Information Fusion 91 (2023) 424–444

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions

Ankita Gandhi [a,*], Kinjal Adhvaryu [a], Soujanya Poria [b], Erik Cambria [c], Amir Hussain [d]

[a] Gujarat Technological University, India
[b] Singapore University of Technology & Design, Singapore
[c] Nanyang Technological University, Singapore
[d] Edinburgh Napier University, UK

ARTICLE INFO

Keywords:
Affective computing
Sentiment analysis
Multimodal fusion
Fusion techniques

ABSTRACT

Sentiment analysis (SA) has gained much traction In the field of artificial intelligence (AI) and natural language processing (NLP). There is growing demand to automate analysis of user sentiment towards products or services. Opinions are increasingly being shared online in the form of videos rather than text alone. This has led to SA using multiple modalities, termed Multimodal Sentiment Analysis (MSA), becoming an important research area. MSA utilises latest advancements in machine learning and deep learning at various stages including for multimodal feature extraction and fusion and sentiment polarity detection, with aims to minimize error rate and improve performance. This survey paper examines primary taxonomy and newly released multimodal fusion architectures. Recent developments in MSA architectures are divided into ten categories, namely 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. A comparison of several architectural evolutions in terms of MSA fusion categories and their relative strengths and limitations are presented. Finally, a number of interdisciplinary applications and future research directions are proposed.

1. Introduction

Since the introduction of Web 2.0, people have become keener to express and share their ideas on the web about day-to-day activities as well as global issues. The evolution of social media has also greatly aided these activities, providing us with a transparent platform to share our points of view with people all over the world. These web-based electronic Word of Mouth (eWOM) comments are widely used by business and service industries to allow customers to voice their opinions. As a result, affective analytics has emerged as a new and exciting research area. Sentiment analysis which is also known as opinion mining and emotion recognition are two types of affective analytics. Sentiment analysis is used to extract and analyse public mood and views. It has been gaining popularity amongst research communities, academia, government, and service sectors. The process of recognising human emotion is known as emotion recognition. People's ability to recognise other people's emotions varies greatly. The use of technology to assist individuals in recognising emotions is a relatively new study field. Affective computing refers to the automatic recognition of an individual's mood or sentiment.

It is an emerging field of research aiming at enabling intelligent systems to perceive, infer, and comprehend human emotions. The multidisciplinary field spans computer science, psychology, social science, and cognitive science. Despite the fact that sentiment analysis and emotion identification are two distinct disciplines, they are lumped together under the Affective Computing umbrella [1]. Emotions and sentiments play a significant role in our daily lives. They aid decision-making, learning, communication, and situation awareness in human-centric environments. For the past two decades, AI researchers have been researching ways to empower machines with cognitive ability to enable them to recognise, analyse, and express emotions and sentiments like humans. All of these endeavours are the result of affective computing. Product, service, and event reviews provided by users have a lot of commercial value. They assist other users in making decisions, such as purchasing a new product, and are extremely beneficial to businesses in terms of product monitoring, promoting better customer relationships, developing better marketing strategies, and improving and innovating their services. Consumers are keen to understand what is being said on multiple web platforms and on social media, and on that basis, they take their decisions to buy or use any products or services.

[*] Corresponding author.

https://doi.org/10.1016/j.inffus.2022.09.025
Received 13 June 2022; Received in revised form 25 September 2022; Accepted 27 September 2022
Available online 28 September 2022

This is why emotion recognition and sentiment analysis has become a growing research trend [2]. However, automatically analysing a large amount of data and producing a summary of aspects is a highly challenging task. Identifying and extracting sentiments from natural language is a difficult endeavour. It necessitates a deep understanding of language syntactic and semantic norms. In addition, opinion texts are typically informal which includes slang, irony, sarcasm, abbreviations, and emoticons. This makes analysis even more difficult. Sentiment analysis employs data mining, information retrieval and natural language processing approaches to identify and recover opinions from large textual sources. While MSA extract people's thoughts, feelings, and emotions from observations of their behaviours. Behavioural clues can be in the form of documented writings, facial expressions, speech, physiological signs, and movements.

Emotion is inextricably linked to humans which is why emotion comprehension is a crucial component of human-like artificial intelligence (AI). A person's mood is frequently reflected in their natural language. Emotion recognition has acquired prominence in the field of NLP due to its multiple applications in sentiment analysis, review-based systems, healthcare, and other fields [3]. The idea of detecting emotion in news headlines has been discussed by a group of researchers [4]. To tackle the challenge of textual emotion recognition, a number of emotion lexicons [5] have been established. Currently, conversational or multimodal emotion recognition is gaining traction in NLP due to its ability to mine opinions from a plethora of publicly available conversational data on platforms such as Facebook, YouTube, Reddit, Twitter, and others. It can also be utilised in other industries such as healthcare (e.g. as a tool for mental health prediction), education (understanding student frustration and for student counselling), deceptive detection in criminology and many others. Emotion recognition in a conversational environment is also necessary to enable emotion-aware interactions encompassing a deep comprehension of users feelings. To meet these demands, conversational emotion identification systems must be both effective and scalable. However, due to various research hurdles, this is a challenging subject to address.

Machine Learning is now a well-established field that encompasses any activity that involves automated learning from data or experience. The ability of a software or machine to enhance the performance of particular tasks by being exposed to data and experiences is at the heart of machine learning. Currently, deep learning is a hot field in machine learning. In the context of Big Data Analytics, the knowledge gained by Deep Learning algorithms has mostly gone unexplored. Deep learning has been used in a number of Big Data domains, primarily to improve classification and modelling outcomes. Improved categorization modelling results generated by modern deep learning algorithms can be applied to a wide range of applications [6]. Convolution Neural Network (CNN) is a deep learning-based algorithm which is widely used for image processing. In a study by [6], a comprehensive analysis of recent improvements in CNN is presented. The large amount of data on the web can usually be in any form like structured, semi structured or unstructured, and it may come from variety of databases. Structured data is normally in well-defined and standardized format and highly organized. Semi structured data doesn't follow the conventional tabular data model though they are organized in some prescribed format like email or XML data. Unstructured data does not have any prescribed data model though they have an internal data structure and can be in textual and non-textual format usually for big data. Processing such enormous amounts of data is a difficult task. Deep learning is gaining growing popularity as complex processing is done via hidden layers which speeds up the automization process. All different time-consuming phases of the sentiment analysis process like feature selection, feature extraction, learning parameters, processing feature vectors and generating predictions can be automated and accelerated using deep learning.

In deep learning-based approaches, hidden layers are used to perform complex processing between input and output layers and act as a black box so data representations in the middle-hidden layers often remain unidentified. In this survey, we focus on different fusion techniques for fusing multiple modalities like visual, auditory and textual features for Multimodal sentiment analysis. The most used and recent advancements in dataset generation for multimodal sentiment analysis are summarized.Popular datasets are described along with their stages for creation. The most up-to-date fusion techniques from current research are discussed. Different concepts for results improvisation employed by all these methods are also discussed. Various multimodal categorization algorithms based on latest machine learning and deep learning innovations are studied. The usage of sentiment analysis in diverse application domains and future scope in various applications are outlined. Utilisation of heterogeneous data like videos, psychological signals, EEG signals and other information are also discussed, along with multi-lingual data, cross-language data, cross-domain data and code-mixed data formats.

The rest of the sections of this paper are organized as follows. Section 2 explains the fundamentals and need for MSA. In Section 3, some of the most prominent datasets for MSA are summarized. Section 4 presents a comprehensive survey of different fusion architectures utilising latest innovations. MSA applications in various domains are defined in Section 5. Section 6 describes limitations of each fusion architecture along with model wise merits and demerits. Section 7 discusses future directions in MSA research, followed by concluding remarks in Section 8.

2. Multimodal sentiment analysis fundamentals

In classic sentiment analysis systems, just one modality is inferred to determine user's positive or negative view about subject. Multimodal sentiment analysis is a subset of traditional text-based sentiment analysis that includes other modalities such as speech and visual features along with the text. Visual features are used because a visual depiction can explain or describe something more effectively than a long list of written or spoken words and may could be highly utilized to rightly predict the associated sentiment with data. For multimodal sentiment analysis a variety of two-modality combinations such as speech+image, image+text, speech+text, speech+EEG signals can be used. System using two modalities are called bimodal sentiment analysis system. System using all three modalities are called trimodal sentiment analysis systems.

Modalities are available in a wide range of shapes and sizes like audio, or uttered words (e.g., laughs, cries, tones), temporal images (e.g., smile, gaze, facial expressions), and textual data (e.g., transcribed data from spoken words) or in the form of EEG signals. These are the most commonly investigated modalities in multimodal sentiment analysis. Only a subset of the features listed below are used for optimal categorization. The text is made up of positive or negative words which is called polar words that clearly defines the polarity e.g., word 'amazing' shows positive polarity while word 'terrible' indicates negative polarity. The text is divided into word groups or phrases, character N-Grams, phoneme N-Grams, and other textual elements. Auditory features include laughs, pauses, tones, cries and voice pitch distribution over a sentence, utterance speed, and so on. Smiles and frowns, gestures, posture, gazes, and other visual elements such as eye contact are all examples of facial expressions [7].

In text indicators, words are considered for sentiment prediction. But only a few words in a sentence are important to know the sentiment, rest of the words are either stop words which are used to grammatically form the sentence. Character n-grams which is groups of characters appear together in comparable emotive terms are considered for sentiment analysis. Phoneme n-grams are also considered which are similar to character n-grams, but with the addition of phonemes.

In Auditory indicators, audio data is very much important to generate the transcripts. Some of the important auditory features are pauses. For example, larger number of pauses in an utterance indicates neutral sentiment. Pitch is also used to reveal subjectivity. A high pitch conveys anxiousness or enthusiasm whereas a low pitch conveys seriousness. Another important auditory feature is intensity or energy with which words are spoken. The polarity of high-energy utterances is typically demonstrated by a focus on a single word or phrase.

