

## **Multimodal Sentiment Analysis: A Systematic review of History, Datasets, Multimodal Fusion Methods, Applications, Challenges and Future Directions**

**Ayaz Ahmed Faridi<sup>1</sup> Tryambak Hiwarkar<sup>2</sup>**

<sup>1</sup>Research Scholar PhD (CS), Sardar Patel University Balaghat (M.P.), INDIA

<sup>2</sup>Professors, Sardar Patel University Balaghat (M.P.), INDIA

**Email-[ayazahmed.faridi@gmail.com](mailto:ayazahmed.faridi@gmail.com)**

### **Abstract**

Sentiment analysis (SA), a buzzword in the fields of artificial intelligence (AI) and natural language processing (NLP), is gaining popularity. Due to numerous SA applications, there is an increasing need to automate the procedure of analysing the user's feelings concerning any products or services. Multimodal Sentiment Analysis (MSA), a branch of sentiment analysis that uses many modalities, is a rapidly growing topic of study as more and more opinions are expressed through videos rather than just text. Recent advances in machine learning are used by MSA to advance. At each stage of the MSA, the most recent developments in machine learning and deep learning are used, including sentiment polarity recognition, multimodal features extraction, and multimodal fusion with reduced error rates and increased speed. This research paper categorises several recent developments in MSA designs into 10 categories and focuses mostly on the primary taxonomy and recently published Multimodal Fusion architectures. The 10 categories are: early fusion, late fusion, hybrid, model-level fusion, tensor fusion, hierarchical, bi-modal, attention-based, quantum-based, and word-level fusion. The primary contribution of this manuscript is a study of the advantages and disadvantages of various architectural developments in MSA fusion. It also talks about future scope, uses in other industries, and research shortages.

**Keywords:** Affective Computing, Sentiment Analysis, Multimodal Fusion, Fusion technique

### **Introduction**

People are more eager to express and share their thoughts online about both daily activities and major world issues after the arrival of Web 2.0. These efforts have also been tremendously facilitated by the development of social media, which has given us a public forum to share our opinions with people around the globe. Customers can express their ideas by using these web-based electronic Word of Mouth (eWOM) remarks, which are widely employed in the commercial and service sectors. As a result, emotional analytics has become a fresh and intriguing field of study. The two categories of emotional analytics are sentiment analysis, commonly known as opinion mining, and emotion recognition. To gather and analyse public sentiment and viewpoints, sentiment analysis is utilised. It has been gaining acceptance in academic circles, government agencies, and service industries. Emotion recognition is the technique of identifying human emotions. The extent to which people are able to discern the emotions of

others varies substantially. Technology's usage to help people identify their emotions is a relatively recent area of research. Automatic detection of a person's mood or sentiment is known as affective computing. It's a brand-new area of study whose goal is to make it possible for intelligent systems to recognise, experience, infer, and understand human emotions. The fields of computer science, psychology, social science, and cognitive science are all included in this multidisciplinary field. Despite being two separate academic fields, sentiment analysis and emotion detection are grouped together under the Affective Computing umbrella [1].

Our daily lives are significantly influenced by our emotions and sentiments. In human-centered environments, they support decision-making, education, communication, and situation awareness. To provide computers cognitive capabilities, AI experts have been working on it for the past 20 years. So, much like people, machines can recognise, evaluate, and convey emotions and sentiments. Affective

computing is the cause of all of these endeavours. User reviews of goods, services, and events are very valuable from a business standpoint. They are very helpful to organisations in terms of product/service monitoring, fostering stronger customer relationships, creating better marketing tactics, and enhancing and inventing their products/services. They also help other users make decisions, like as buying a new product. Customers carefully consider what is being said on various web platforms and on social media before making decisions about whether to use or purchase any goods or services. Because of this, emotion recognition and sentiment analysis have gained popularity in research [2]. However, it is difficult to automatically analyse a lot of data and produce an aspect summary. It might be challenging to recognise and extract sentiments from natural language. It demands a thorough comprehension of language's syntactic and semantic rules. Additionally, as opinion writings are frequently informal, they frequently contain slang, irony, sarcasm, acronyms, and emoticons. This complicates analysis even further. Sentiment analysis uses information retrieval, data mining, and natural language processing techniques to find and extract opinions from massive amounts of text. While MSA derives peoples' emotions, thoughts, and sentiments from behavioural observations. Physiological indicators, speech, written records, motions, and facial expressions are all examples of behavioural hints.

