is the receipt of the best paper publication award by many organizations. His research interest includes Data Mining, Information Retrieval, Network Security, and Natural Language Processing.