In Visual indicators, visual sentiments are used along with text data to analyse the sentiments. Facial expression plays a crucial role in identification of sentiments. Smile is the obvious feature which indicates positive sentiments. Smiling is a positive emotion. Due to recent advancements, cameras can now detect smiles and even can detect smile intensity and can assign a score to it. Same way gaze, eye contact denotes positive sentiment whereas if person is looking away represents negative or neutral sentiment.

As a result, the speaker's face orientation can be used to determine eye contact or gaze.

Fig. 1 depicts the different phases of process for sentiment analysis that uses multimodal fusion by combining different audio-visual features from heterogeneous sources. First the varieties of data like structured, semi structured or unstructured from internet or web are taken as input and then data is pre-processed. In pre-processing phase, data is cleaned and selected using dimensionality reduction as per problem requirement. Then features are extracted using various feature extraction algorithms. Audio features are extracted from spoken words in video and temporal images are extracted from video. Transcripts for the text data is generated from speech in video. Next from extracted features multimodal feature vector is generated and using various calcification algorithm data is classified. Then classified result is sent to particular application [7].

2.1. Importance of modalities

In multimodal sentiment analysis, different modalities are used for finding affective states from the conversation. Most widely used modalities are text, audio and visual modality. Each contributes for better prediction of sentiments and the literature says that bimodal and trimodal system improves result as compared to unimodal system. Each modality having some important contribution to improve the accuracy.

Text Modality: Text modality is dominant amongst all the modalities. It plays a key role in identifying the hidden sentiments. Textual sentiment analysis is generating very good results but now a days most of the opinionated data are shared in the form of videos rather than text.

Visual Modality: Visual features helps in better identification of underlying sentiments or opinions. For example, if there is a text "this is pretty good mouse". Using only textual data it is difficult to identify if this is about original animal mouse or computer mouse. Visuals helps in this scenario and combination of text and visuals which makes bimodal system and generates better results as compared to unimodal systems.

Audio Modality: Acoustic features are used to generate the textual data from videos and as well as the tone of the speakers can be identified. Combination of all the three modalities generates a better analysis model. In the case of humour, sarcasm and common-sense detection visuals may be do wrong predication but combination of modalities can correctly identify the sentiments.

3. Popular datasets

The following are the stages involved in creating a dataset for multimodal sentiment analysis.

Data Acquisition: Data acquisition is made up of two words: data and acquisition. Data refers to raw facts and numbers that can be structured or unstructured. Acquisition refers to gathering data for a specific goal. The term 'data acquisition' refers to the act of gathering data from relevant sources before it is stored, cleansed, pre-processed, and used in other methods. For this purpose, videos are collected from several internet video sharing platforms for the creation of multimodal datasets. To crewel the movies from the web using specified search phrases, automatic or semi-automatic tools are utilised. To confirm that the video is a monologue, web videos are analysed for the presence of one speaker in the frame using facial detection. 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 have missing data, incorrect or spurious values, and may not contain relevant, specific properties. Data pre-processing is necessary to increase the data's quality. A broad age range from the mid-20 s to the late-50 s is envisaged for MSA dataset generation. Majority of speakers in dataset are being English natives from the United Kingdom or the United States of America. A small percentage are non-native but fluent English speakers are included in the dataset. Glasses are worn by some of the speakers. There aren't many videos with speakers that have a dialect or an accent.

Data Post Processing: It is well known that linguistic information from spoken language can aid in the acquisition of emotions and is a crucial component of context interpretation. The audio data is automatically transcribed to facilitate the interaction of the auditory, visual, and text modes. Both the Google Cloud Speech API and Amazon Transcribe transcriptions of the videos are of appropriate quality for this task. Nonverbal cues and aural aspects, such as laughing, music, and theme are included. The spoken language transcriptions include punctuation (e.g., period, question mark, exclamation mark and each transcribed word has a start and finish timestamp as well as duration. These metadata aid in the alignment of the text with annotations (sample rate differences) and other modalities.

Data Annotation: Data annotation is the process of labelling images, video frames, audio, and text data that is mainly used in supervised machine learning and semi supervised machine learning to train the datasets. It helps machine to understand the input and act accordingly. For each clip, every annotator decides its sentimental state as -1 (negative), 0 (neutral) or 1 (positive). There are independent human resources who are used for making annotations. Then, in order to do both regression and multi-classifications tasks, average labelled result is used.

The most used and recent advancements in dataset generation for multimodal sentiment analysis are summarized in Table 1. The first column displays the name of dataset, followed by the year it was published in the second column. Then third column shows the reference to research by which dataset is introduced while various modalities

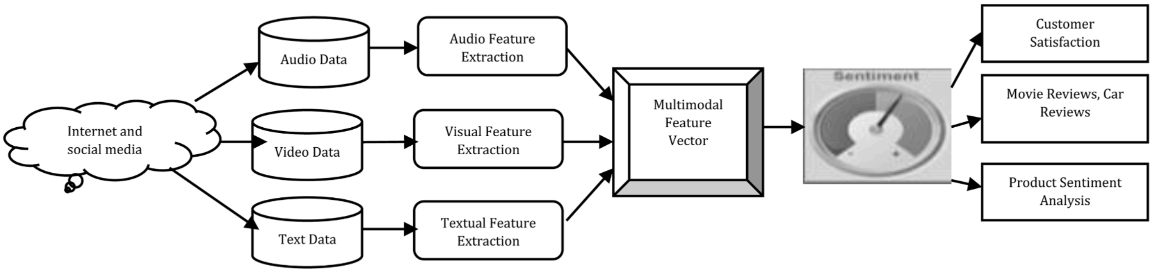
Internet and social media
Audio Data
Video Data
Text Data
Audio Feature Extraction
Visual Feature Extraction
Textual Feature Extraction
Multimodal Feature Vector
Sentiment
Customer Satisfaction
Movie Reviews, Car Reviews
Product Sentiment Analysis

Fig 1. MSA process model [7].

Table 1
Summary of popular datasets for multimodal sentiment analysis.

Name	Year	Reference	Modalities	No. of videos	Source	No. of speakers	Language	Topics Covered	Available At
MOSI	2016	[10]	$A+V+T$	93	YouTube	89 41-female 48-male		General Indexed by #vlog	*https://www.amir-zadeh.com/datasets*
CMU-MOSEI	2018	[11]	$A+V+T$	3228	YouTube	1000	English	250 Reviews Debate Consulting	*https://www.amir-zadeh.com/datasets*
MELD	2019	[90]	$A+V+T$	–	TV Series -Friends	Multi Speaker	English	Dialogues from TV series-Friends	*https://affective-meld.github.io.*
Memotion Analysis	2020	[12]	$V+T$	–	Reddit, Facebook, etc.	–	English	52 Hillary Trump Minion	*https://github.com/terenceylchow124/Meme-MultiModal/tree/main/data/memotion*
CH-SIMS	2020	[13]	$A+V+T$	2281			Chinese	General	*https://github.com/thuiar/MMSA.*
CMU-MOSEAS	2021	[14]	$A+V+T$	4000	YouTube	1645	Spanish Portuguese German French	General 250 topics	–
MuSe-CaR	2021	[15]	$A+V+T$	291	YouTube	70	English	Vehicle Review	*https://www.muse-challenge.org/challenge/data*
B-T4SA	2021	[16]	$V+T$	470k tweets	Tweeter	–	Other than English	General	*https://www.t4sa.it/*
FACTIFY	2022	[17]	$V+T$	50,000 tweets	Tweeter	–	English	20 Politics Governance	–
MEMOTION 2	2022	[18]	$V+T$	10,000 images	Reddit, Facebok, etc.	–	English	Politics Religion Sports	–

covered in each dataset is shown in fourth column. Fifth column shows the number of review videos are used in dataset while sixth column shows source platform of videos from where the videos are collected. The number of speakers with male and female speakers in the dataset can be seen in the seventh column followed the language spoken in videos in eighth column, while the ninth column lists the many themes covered in videos, such as movie reviews and product reviews. The tenth column displays a link to a dataset that is available in the public domain.

The following two datasets were heavily used for research purposes before the more recent datasets described in Table 1. The first is the YouTube Opinion Dataset, which was generated by [8]. It is a sentiment dataset for multimodal analysis. It features 47 videos from YouTube that are not tied to any certain theme. There are 27 male speakers and 20 female speakers. The dataset includes text that has been manually transcribed as well as audio and visual elements that have been automatically extracted. It also has utterances that have been automatically extracted. And the second is Spanish Multimodal Opinion Dataset created by [9] is a multimodal sentiment analysis dataset in Spanish language. It includes 105 videos that have been annotated for sentiment polarity at the utterance level. Long gaps are used to automatically extract utterances, with most films having 6–8 utterances. In total, there are 550 utterances in the dataset. There are no sentiment intensity annotations in any of the suggested datasets. They mainly focus on polarity. Also, as indicated in the introduction, they tend to focus on video or utterance analysis rather than fine-grained sentiment analysis.

3.1. MOSI dataset

MOSI is created by [10]. MOSI stands for Multimodal Opinion Level Sentiment Intensity. It includes the following features like multimodal observations, transcript generated from spoken words and visual gestures, automatic audio and visual features. It also includes subjectivity segmentation at opinion-level, sentiment intensity annotations with high coder agreement and alignment between words, visual, and acoustic features. It's made up of 93 YouTube videos grouped together under the hashtag #vlog. It has 89 different speakers, 41 of whom are female and 48 of them are male. It's the first multimodal sentiment analysis dataset to include subjectivity and sentiment intensity annotations at the opinion level.

3.2. CMU-MOSEI dataset

CMU Multimodal Opinion Sentiment and Emotion Intensity (**CMU-MOSEI**) is the largest dataset of sentence level sentiment analysis and emotion recognition in online videos [11] .On CMU-MOSEI, you may find over 65 h of annotated video from over 1000 speakers and 250 subjects. Each video segment includes a phoneme-level manual transcription that is synchronised with the audio. The videos were taken from YouTube, an online video sharing service. According to statistics The number of videos you can get from any YouTube channel is limited to ten. According to statistics, it has a total of 23,453 sentences from 3228 videos,. The dataset covers a wide range of topics, but the top three are reviews for different product and services (16.2%), debate on various topics (2.9%), and consulting (2.9%). The remaining topics are nearly evenly spread over the dataset.