Since humans and emotion are closely intertwined, emotion comprehension is essential to the development of artificial intelligence that resembles humans (AI). Natural language frequently reflects a person's emotions. Due to its numerous uses in sentiment analysis, review-based systems, the application in the medical industry, and other domains, emotion recognition has gained popularity in the field of NLP [3]. A group of scholars talked about the notion of identifying emotion in news headlines [4]. A number of emotion lexicons [5] have been developed in order to address the problem of textual emotion recognition.

Due to its capacity to extract opinions from a wealth of publicly available

conversational data on platforms like Facebook, YouTube, Reddit, Twitter, and others, conversational or multimodal emotion recognition is already gaining traction in NLP. It could also be used in a variety of other fields, including criminology, health care (as a tool for mental health prediction), education (for student counselling), understanding student frustration, and many more. Establishing emotion-aware interactions that demand a thorough understanding of the user's feelings also requires emotion recognition in a conversational setting. Conversational emotion identification systems must be efficient and scalable to achieve these requirements. However, it is a difficult topic to handle because of numerous research roadblocks. Any activity that involves automated learning from data or experience is now included in the well-established field of machine learning. The core of machine learning is the capability of a programme or machine to improve performance of specific tasks by being exposed to data and experiences. Deep learning is a machine learning area that is receiving a lot of attention right now. . The knowledge generated by Deep Learning algorithms has largely gone unused in the context of Big Data Analytics. Many Big Data applications have made use of deep learning, particularly to enhance classification modelling results. Modern deep learning algorithms can produce improved categorization modelling results that can be used for a variety of applications [6]. A popular deep learning-based approach for image processing is the convolution neural network (CNN). In research by [6] , CNN has a thorough overview of all recent advancements. The vast amount of data on the internet might be structured, semi-structured, or unstructured, and it can originate from a number of different databases. Structured data is well ordered and in a clearly defined, standardised format. Semi-structured data is arranged in a certain format, such as email or XML data, but does not adhere to the traditional tabular data paradigm. Although they have an internal data structure and can be in textual or non-textual formats, unstructured data, which typically refers to big data, lacks a prescribed data model. It's

challenging to process such vast amounts of data. Deep Learning techniques are quite popular right now since they speed up automation by using hidden layers to handle difficult processing. Its design is adaptable enough to handle a variety of tasks, including sentiment analysis for audio-visual and text data. Deep learning can automate and speed up all of the laborious stages of the sentiment analysis process, including feature selection, feature extraction, learning the appropriate parameters, processing feature vectors, and providing predictions. In DL-based techniques, hidden layers are utilised to perform complicated processing similar to that of the human brain between the input and output layers and serve as a "black box," preventing the identification of data representations in the middle hidden levels. More complex facts can be learned via the DL approach. This survey focuses on various fusion methods for combining textual, visual, and auditory features for multimodal sentiment analysis. Here is a summary of the most popular and most current developments in dataset generation for multimodal sentiment analysis. The common datasets and their stages of development are described. Current research's most advanced fusion approaches are discussed. All of these strategies use various notions for outcome improvisation, which are also described. On the basis of the most recent advancements in machine learning and deep learning, various multimodal categorization techniques are investigated. The implementation of sentiment analysis in many application domains and its potential future growth are highlighted. The use of diverse data, including videos, psychological signals, EEG signals, and other data, is also covered. There is also discussion of multilingual data, cross-language data, cross-domain data, and code-mixed data formats. The remainder of this essay is divided into the following sections. The foundations and requirement for multimodal sentiment analysis are explained in Section 2. Some of the most well-known datasets for multimodal

sentiment analysis are compiled in Section 3. A thorough overview of various fusion architecture using the most recent advances may be found in Section 4. Section 5 defines the MSA applications in several fields. The limitations of each fusion design are discussed in Section 6, along with the benefits and drawbacks of each model. MSA's future directions are discussed in Section 7, which is followed by a conclusion in Section 8.