3.3. MELD dataset

It is Multimodal EmotionLines Dataset which is the extension of EmotionLines dataset [90]. It is multimodal multi-party conversational emotion recognition dataset. It is most widely used dataset for emotion recognition. MELD includes the same dialogue instances as EmotionLines, but it also includes audio and visual elements in addition to text. MELD contains around 1400 dialogues and 13,000 utterances from the television series named Friends. It is multiparty dataset includes multiple speakers. Each utterance is annotated with emotion and sentiment labels, and encompasses audio, visual, and textual modalities. Any of these seven emotions has been assigned to each syllable in a dialogue: Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear. Each speech in MELD additionally has a sentiment annotation (positive, negative, or neutral). Its main purposes are to train contextual modelling in a conversation for emotion recognition.

3.4. Memotion analysis dataset

It is multimodal dataset for online meme sentiment analysis. Memes are usually humorous and attempt to be relatable. Many of them hope to connect with their audience by expressing solidarity during different stages of life. Some memes are purely amusing while others take a critical stab at current events. For collecting meme data, a total of 52 distinct and worldwide popular categories are evaluated such as Hillary, Trump, Minions, Baby godfather, and so on. Google's image search service was used to obtain the meme photos. Amazon Mechanical Turk (AMT) workers is used to annotate the emotion class labelled as humorous, sarcasm, offensive, motivation and quantify the intensity with which a particular effect of a class is communicated. Overall feelings in the dataset spans across four classes (very negative, negative, neutral, positive, very positive) and collected from approximately 10k samples [12].

3.5. CH-SIMS dataset

CH-SIMS is a Chinese single and multimodal sentiment analysis dataset with independent unimodal annotations that contains 2281 refined video segments in the wild. It has both multimodal and independent unimodal annotations. It enables researchers to investigate the relationship between modalities or conduct unimodal sentiment analysis using independent unimodal annotations. It contains 2281 video segments from 60 raw videos. SIMS boasts a diverse cast of characters, a wide age range, and excellent production values. [13].

3.6. CMU-MOSEAS dataset

The CMU-MOSEAS (CMU Multimodal Opinion Sentiment, Emotions and Attributes) dataset includes four languages: Spanish (over 500 million total speakers worldwide), Portuguese (over 200 million speakers worldwide), German (over 200 million speakers worldwide), and French (over 200 million speakers worldwide). These languages are Roman or German in origin. They are from Europe which is also where we get the majority of our video. The languages are also spoken in parts of Africa and the Caribbean as well as in the north and south of the American continent with different dialects. However, the European dialect is mostly comprehensible across different regions with some exceptions. With 40,000 total samples encompassing 1645 speakers and 250 themes, the CMUMOSEAS dataset is the largest of its sort in all four core languages (French, German, Portuguese, and Spanish). Sentiment and subjectivity, emotions, and personality traits are amongst the 20 annotated labels in CMUMOSEAS. As multimodal learning progresses, the dataset and accompanying descriptors is made publicly available and new feature descriptors will be added on a regular basis [14].

3.7. MuSe-CaR dataset

The MuSe-CAR database is a big, multimodal (video, audio, and text) dataset that was obtained in the field with the goal of learning more in the wild. Multimodal Sentiment Analysis such as the emotional involvement that occurs during product reviews (e.g., vehicle reviews) where a sentiment is linked to a topic or entity. Professional, semi-professional (influencers), and casual reviewers are to be in their late 20 s to late 50 s. The majority are native English speakers from the United Kingdom or the United States. It also has a small number of non-native yet fluent English speakers. MuSe-CaR dataset was acquired in the wild with the goal of creating appropriate methods and learning more about Multimodal Sentiment Analysis. MuSe-CaR was built with a variety of computational tasks in mind, including emotion and entity recognition, and primarily to improve machine comprehension of how sentiment (i.e., emotion) is linked to an entity and components of such reviews. It has a high-quality subset of the MuSe-CaR dataset for the MuSe 2020 Challenge, consisting of 36 h 52 min 08 s of video data from 291 movies and 70 host speakers (plus an extra 20 narrators) collected from YouTube. The subject of MuSe-CaR videos is limited to vehicle reviews. With the number of vehicle manufacturers limited to premium brands (BMW, Audi, Mercedes-Benz) that equip their vehicles with the most up-to-date technology, ensuring that discussed entities and aspects (e.g., semi-autonomous vehicle functions) appear across a wide range of videos (and different manufacturers). The majority of the evaluations are written by semi-professional or professional reviewers (such as YouTube's "influencers"). All of the YouTube channels that were used in MuSeCaR gave their clear consent for their data to be used in academic research [15].

3.8. B-T4SA dataset

B-T4SA is a subset of T4SA that has 470 thousand samples. Each of the samples contains both text and image data. The train set accounts for roughly 80% of the dataset while the validation and test set each account for 10%. B-T4SA was designed to address T4SA's issues with duplicated entries, tiny phrases, malformed graphics, and unbalanced classes [16].

3.9. FACTIFY dataset

It is multimodal fact checking dataset [17]. It is the world's largest multimodal fact-checking public dataset. It has 50 K data points covering news from India and the United States. FACTIFY is a collection of photographs, textual assertions, reference textual sources and images. It is categorised into three categories: support, no evidence, and refute. It is based on accessibility, popularity, and posts per day, and it is gathered date-wise from twitter handles of Indian and US news sources: Hindustan Times 1, ANI2 for India and ABC3, CNN 4 for the US. Furthermore, these Twitter accounts are notable for their impartial and unbiased stance. The tweet text and image were retrieved from each tweet. There are 50,000 samples in total in the dataset, with equal numbers in each of the five categories. The Train-Val-Test split for the dataset is 70:15:15. The majority of the claims in the dataset are related to politics and government. Political parties and leaders are mentioned in claims from both the United States and India, as indicated by the top 20 most frequent entities.

3.10. MEMOTION 2 dataset

This is the first large-scale multimodal dataset for meme classification [18]. Humans express a wide range of emotions, including fury, hatred, grief, tranquilly, fear, and so on with varying degrees of intensity. Memotion 2.0 is a dataset that focuses on categorising emotions and their intensities into discrete labels. It also features labels that correspond to a meme's sentiment. Memotion 2.0 expands on the previous version (Memotion 1.0) by including a new set of 10,000 memes collected from various social media sites. It is collected from a variety of sources. Memes are manually downloaded after narrowing down numerous themes of interest, such as politics, religion, and sports. The dataset comprises of 10,000 images separated into a train-val-test split with images ranging from 8500 to 1500 to 1500. Overall Sentiment (positive, neutral, negative), Emotion (humour, sarcasm, offence, motivation), and Scale of Emotion are all annotated for each meme (0–4 levels).

4. MSA fusion techniques

Fusion of different modalities is at the centre of Sentiment Analysis using multiple modalities. Multimodal fusion is process of filtering, extracting and combining required features from data received from a variety of sources. These data are then analysed further in order to extract opinions and assess attitudes. Table 2 lists many fusion procedures as well as their explanations. The early studies [87–89] show that employing any combination of three modalities, a bimodal system and

Table 2
Performance summary of various multimodal fusion variants with its architectural categories.

Ref.	Model Name	Year	Fusion Type	Method	Dataset	Main Objective	Accuracy/ F1 Score
[19]	Tri-modal HMM	2011	Early Fusion	HMM	YouTube	Introduces trimodal sentiment analysis Five multimodal features were identified: polarised words, smiles, gazes, pauses, and voice tone. A new YouTube dataset has been introduced.	55.3%
[20]	Text-audio-Visual	2013	Early Fusion	SVM	Spanish Multimodal Opinion Dataset	The usage of multiple modalities together boosts performance. A new Spanish Multimodal Opinion Dataset has been introduced. Check the portability of model on English dataset	75%
[21]	Proposed Multi Method	2015	Early Fusion	SVM	eNTERFACE	Novel multimodal information extraction agent	87.95%
[22]	Multimodal Model	2016	Early Fusion	SVM	POM	Persuasiveness prediction in Online Social Multimedia	70.85%
[23]	MARN	2018	Early Fusion	LSTHM	ICT-MMMO	Novel architecture to understand Human Communication Comprehension	86.3%
[24]	Audio-Visual	2013	Late Fusion	HMM, CERT	AVEC	Classifier Fusion using Kalman Filter	68.5%
[25]	Ref	2014	Late Fusion	SMO (Sequential Minimal Optimization)	YouTube	Measuring personality trait using behavioural signal processing Automatic recognition of personality trait using Vlogs Model's performance also compared with the emotional feature set which is poor	62.6%
[26]	Multi CNN	2015	Late Fusion	CNN	Real World Twitter Dataset TD1 and TD2	Sentiment Analysis of Multimodal tweets using CNN	79%
[27]	ICT-MMMO	2013	Hybrid Fusion	BLSTM, SVM	ICT-MMMO	Multimodal sentiment analysis in online review videos Hybrid Fusion by combining Early and Late Fusion	71.3%
[28]	Bimodal with Unimodal	2015	Hybrid Fusion	Deep CNN, MKL		Parallelizable decision-level data fusion method Multiple kernel learning for training classifier	86.27%
[29]	Three modalities	2016	Hybrid Fusion	ELM	YouTube	Explanation of feature extraction as well as model building Comparison using various machine learning methods	80%
[30]	HMM-BLSTM	2012	Model-Level Fusion/ Utterance-Level	HMM, BLSTM	IEMOCAP	At the utterance level flow of affective expressions Context-sensitive approaches for emotion recognition within a multimodal, hierarchical approach indicate potentially relevant patterns	72.35
[31]	Context-aware BLSTM	2013	Model-Level Fusion	BLSTM	SEMAINE	Context-sensitive LSTM-based audio-visual emotion recognition	65.2%
[32]	TFN	2017	Tensor Fusion	TFN	CMU-MOSI	End-to-end learning of intra-modality and inter-modality dynamics	77.1%
[33]	MRRF	2019	Tensor Fusion	MRFF	CMU-MOSI	Modality-based Redundancy Reduction Fusion (MRRF) modality-based tensor factorization Modality-based Redundancy Reduction Fusion in MRRF is a tensor fusion and factorization method that allows for modality-specific compression rates while also reducing parameter complexity.	77.46%
[34]	T2FN	2019	Tensor Fusion	T2FN	CMU-MOSI	Regularization method based on tensor rank minimization Random drop settings	–
[35]	MTFN—CMM	2021	Tensor Fusion	MTFN—CMM	CMU-MOSI CMU-MOSEI	Emotional fusion in multimodal data is effective Cross-modal modelling with a multi-tensor fusion network	80.9%
[36]	CHFusion	2018	Hierarchical Feature Fusion	CNN	CMU-MOSI	Hierarchical Fusion, in which the two modalities are fused first, followed by the three modalities. Context-Aware Hierarchical Fusion (CHFusion)	80%
[37]	HFNN	2019	Hierarchical Feature Fusion	BiLSTM	CMU-MOSI	'Divide, conquer, and combine' is a strategy used for multimodal fusion. For a thorough interpretation of multimodal embeddings, fusion have been done hierarchically so that both local and global interactions are taken into account. Global fusion is used to obtain an overall view of multimodal embeddings via a specifically designed ABS-LSTM. Two levels of attention mechanism are used: Regional Interdependence Attention and Global Interaction Attention	80.19%
[38]	BBFN	2021	Bimodal Fusion		CMU-MOSEI	An innovative end-to-end Bimodal Fusion network that conducts fusion (relevance increment) and separation (difference increment) on pairs modality representations is introduced. Modality-specific feature space separator and gated	86.2%

(continued on next page)

Table 2 (*continued*)

Ref.	Model Name	Year	Fusion Type	Method	Dataset	Main Objective	Accuracy/ F1 Score
						control mechanism is used	
[39]	BIMHA	2022	Bimodal Fusion		CH-SIMS	Bimodal Information-augmented Multi-Head Attention Inter-modal interaction and inter-bimodal interaction	82.71%
[40]	CATF-LSTM	2017	Attention Based Fusion	CATF-LSTM	CMU-MOSI	Attention-based fusion mechanism, termed AT-Fusion	81.30%
[41]	MMHA	2020	Attention Based Fusion	MMHA	MOUD, CMU-MOSI	Learn how unimodal features are related, and capture the internal structure of unimodal features Learn the connections between the various modalities, and focus on the contributing elements.	82.71%
[42]	Bi-LSTM with attention model	2021	Attention Based Fusion	Bi-LSTM	CMU-MOSI	Before fusion, a unique attention-based multimodal contextual fusion technique extracts contextual information from the utterances	80.18%
[43]	QMR	2018	Quantum Based Fusion		Flicker GI	Quantum-inspired Multimodal Sentiment Analysis fill the 'semantic gap' and model the correlations between different modalities via density matrix Quantum Interference inspired Multimodal decision Fusion (QIMF)	88.24%
[44]	QMN	2020	Quantum Based Fusion	QT+LSTM	MELD	Quantum-like multimodal network (QMN), which combines quantum theory (QT) mathematical framework with a long short-term memory (LSTM) network Dynamics of intra- and inter-utterance interaction modelling choice correlations between multiple modalities using a quantum interference-inspired decision fusion approach. To create better predictions about social impact amongst speakers, a quantum measurement-inspired strong-weak influence model was developed.	75.60%
[45]	QMF	2020	Quantum Based Fusion		CMU-MOSEI	The word interaction within a single modality and the interaction across modalities are formulated with superposition and entanglement respectively at different stages. Quantum-theoretic multimodal fusion framework	79.74%
[46]	MFN	2018	Word-level Fusion	MFN	MOUD	MFN shows a consistent trend for both classification and regression	80.4%
[47]	DFG	2018	Word-level Fusion	MFN+DFG	CMU-MOSEI	Multi-view sequential learning that consists of three main components System of LSTMs consists of multiple Longshort Term Memory (LSTM) networks, Delta-memory Attention Network	85.2%
[48]	RMFN	2018	Word-level Fusion	LSTHM	CMU-MOSI	Gated Modality mixing Network Multimodal Shifting	85.4%
[49]	RAVEN	2019	Word-level Fusion	RAVEN	CMU-MOSI	Recurrent Attended Variation Embedding Network (RAVEN) Nonverbal Sub-networks Gated Modality mixing Network Multimodal Shifting	78.0%

trimodal system have superior accuracy than a unimodal system.

Data fusion, feature fusion, and decision fusion are three methods for combining or fusing data. The majority of the work employs decision fusion. A joint feature vector is formed by combining independent input features in feature level fusion. Feature level fusion mixes prosodic and face expression elements. Multimodal sentiment analysis is a relatively recent field of research. Combining input modes improves the precision of the analysis. A generic feature vector is created by combining features from several modalities, such as text, audio, and visual elements. This vector is then categorised after being normalised on the same scale. After that, the produced combined features vector is transmitted to be analysed. These basic algorithms are used to extract redundant information from several modalities such as audio, video, and text as well as contextual information between video utterances. Fig. 2 summarizes different state-of-the art models for all fusion categories.

4.1. *Early fusion-feature level fusion*

Feature-level fusion (sometimes called early fusion) combines all of the features from each modality (text, audio, or visual) into a single feature vector. It is then inputted to a classification algorithm. The benefit of feature-level fusion is that it allows for early correlation between distinct multimodal features which could lead to better task completion. The integration of disparate elements is one of the challenges in applying this strategy. The downside of this fusion method is time synchronisation as the characteristics collected belongs to several modalities and can differ greatly in many areas. Therefore, the features need to be converted into desired format before the fusion process can take place. Modality fusion at the feature level poses the challenge of integrating widely dissimilar input features. It implies that synchronising various inputs while retraining the modality's classification system is a difficult process. How to create acceptable joint feature vectors made of characteristics from distinct modalities with varying time scales, metric levels, and temporal structures is an unresolved question.

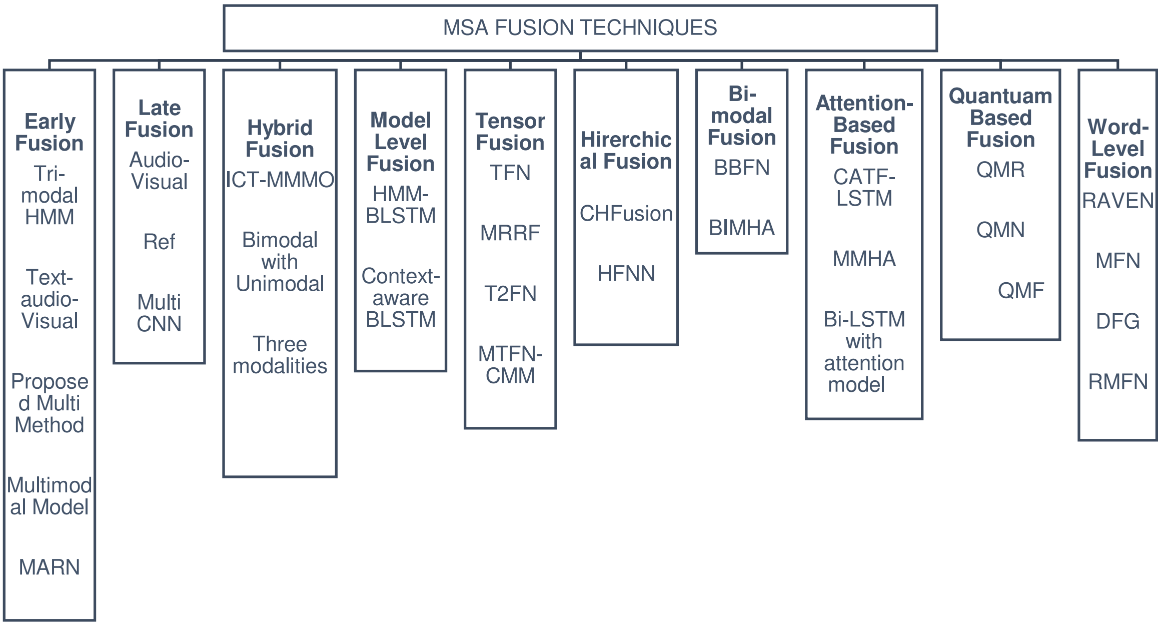

Fig. 2. Multimodal fusion models for multimodal sentiment analysis.

Concatenating audio and video features into a single feature vector as are done in existing human effect analysers that use feature level data fusion is clearly not the answer. The intra-modality dynamics cannot be effectively represented using this fusion approach. It is unable to filter out conflicting or redundant data gathered from several modalities. An early fusion architecture is represented in Fig. 3.

4.2. Late fusion-decision level fusion

In late fusion, first the features of each modality are independently processed and classified. Then classification results are fused to form a final decision vector which subsequently yields the sentiment prediction. Because fusing occurs after classification, this procedure is known as late fusion. Due to early fusion challenges, most researchers choose for decision-level fusion in which each modality's input is modelled separately and the results of single-modal recognition are integrated at the end. In the disciplines of machine learning and pattern recognition, decision-level fusion which is also known as classifier fusion is now a hot topic. Many studies have proved the superiority of classifier fusion over separate classifiers due of the uncorrelated mistakes from different classifiers. Because the decisions arising from many modalities usually have the same form of data, fusion of decisions received from various modalities is easier than feature-level fusion. Another benefit of this fusion process is that each modality can learn its characteristics using the best classifier or model available. The learning procedure of all these classifiers at the decision-level fusion step becomes difficult and time consuming when different classifiers are required for the analysis task. Fig. 4 represents late fusion architecture.

4.3. Hybrid fusion

It's a hybrid of early fusion and late fusion techniques. This fusion method incorporates both feature-level and decision-level fusion techniques. Researchers use hybrid fusion to take use of the benefits of both feature and decision-level fusion procedures while avoiding the drawbacks of each.