### **Importance of Modalities**

Different modalities are employed in multimodal sentiment analysis to extract affective states from the conversation. The three most frequently used modalities are text, audio, and visual. Each helps to improve sentiment prediction, and research shows that bimodal and trimodal systems produce better results than unimodal systems. Every modality makes a significant contribution to increasing accuracy.

**Text Modality:** Out of all the modalities, text is the most prevalent. It is crucial in revealing the buried emotions. Although textual sentiment analysis produces excellent results, most opinionated data are now exchanged as videos rather than texts.

Visual elements aid in a better understanding of underlying feelings or opinions. For instance, if it says, "This is a fairly decent mouse." It is challenging to determine whether this is about the original animal mouse or the computer mouse using simply textual information. In this situation, visuals are helpful, and the combination of text and visuals creates a bimodal system that produces better results than unimodal systems.

**Audio Modality:** Acoustic features are used to extract text from videos, and they can also be used to determine the speaker's tonality. An analytical model is produced that is more accurate when all three modalities are combined. Sarcasm and common sense detection visuals may make incorrect predictions in the case of humor, but a combination of modalities can accurately identify the sentiments.

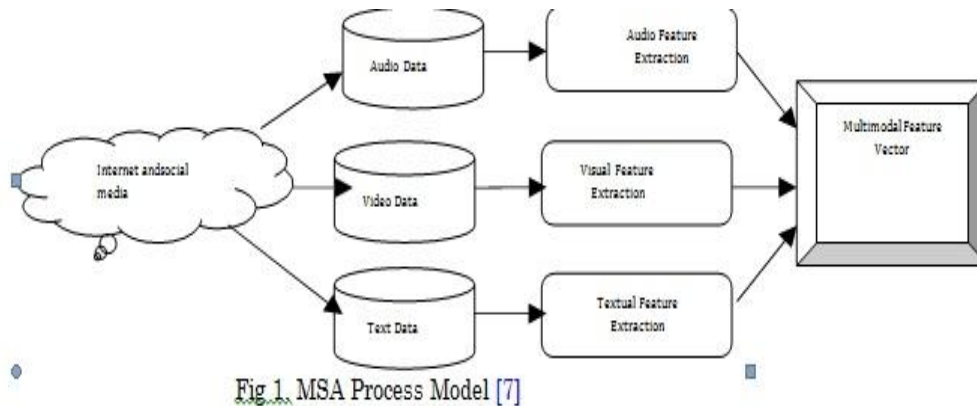


Fig 1. MSA Process Model [7]

### Popular Datasets

The steps for constructing a dataset for multimodal sentiment analysis are as follows.

**Data gathering** The phrases data and acquisition together make up the phrase "data acquisition." Raw, structured or unstructured facts and numbers are referred to as data. Data collection for a specific purpose is referred to as acquisition. The act of acquiring data from pertinent sources before it is saved, cleaned up, pre-processed, and used in other processes is referred to as "data acquisition." In order to create multimodal datasets for this purpose, videos are gathered from several online video sharing platforms. Automatic or semi-automatic tools are used to gather movies from the web using specific search terms. In order to determine whether a video is a monologue, web videos are examined for the. Typically, videos are chosen in which the speaker's focus is solely on the camera. Videos are gathered based on frequently searched themes, and videos from each channel are limited to a certain quantity for greater variety. **Data pre-processing:** Data may be absent, include false or inaccurate values, or lack pertinent or specified attributes. Pre-processing data is essential to improving the quality of the data. For the creation of the MSA dataset, a broad age range from the mid-20s to the late-50s is anticipated. The majority of speakers in the sample are English-speaking citizens of the United Kingdom or the United States. The dataset includes a tiny proportion of non-native yet proficient English speakers. Some of the speakers are wearing glasses. **Post-processing of the data:** Linguistic data from spoken language can help with emotional learning and is an essential part of context

interpretation. To make it easier for the auditory, visual, and textual modes to interact, the audio data is automatically transcribed. The video transcriptions from Amazon Transcribe and Google Cloud Speech API are of sufficient quality for our task. Aural elements and nonverbal cues like laughing, music, and theme are all present. Each word that is transcribed has a start and finish timestamp as well as its duration, and the spoken language transcriptions also include punctuation (such as periods, question marks, and exclamation points).