4.4. Model-level fusion

The features of distinct modalities are studied to see if there is a link between them. The desired Model is then created based on the study domain and problem need. It's a technique that combines data from several modalities and uses correlation to create a relaxed fusion. Researchers created models that met their study needs while also taking into account the problem space.

4.5. Tensor fusion

This approach builds a 3-fold Cartesian product employing modality embeddings using a tensor fusion layer that explicitly mimics unimodal, bimodal, and trimodal interactions. It minimises the number of training samples required. The architecture of one of the tensor fusion techniques i.e., MTFN can be seen in Fig. 5.

4.6. Hierarchical fusion

It is a unique feature fusion approach that works in a hierarchical

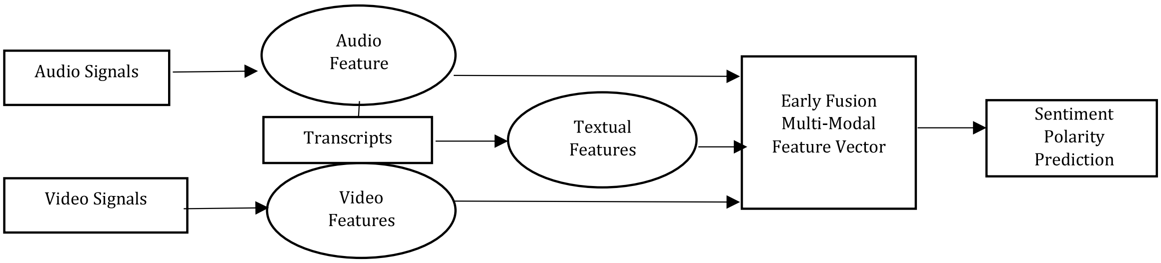

Fig. 3. Early Fusion for Multimodal Sentiment Analysis.

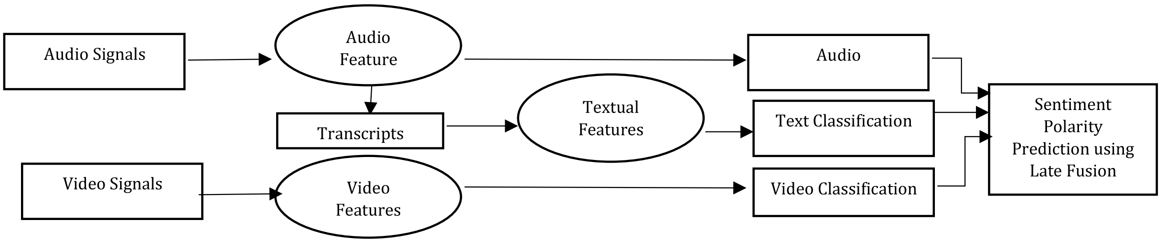

Fig. 4. Late fusion for multimodal sentiment analysis.

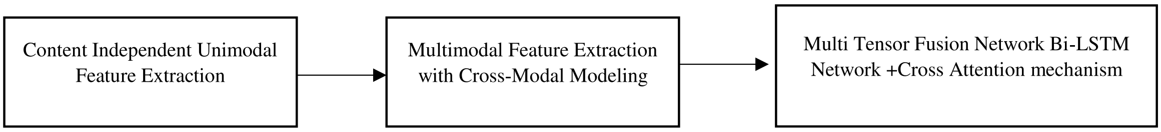
Content Independent Unimodal Feature Extraction
Multimodal Feature Extraction with Cross-Modal Modeling
Multi Tensor Fusion Network Bi-LSTM Network +Cross Attention mechanism

Fig. 5. MTFN architecture for tensor fusion.

order, merging the modalities two in two first, and then all three modalities last. The difficulty with this unsophisticated method is that it is incapable of filtering out contradictory or redundant data gathered from other modalities. We develop a hierarchical technique to handle this major issue, progressing from unimodal to bimodal vectors, then bimodal to trimodal vectors. In fusion of two modes, for each bimodal combination, such as $T + V$, $T + A$, and $A + V$, it fuses the utterance feature vectors. It uses three bimodal features to generate a trimodal feature. One of the hierarchical fusions i.e., HFFN model architecture is represented in Fig. 6.

4.7. Bimodal fusion

In bimodal fusion, on pairwise modality representations, a novel end-to-end network achieves fusion (relevance increment) and separation (difference increment). The two components are trained at the same time so that they can fight each other in a simulated battle. Due to the known information imbalance amongst modalities, the model takes two bimodal pairings as input. One of the bimodal fusion architectures i.e., BBFN architecture model is shown in Fig. 7.

4.8. Attention mechanism-based fusion

Contextual information extraction and multimodal fusion are the two most important difficulties in multimodal sentiment analysis and emotion identification. Multi-level contextual feature extraction using a bidirectional recurrent neural network-based model is called attention mechanism-based fusion. At the utterance level, each modality contributes differently to sentiment and emotion classification. As a result, the model suggests attention-based inter-modality fusion for multimodal fusion to accommodate the importance of each inter-modal utterance. Contextual attentive unimodal features are joined two by two to form bimodal features, which are then merged all together to form trimodal feature vectors. Contextual features are extracted after each step of fusion. One of the attentions mechanism-based fusion i.e., MMHA architecture is depicted in Fig. 8.

4.9. Quantum based fusion

Quantum based fusion uses quantum interference and quantum measurement theory. Interactions inside each utterance (i.e., correlations between distinct modalities) are captured using quantum interference and a strong-weak influence model to detect interactions between consecutive utterances in this method (i.e., how one speaker influences another) is developed using quantum measurement. It also makes use of the decision-level or late fusion approach. One of the quantum-based architecture i.e., QMN architecture is depicted in Fig. 9.

4.10. Word level fusion

In this strategy, interaction between multiple modalities is examined in order to obtain a superior sentiment tendency. Transformer is used to learn joint representations for utterances and to translate across different modalities. The Memory Fusion Network (MFN) is a recurrent model for multi-view sequential learning that is made up of three parts: (1) Long Short-Term Memory (LSTM) networks, which encode the dynamics and interactions unique to each view. (2) The Delta-memory Attention Network is a particular attention mechanism in the System of LSTMs that is meant to find both crossview and temporal relationships across different dimensions of memories. (3) Multi-view Gated Memory (MVGM) is a unified memory that records cross-view interactions across time. The RMFN pipeline and its components are depicted in Fig. 10. MFN takes as input a multi-view sequence consisting of N views of length T each.

4.11. Latest variable multimodal sentiment analysis models

Around 177 articles in the last five to seven years have demonstrated

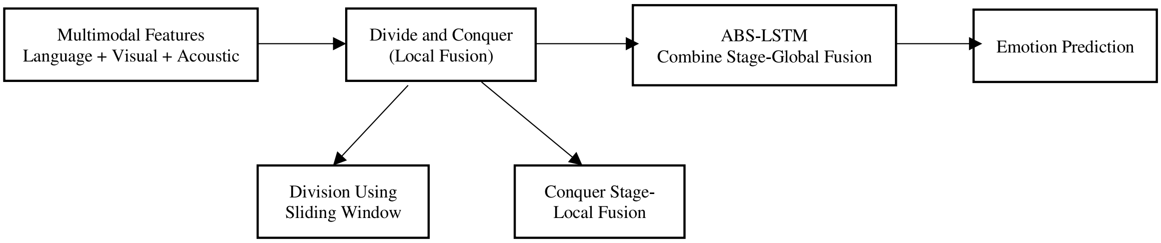
Multimodal Features Language + Visual + Acoustic
Divide and Conquer (Local Fusion)
ABS-LSTM Combine Stage-Global Fusion
Emotion Prediction
Division Using Sliding Window
Conquer Stage-Local Fusion

Fig 6. HFFN architecture for hierarchical fusions.

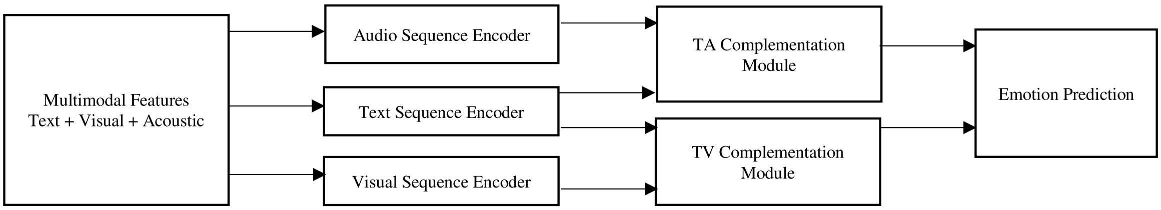

Fig. 7. BBFN architecture for bimodal fusion.

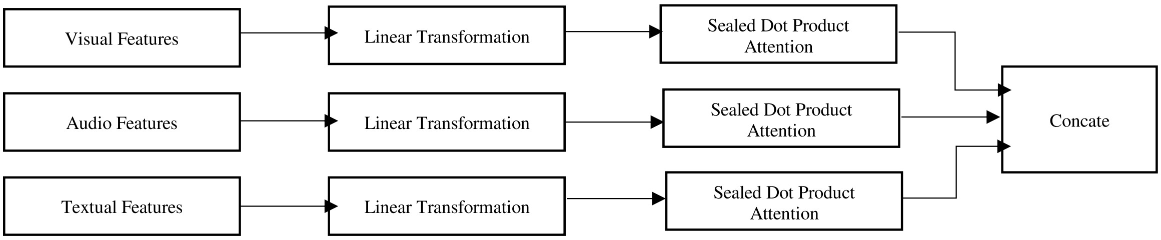

Fig. 8. MMHA architecture for attention based fusions.