The alignment of the text with annotations (sample rate differences) and other modalities is made easier by these information.

**Data Annotation:** To train datasets for supervised machine learning and semi-supervised machine learning, data annotation is the process of labelling photos, video frames, audio, and text data. It enables the machine to comprehend the input and respond appropriately. Each annotator selects one of three emotive states for each clip: -1 (negative), 0 (neutral), or 1. (positive). For making annotations, separate human resources are employed. The average labelled result is then utilised to perform both regression and multi-classification tasks. Table 1 provides an overview of the most popular and most current developments in dataset generation for multimodal sentiment analysis. The name of the dataset is listed in the first column, and the year it was published is listed in the second column. The third column then lists the research that introduced the dataset, and the fourth column lists the various modalities that were covered by each dataset. The fifth column displays the total

number of review videos included in the dataset, and the sixth column displays the platform used to collect the videos.

The dataset's gender breakdown can be seen in the seventh column, which is followed by the language used in the videos in the eighth. The ninth column lists the various topics covered in videos, such as movie reviews and product reviews. A reference to a publicly accessible dataset is shown in the tenth column. Before the more recent datasets listed in Table 1, the next two datasets were frequently utilised for research purposes. The first was created by [8] and is known as the YouTube Opinion Dataset. It is a dataset for multimodal sentiment analysis. There are 47 YouTube videos included in it that are not related to any one theme. Twenty female speakers and 27 male speakers are present. The dataset consists of automatically extracted audio and visual components as well as text that has been manually transcribed.

Additionally, it contains utterances that were automatically extracted. The second is a multimodal sentiment analysis dataset in Spanish that was produced by [9]. It is called Spanish Multimodal Opinion Dataset. There are 105 videos in it that have had their sentiment polarity at the utterance level annotated. Most movies have 6–8 utterances, which are mechanically extracted from long pauses. The dataset has 550 utterances in total. Any of the offered datasets lack sentiment intensity annotations. They emphasise polarity a lot. Also, as indicated in the introduction, they tend to focus on video or utterance analysis rather than fine-grained sentiment analysis.

#### 4. Conclusion

Multimodal sentiment analysis has gotten a lot of attention in the last decade. They have a significant influence on sentiment predictions and emotion recognition, which has aroused the interest of researchers. Many scholars have made significant contributions to this field, changing the MSA fusion method to improve its efficiency. Changes in number of modalities that is bimodal or trimodal, context aware and speaker independent, humour and sarcasm detection, fusion techniques, application-specific modifications in architecture, developing various learning algorithms, and recommendation systems are some of the categories in which researchers have made

advancements in MSA. This manuscript summarizes recent developments in MSA architectures. The ten fundamental architectural advances in MSA are early fusion, late fusion, hybrid fusion, model-level fusion, tensor fusion, hierarchical fusion, bi-modal fusion, attention-based fusion, quantum-based fusion and word-level fusion. By examining several architectural alterations in MSA, it can be stated that MSA's most powerful architecture for multimodal sentiment analysis is the word level architecture, which is defined by classifying the target utterance using contextual information from neighbouring utterances in a video. This architecture is essentially composed of two components, the order of which varies by model. The first module is the Context Extraction Module, which is used to model the contextual link between neighbouring utterances in the video and highlight which utterances of the relevant contextual utterances are more important to predict the target one's emotion. In most recent models, this module is a bidirectional recurrent neural network-based module. The second module is the Attention-Based Module, which is responsible for merging the three modalities (text, audio, and video) and picking only the most important ones.