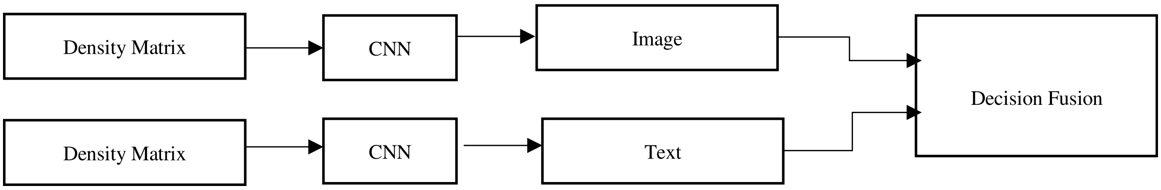

Fig. 9. QMN architecture for quantum based fusion.

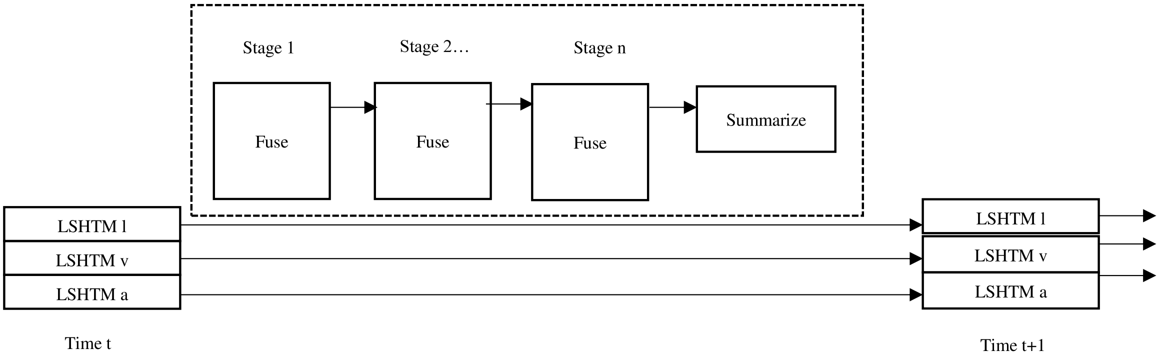

Fig. 10. RMFN architecture for word-level fusion.

an architectural shift in the Multimodal Sentiment Analysis framework, according to dblp (computer science bibliography). The scope and depth of research in this field has grown to the point where displaying the entire model has become impossible. There were around 36 papers in 2020, nearly 67 papers in 2021, and 13 papers in 2022 till May 2022. This paper presents some of the MSA models that have been benchmarked.

4.11.1. GESentic (real-time GPU–ELM multimodal sentiment analysis) (2017)

Sentic memes (fundamental inputs of sentiments that can generate most human emotions) are used in an ensemble application of ELM and GPU for real-time multimodal sentiment analysis. To improve the performance of the feature extraction process from diverse modalities, the method leverages a variety of GPU-friendly approaches. In addition, the sentiment analysis model is built using sophisticated ELM classifiers based on the retrieved features [55].

4.11.2. CSFC (Convolutional Fuzzy Sentiment Classifier) (2019)

Convolutional Fuzzy Sentiment Classifier is a model that combines deep convolutional neural networks and fuzzy logic. The majority of phrases contain mixed emotions like sarcasm which can only be successfully described via fuzzy membership functions. Deep learning was used to extract information from each modality which were then projected into a single affective space that was clustered into different emotions. It utilises a fuzzy logic classifier to predict the degree of a certain emotion in affective space since people in the actual world have partial or mixed feelings regarding an opinion target [54].

4.11.3. HALCB (Hierarchal Attention-BiLSTM (Bidirectional Long-Short Term Memory) model based on Cognitive Brain limbic system (2020)

It is a cognitive neuroscience-inspired multi-modal fusion technique. HALCB divides multimodal sentiment analysis into two modules. Each of which is responsible for one of the tasks: binary classification or multi-classification. The former module recognises the polarity of the input items and splits them into two groups before sending them to the latter module separately. The Hash technique is used in this module to increase retrieval accuracy and speed. A positive sub-net is dedicated to positive inputs whereas a negative subnet is dedicated to negative inputs in the latter module. For matching its respective role, each of these binary modules and two subnets in the multi-classification module has a separate fusion approach and decision layer. At the final link, a random forest is additionally added to collect outputs from all modules and fuse them at the decision-level [53].

4.11.4. AHRM (Attention-based Heterogeneous Relational Model) (2020)

The Attention-based Heterogeneous Relational Model (AHRM) was used to classify multi-modal sentiments that included both content and social relationships. A progressive dual attention module is used in this method to gather image-text correlations before learning a joint image-text representation from the perspective of content information. A channel attention schema is proposed here to highlight semantically-rich image channels, and a region attention schema is proposed based on the attended channels to emphasise emotional areas. Then, in addition to learning high-quality representations of social images, a heterogeneous relation network is developed to aggregate content information from social situations, which extends the Graph Convolutional Network. It works with Flicker as well as the GI dataset. It is divided into four sections: (1) Single-modal Representation Learning, which learns single-modal picture representations using both visual and text perspectives. (2) Progressive Dual Image-Text Attention, which embeds image-text correlations into joint image-text representations via two innovative cross-modal attentions (channel attention and region attention). (3) Heterogeneous Relation Fusion, which creates a heterogeneous relation network from social relationships and extends GCN to collect content information from social settings as a complement to developing high-quality image representations. (4) Sentiment Prediction, which is ultimately responsible for sentiment classification [56].

4.11.5. SFNN: Semantic Features Fusion Neural Network (2020)

It's a semantic feature fusion neural network (SFNN). The model first obtains the effective emotional feature expressions of the image using convolutional neural networks and attention mechanisms. Then maps the emotional feature expressions to the semantic feature level. Finally, the emotional polarity of the comment is efficiently examined by merging the emotional features of the physical level of the image with the semantic features of the visual modal. The disparity between heterogeneous data can be reduced via feature fusion based on semantic level [60].

4.11.6. SWAFN: Sentimental Words Aware Fusion Network (2020)

It guides the learning of combined representation *of multimodal aspects by incorporating sentimental terms knowledge into the fusion* network. This approach is divided into two parts: shallow fusion and aggregation. To get the fused shallow representations, it applies a cross modal contention technique to obtain bidirectional context information from each two models. To support and guide the deep fusion of the three modalities and acquire the final sentimental words aware fusion representation, it builds a multitask of sentimental words classification for the aggregation component. CMU-MOSI, CMU-MOSEI and YouTube datasets are used for analysis [61].

4.11.7. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (2020)

MISA is a unique framework that divides each modality into two subspaces before fusing them to predict affective states. The first subspace is modality invariant. In this subspace, representations from different modalities learn to share commonalities and close the gap between them. All of an utterance's modalities are mapped to a common subspace with distributional alignment. Though multimodal signals come from various sources, they all share the speaker's motivations and aims, which are accountable for the utterance's overall affective state. The second subspace is modality-specific, which is unique to each modality and contains its distinguishing characteristics. As aligned projections on the shared subspace, the invariant mappings aid in capturing these underlying commonalities and linked features. It makes use of the MOSI and MOSEI datasets [91].

4.11.8. MAG-BERT (Multimodal Adaption Gate-Bidirectional Encoder Representations from Transformers) (2020)

This MAG-BERT model combines the core architectures of BERT and XLNet with a properly developed Multimodal Adaptation Gate (MAG). BERT and XLNet may now receive multimodal nonverbal data during fine-tuning due to MAG. It accomplishes so by causing BERT and XLNet to change to internal representation, which is dependant on the visual and audio modalities. MAG sentially translates the informative visual and audio components to a vector with a trajectory and magnitude using an attention conditioned on nonverbal behaviours. For multimodal sentiment analysis, it employed the CMUMOSI and CMU-MOSEI datasets. Finetuning MAGBERT and MAG-XLNet, as well as language-only finetuning of BERT and XLNet, greatly improves sentiment analysis performance over earlier baselines [93].

4.11.9. M2Lens (2021)

It's a new explanatory visual analytics tool that can help multimodal machine learning model developers and users to better understand and diagnose Multimodal Models for sentiment analysis. M2Lens analyses intra-modal and inter-modal interactions learned by a multimodal language model from the global subsets and local levels by taking feature importance into account using post-hoc explainability techniques. Furthermore, it allows for a multi-faceted investigation of multimodal elements and their effects on sentiment analysis model judgments. M2Lens is made up of four main perspectives. It illustrates the impact of three different forms of interactions (dominance, complement, and conflict) on model predictions. Furthermore, multimodal feature templates and visualisation glyphs make it easier to explore a group of frequently used and influential feature sets [51].

4.11.10. Persian multimodal sentiment analysis framework (2021)

It is an innovative Persian multimodal sentiment analysis framework for merging auditory, visual and textual elements in a contextually relevant way. It creates first multimodal dataset for Persian language utterances and sentiment polarity derived from YouTube videos. Automated feature extraction approaches such as convolutional neural network (CNN) and long short-term memory are used in this context-aware multimodal fusion framework to extract unimodal and multimodal characteristics (LSTM). Different features are combined to do multimodal sentiment analysis. The use of a context-aware multimodal strategy to overcome the limits of ambiguous words in Persian language is demonstrated [52].

4.11.11. TCM-LSTM (Temporal Convolutional Multimodal LSTM) (2021)

It is used to study inter-modality dynamics from a new perspective using audio and visual LSTM where language aspects are the most important. A well-designed gating mechanism is introduced inside each LSTM variant to enhance the language representation via the associated auxiliary modality. It is made up of two primary parts: (1) LSTMs with auditory and visual components to improve the representation of spoken language. A well-designed gating mechanism is included in each LSTM variation to assess if auditory and visual augmentation should be performed based on the discriminative information expressed in each modality. (2) by incorporating the 'Channel Interdependency Learning' module into the conventional TCN, a 'channel-aware' temporal convolution network is generated [57].

4.11.12. MMIM (MultiModal InfoMax) (2021)

Multimodal Infomax (MMIM) is a framework for multimodal fusion that maintains task-related information by hierarchically maximising Mutual Information (MI) between unimodal input pairings (inter-modality) and multimodal fusion result. To avoid the loss of critical task-related data, MI maximisation takes place at the input and fusion levels. As far as we know, this is the first attempt to link MI with MSA. To address the intractability problem, the formulation includes parametric learning and non-parametric GMM with stable and smooth parameter estimations [58].

4.11.13. OMSKELM (Optimal Multimodal Sentiment classification using the Kernel Extreme Learning Classifier) (2021)

This system looks at the relationship between text, audio, and video before doing multimodal sentiment analysis. A distinct set of features is extracted. Extracted features are then optimised using a new hybrid grass bee optimization algorithm (HGBEE), resulting in a feature set with an optimal value for improved precision and reduced computing time. Then utterance level multinodular fusion is created using text, audio, and video information. Finally, for sentiment classification, the system employs a multikernal extreme learning classifier (MKELM) [59].

4.11.14. TIMF (Two Level Multimodal Fusion) (2021)

In this system, a two-level multimodal fusion (TlMF) method incorporating both data-level and decision-level fusion is proposed to complete the sentiment analysis task, A tensor fusion network is used in the data-level fusion stage to create text-audio and text-video embeddings by fusing the text with audio and video properties, respectively. The soft fusion approach is used during the decision-level fusion step to fuse the upstream classifiers' classification or prediction results so that the final classification or prediction results are as accurate as possible. The CMU-MOSI, CMU-MOSEI, and IEMOCAP datasets are used to test it [62].

4.11.15. AWMA (Asymmetric Window Multi Attention) (2021)

The gated recurrent unit (GRU) module, asymmetric window attention (AWA) module and inter-modality attention (IMA) module make up the asymmetric window multi-attentions (AWMA) neural network. In multimodal sentiment analysis, the GRU and AWA modules are used to capture intra-modality dynamics. The IMA module is used to describe inter-modality dynamics. Then three modules are gradually merged. First the GRU module receives the input sequence followed by the AWA module and finally the IMA module. The use of asymmetric windows to express the asymmetric weights of historical and future contexts at a specific timestamp of the input data is a novel feature of proposed method. It believes that the previous and future contexts of input data at a specific timestamp should be distinguished. Furthermore, asymmetric windows can be used to indicate the implicit weights of the contexts [63].

4.11.16. Auto-ML based Fusion (2021)

Pre-processing, individual classifications, and fusion are the three stages of this approach. The fusion method focuses on performing a final classification using the independent classifications of the text and image components. The goal of the fusion strategy is to surpass individual classifications by leveraging context knowledge from both sources. The B-T4SA dataset is used, which is made up of Twitter data. Each sample from dataset contains both text and image [64].

4.11.17. Self-MM (Self-Supervised Multi-Task Multimodal sentiment analysis network) (2021)

Self-MM model proposed a self-supervised unimodal label generating process that saves a lot of time and money. Extensive testing has proven that the auto-generated unimodal labels are reliable and stable. The relative distance value is calculated using the distance between modality representations and class centres, which is positively connected with model outputs. It also devises a novel weight-adjustment approach to balance various task-related weight-loss restrictions. Self-goal MM's is to learn one multimodal task and three unimodal subtasks together to create information-rich unimodal representations. The MOSI, MOSEI and SIMS datasets are used [92].

4.11.18. DISRFN (Dynamic Invariant Specific Representation Fusion Network) (2022)

The redundant information of each mode may be utilised more effectively by joining modality-invariance and modality-specificity representations of each mode. A simple dynamic fusion method can acquire the interaction between modes more quickly in this model. The DISRFN framework is made up of two parts: enhanced JDSN and HGFN. To begin, an upgraded JDSN module is used to produce modal invariant-specific joint representations of each mode. It allows the complementary information amongst modes to be efficiently utilised and the heterogeneity gap between modes to be reduced. The MOSI and MOSEI data sets were used in the performance analysis experiment and the results were satisfactory [50].

4.11.19. MH-GAT (Multi-Feature Hierarchical Graph Attention Model) (2022)

In order to analyse sentiment, this research proposes a brain-inspired Multi-Feature Hierarchical Graph Attention Model (MH-GAT) based on co-occurrence and syntactic dependency graphs. It does this by simultaneously taking into account various structural information, part of speech information, and position association information. It consists of bi-graph hierarchical attention and multi-feature fusion. An input layer contains numerous features, such as part of speech, position, syntactic dependency, and co-occurrence information. It constructs bi-graph hierarchical attention model, hierarchical graph for every text and a graph attention network is developed. The suggested model's sentiment analysis accuracy has improved by an average of 5% compared to the most recent versions of the AT&T BLSTM, Text-Level-GNN, and TextING [98].

5. MSA applications

Sentiment Analysis is a technique for identifying and categorising opinions about a product or service. Multimodal sentimental analysis provides ways for doing opinion analysis using a combination of video, audio and text, which goes much beyond traditional text-based sentimental analysis in analysing human behaviour. Product reviews, opinion surveys, movie reviews on YouTube, news video analysis and health care applications, All benefit from the examination of these sentiments.

Multimodal Sentiment Analysis is used for Object recognition and fraud detection [65], Market Prediction for trading system [66], Tourism sentiment analysis, Sentiment Analysis of #MeToo tweets [67] . MSA has demonstrated remarkable success in human-machine interaction [68], health care applications [69], sentiment analysis in education and learning [70], recommendation systems [71], Sentiment towards any current issue [72], etc. Selected use cases are described next.

5.1. Multimodal action recognition and fraud detection

For fraud detection, a variety of modalities can be used. The task of avoiding fraudulent facial verification by utilising a photo, video, masque, or other substitute for an authorised person's face is known as facial anti-spoofing. It is used to detect erroneous inputs. The face anti-spoofing benchmarks use RGB, depth and infrared sensors to detect false face inputs. It's easy to deceive a single modality system. A printed face on paper or a 3D masque was able to trick an RGB-based system in a famous incident known as 'facegate'. On the other hand, deceiving many modalities at the same time is significantly more challenging. In the areas of action recognition, sentiment analysis and face anti-spoofing, a variety of datasets with different modalities have been presented. There are three different modalities to utilise in action recognition datasets:

visual (RGB sequence), motion (optical flow) and auditory. Only visual and motion modalities were employed until recently. A popular method was there to train one model for each modality and ensemble via late-fusion. Sensors cannot defend against 3D masks in face identification. Some actions (e.g., snapping) can only be recognised with audio modality while some sentiments can only be expressed through tone (audio), words (language) or facial expressions (visual).

5.2. Market prediction for trading system

Reinforcement learning, sentiment analysis and multimodal learning are all used in the development of a multimodal reinforcement trading system. When making a trading choice, the agent examines not only price movements but also news information. Multimodal learning, which can combine diverse modalities of data to improve the model's performance, and sentiment analysis, which can be used to understand the sentiment of news, are also introduced. Multiple modalities are concatenated with feature vectors or neural networks to produce joint representation. Financial emotion words are substantially connected with risk prediction according to the results of the experiment. Some experts believe that news events have an impact on stock price movements. The focus in 2020 is on mood classification, which has been shown to have an impact on stock market values. A new tool, a newspaper-based sentiment index, was proposed in 2021, allowing for real-time monitoring of economic activity in Spain. The measure not only exceeds the European Commission's famous economic attitude indicator but it also performs well in forecasting the Spanish gross domestic product (GDP).

5.3. Tourism sentiment analysis

Passengers frequently express their sentiments through social media platforms such as Weibo (a famous Chinese microblog social network). Using Weibos to analyse passenger opinions can assist comprehend current passenger moods, deliver timely and satisfactory service and improve the passenger experience. Multiple modalities including text and images are used to model travel events and sentiment. To learn discriminative multi-modal representations, both single-modality content and cross-modal relations are taken into account.

5.4. Sentiment analysis of Instagram posts

Instagram is a good social media platform for users to communicate their ideas, feelings, and opinions as well as like and comment on other people's posts. In the field of natural language processing and sentiment analysis, analysing the sentiment of posts has a variety of applications. A multimodal deep fusion method was proposed in the study to efficiently analyse the sentiment of posts. For picture and text sentiment analysis, the approach has two parallel branches. MPerInst, a multimodal dataset of Persian comments and their related image, was painstakingly gathered. The strategy outperforms similar multimodal deep models and classical machine learning methods, according to the findings of experiments.

5.5. Sentiment analysis of #MeToo tweets

People are talking more openly about their experiences of harassment as a result of the #MeToo movement. This project aims to bring together such sexual assault experiences in order to better understand social media constructions and effect societal change. The researchers used deep neural networks to combine visual analysis and natural language processing in a multimodal sentiment analysis approach. The method aims to determine a person's position on the subject and deduce the emotions expressed. A Multimodal Bi Transformer (MMBT) model that uses both picture and text information to predict a tweet's position and thoughts on the #MeToo campaign.

5.6. Human-computer interaction

Pepper is a robot that looks like a human. People can interpret communication with a robot in a good light in a variety of ways. Sick children, in particular, who may have difficulty communicating with others. The device, which consists of a Raspberry Pi, a camera, and a microphone, analyses audio, text, and video to provide a standalone source of information on an interlocutor's emotions, facial expressions, body language, and other factors.

5.7. Health care

Patients can now rate the quality of healthcare treatments they receive by publishing online evaluations due to the rise of social media. The online reviews' extensive text and visual material provides insights into patients' interactions with doctors and their satisfaction with healthcare service delivery. Various studies have used textual content to analyse patients' opinions. This research introduces a novel multimodal technique to analysing patients' perceptions of healthcare service quality (high vs. low). Not only original written contents, but also image contents from the Yelp.com platform is evaluated in this study. It is more difficult due to feature extraction from multiple modalities.

5.8. Education and learning

Sentiment discovery and analysis (SDA) is a technique for automatically detecting underlying attitudes, feelings, and subjectivity toward a specific thing such as learners and learning resources. SDA has been hailed as an effective technique for recognising and classifying feelings from multimodal and multisource data throughout the educational process due to its enormous potential. Multimodal SDA can provide complementing insights into educational stakeholders' sentiments in the classroom or through e-learning. SDA faces issues in unimodal feature selection, sentiment classification and multimodal fusion for large educational data streams. As a result, a substantial amount of research examines various ways to SDA for instructional purposes.