#### References

- [1] J. A. & V. J. D. Balazs, "Opinion mining and information fusion: a survey." *Information Fusion*, 27, pp. 95-110., 2016.
- [2] S. L. C. & C. J. Sun, " (2017). A review of natural language processing techniques for opinion mining systems," *Information fusion*, 36,, pp. 10-25, 2017.
- [3] Kratzwald, B., Ilic, S., Kraus, M., Feuerriegel, S., & Prendinger, H., "Decision support with text-based emotion recognition: Deep learning for affective computing," *Decesion Support Systems*, pp. 24-35, 2018.
- [4] C. & M. R. Strapparava, "Semeval-2007 task 14: Affective text," in *In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 2007.
- [5] S. & T. P. Mohammad, "Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon," in *In Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* , 2010.
- [6] Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., Ghayvat, H. "CNN Variants for Computer Vision: History,

- Architecture, Application, Challenges and Future Scope," *Electronics*, 10(20), 2021.
- [7] A. Gandhi, K. Adhvaryu and V. Khanduja, "Multimodal Sentiment Analysis: Review, Application Domains and Future Directions," in *2021 IEEE Pune Section International Conference (PuneCon)*, Pune, India, 2021.
- [8] Morency, L. P., Mihalcea, R., & Doshi, P., "Towards multimodal sentiment analysis: harvesting opinions from the web," in *ICMI '11: Proceedings of the 13th international conference on multimodal interfaces*, Alicante Spain, 2011.
- [9] Verónica Pérez Rosas, Rada Mihalcea, Louis-Philippe Morency, "Multimodal Sentiment Analysis of Spanish Online Videos," *IEEE INTELLIGENT SYSTEMS*, pp. 38-45, 2011.
- [10] Zadeh, A., Zellers, R., Pincus, E., & Morency, L. P. (2016). Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259..
- [11] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, Louis-Philippe Morency, "Multimodal Language Analysis in the Wild:CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, Melbourne, Australia, 2018.
- [12] Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, Bjorn Gambach, "SemEval-2020 Task 8: Memotion Analysis- The Visuo-Lingual Metaphor!," in *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain, 2020.
- [13] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, Kaicheng Yang, "CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020.
- [14] AmirAli Bagher Zadeh, Yansheng Cao, Simon Hessner, Paul Pu Liang, Soujanya Poria, Louis-Philippe Morency, "CMU-MOSEAS: A Multimodal Language Dataset for Spanish, Portuguese, German and French," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, 2020.
- [15] L. Stappen, A. Baird, L. Schumann and S. Bjorn, "The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements," *IEEE Transactions on Affective Computing ( Early Access )*, pp. 1-1, 2021.
- [16] A. G. Vasco Lopes, L. A. Alexandre and J. Cordeiro, "An AutoML- based Approach to Multimodal Image Sentiment Analysis," in *2021 International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China, 2021.
- [17] Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal and Chaitanya Ahuja, "FACTIFY: A Multi-Modal Fact Verification Dataset," in *De- Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2022*, Canada, 2022.
- [18] Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal and Chaitanya Ahuja, "Memotion 2: Dataset on Sentiment and Emotion Analysis of Memes," in *De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2022*, Canada, 2022.
- [19] Morency, L. P., Mihalcea, R., & Doshi, P., "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *In Proceedings of the 13th international conference on multimodal interfaces*, 2011.
- [20] Rosas, V. P., Mihalcea, R., & Morency, L. P., "Multimodal sentiment analysis of spanish online videos," *IEEE Intelligent Systems*, 28(3), pp. 38-45, 2013.
- [21] Poria, S., Cambria, E., Hussain, A. and Huang, G.B., "Towards an intelligent framework for multimodal affective data analysis.," *Neural Networks*, 63, pp. 104-116, 2015.
- [22] Park, S., Shim, H. S., Chatterjee, M., Sagae, K., & Morency, L. P., "Multimodal analysis and prediction of persuasiveness in online social multimedia," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(3), pp. 1-25, 2016.
- [23] Zadeh, A., Liang, P. P., Poria, S., Vij, P., Cambria, E., & Morency, L. P., "Multi-attention recurrent network for human communication comprehension.," in *In Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.