5.9. Recommendation systems

The most common application of a recommendation system in e-commerce (suggesting items to purchasers that may be of interest to them) is in online advertising (suggest to users the right contents, matching their preferences). Recommender systems, in a broad sense, are algorithms that try to propose relevant items to consumers (items being movies to watch, text to read, products to buy or anything else depending on industries). In some businesses, recommender systems are crucial since they can produce a large amount of revenue or serve as a method to differentiate yourself from competitors.

5.10. Sentiment towards any current issue

Sentiment analysis may be used to determine public opinion on a variety of topics as well as the peaks and valleys in the sentiment trend. COVID-19 immunisation is a big worry for every country in the world at this time of pandemic. The vaccine has been widely distributed since the end of 2020 and as a result, vaccinated people have seen fewer illnesses, hospitalizations, and fatalities. Vaccine apprehension, on the other hand, remains a concern, particularly in light of post-immunization side effects. Social media analysis has the ability to provide policymakers with information about side effects that are being discussed in the public as well as public perceptions of the national immunisation programme. Recent research has highlighted the potential for artificial intelligence-enabled social media analysis to supplement traditional assessment methods. For example, public surveys and provide information to governments and organisations on public sentiments. Social media analysis has yet to be applied to investigate regularly reported AEFI with a

COVID-19 vaccination. It could help to find potential safety signals that have gone unnoticed elsewhere (e.g. rarely reported side effects).

6. MSA challenges

According to the findings of the above study, future research should address the following issues: there is a need to establish a robust multimodal dataset in many languages. The dataset should be well-annotated and fine-graded. Co-reference resolution problem needs to be focused, hidden emotions, irony, and sarcasm detection is still an open research problem using multiple modalities. The dataset should be prepared and analysed ethically and widely available to public domains for better research and equal opportunities. The implicit and explicit meaning of different modalities, as well as code-mixed text data, short forms used in text, noisy and low-resolution photos and videos, and the implicit and explicit meaning of different modalities, must all be investigated. Along with all of the aforementioned, open research topics in multilingual directions include handling multilingual data, cross-domain accuracy improvement, cross-dataset accuracy improvement, and sentiment analysis using contextual backdrop.

The problem of identifying hidden emotions such as sarcasm or irony has long been a source of frustration for scholars in the discipline. Because these feelings are not immediately represented in the text, they are referred to as concealed emotions. These emotions are perceived by humans using two cues: nonverbal communication and context. Using multimodal sentiment analysis to detect nonverbal cues and context, the same cues may be utilised to detect concealed emotions. Multimodal sentiment analysis is a new field of study. The majority of previous research has been successful in demonstrating that combining two or three input types improves the accuracy of the analysis. The function of context in sentiment analysis has been investigated, with the conclusion that context enhances classification accuracy. A possible future study approach is to combine multimodal clues and context with a focus on identifying concealed emotions.

Feature extraction in sentiment analysis faces a number of issues, including domain specificity (which means it won't function in other domains), redundancy, high dimensionality, slag words, code-mixed data, biassing, and context-dependency. Some of the difficulties are listed below.

Cross Domain: Sentiment categorization is known to be domain-sensitive, as the modes of expressing opinions differ across domains. In sentiment analysis, domain adaptation tackles this difficulty by learning the features of the unknown domain. The meaning of a word spoken in one domain may differ from that of a word spoken in another area [73].

High Dimensionality: It refers to big feature sets that degrade performance due to computational issues, necessitating the employment of correct feature selection methods [74].

Code-Mixed Data: It is information in which two or more languages are employed in the same statement. Code-Mixing is the process of embedding linguistic features from one language, such as phrases, words, and morphemes, into an utterance from another language. 'Main kal movie dekhne jaa rahi thi and raaste me I met Rima' is example of Code-Mixing. The rule- and deep learning-based approaches are also challenged by code mixing. There has been very little development done in this area [75].

Biasing: Sentiment analysis tools are frequently employed in fields such as healthcare, which deals with sensitive themes such as counselling. Customer calls and marketing leads from varied backgrounds are frequently checked for emotion indicators, and acquired data drive important decision-making. As a result, it's crucial to recognise bias, especially when it comes to demographics. Bias can take many forms, including gender, colour, age, and so on [76].

Context Dependency: The use of sentiment words varies depending on the topic. When conjugated with other words or phrases, words that appear neutral on the surface can carry sentiment. When someone wants to buy a big house for leisure, the word big, for example, can have a positive connotation. When employed in this context, however, the same word might elicit negative feelings, a big house is difficult to clean. Unfortunately, little attention has been paid to this element of sentiment analysis research [76].

Early fusion/feature fusion: Since multimodal sentiment analysis takes into account the fusion of different modalities can be examined by various multimodal fusion techniques. Major challenges in early fusion-based MSA architecture are listed in Table 3.

Late fusion: Different features of each modality are examined and classified independently and then results are fused to generate final decision vector. Major problem with late fusion or decision fusion is as different classifiers are used for the analysis task, the learning process of all these classifiers at the decision-level fusion stage, becomes tedious

Table 3
Major issues related with the multimodal fusion based on early fusion.

Model Name	Merits	Demerits
Tri-modal HMM	• First time addresses the task of tri-modal sentiment analysis • Improvements are observed for both precision and recall • Tri-modal HMM is able to learn the hidden interaction between all three modalities and take advantage of their respective discriminative power	• Domain specific model • Small scale dataset
Text-audio-Visual	• Joint use of visual, audio, and textual features greatly improves over the use of only one modality at a time • Significant improvements are also obtained on a second dataset of English videos	• Domain specific • Occluded frames in videos
Proposed Multi Method	• System outperformed all state-of-the-art systems on the eNTERFACE dataset. • With multimodal fusion, system outperformed the best state-of-the-art system by more than 10%.	• Time complexity of the methods needs to be reduced to a minimum.
Multimodal Model	• Idea of thin slices can be used to observe a short window of a speaker's behaviour to achieve comparable prediction compared to observing the entire length of the video. • New dataset time complexity of the methods needs to be reduced to a minimum introduced.	• Same-gender design in evaluating a speaker's persuasiveness and other high-level attributes. • Inherent variability in human perception and judgement. • Investigating more ways of computationally capturing various indicators of persuasiveness and better algorithmic methods of fusing information from multiple modalities. • Deeper analysis to understand relationship between persuasiveness and relevant high-level attributes including personality.
MARN	• Main strength of our model comes from discovering interactions between modalities through time using a neural component called the Multi-attention Block (MAB) and storing them in the hybrid memory of a recurrent • Component called the Long-short Term Hybrid Memory (LSTHM).	• Under performs for cross-view dynamics • Performance varies as different tasks and datasets require different number of attentions

and time consuming. Major challenges in late fusion-based MSA architecture are listed in Table 4.

Hybrid fusion: This type of fusion is the combination of both feature-level and decision-level fusion methods. It exploits the advantages of both feature and decision-level fusion strategies and overcome the disadvantages of each. Major challenges in hybrid fusion-based MSA architecture are listed in Table 5.

Model level fusion: It's a technique that combines data from several modalities and uses correlation to create a relaxed fusion. The features of distinct modalities are studied to see if there is a link between them. The desired Model is then created based on the study domain and problem needs. Major challenges in model-level fusion-based MSA architecture are listed in Table 6.

Tensor fusion: In tensor based fusion method, a 3-fold Cartesian product from modality embeddings which explicitly models the unimodal, bimodal, and trimodal interactions is built by tensor fusion layer. Major challenges in tensor fusion-based MSA architecture are listed in Table 7.

Hierarchical fusion: Its hierarchical approach which proceeds from unimodal to bimodal vectors and then bimodal to trimodal vectors. Major challenges in hierarchical fusion-based MSA architecture are listed in Table 8.

Bimodal fusion: On pairwise modality representations, a novel end-to-end network which is bimodal fusion achieves fusion (relevance increment) and separation (difference increment). Major challenges in bimodal fusion-based MSA architecture are listed in Table 9.

Attention based fusion: Major challenges in attention based fusion MSA architecture are listed in Table 10.

Quantum based fusion: Quantum interference captures interactions inside each utterance (i.e., correlations across different modalities), and quantum measurement is used to create a strong-weak influence model to detect interactions between successive utterances (i.e., how one speaker influences another). Major challenges in quantum based fusion MSA architecture are listed in Table 11.

Word-level fusion: Table 12 describes major challenges in world-level fusion based MSA architectures

A bar chart depicting the greatest binary accuracy attained by each architecture is shown in Fig. 11.

7. Future scope

MSA's design incorporates a number of unique concepts that have moved research objectives, particularly in the area of sentiment analysis and emotion identification. It also serves as the foundation for review and recommender systems as well as health prediction systems. Studying MSA's architecture advancements is an exciting research topic that has the potential to become one of the most widely used natural language processing approaches.

Table 5
Major issues related with the multimodal fusion based on hybrid fusion.

Model Name	Merits	Demerits
ICT-MMMO	• language-independent audio-visual analysis is almost as effective as in- and cross-domain linguistic analysis, even though no textual information is used. • Scores retrieved automatically from the Web is a promising method to classify spoken reviews, such as those contained in YouTube videos	• BoNG features slightly yet not significantly outperform BoW features • With OKS, we estimate an accuracy of 59.6%, which is significantly above chance level, but significantly below the performance of in- or cross-domain analysis
Bimodal with Unimodal	• Much faster compare to SOTA models • Combine evidence from the words they utter, the facial expression, and the speech sound	• Less accurate • The parameter selection for decision level fusion produced suboptimal results. • Less scalable and stable
Three modalities	• Accuracy improves dramatically when audio, visual and textual modalities are used together • Gaze- and smile-based facial expression features are usually found to be very useful for sentiment classification	• Consumes more time

7.1. Mental health prediction

As more people are diagnosed with depression and other mental illnesses, automated mental health prediction, which identifies depression and other mental illnesses, is becoming a major research area. One out of every four adults in the United States suffers from a mental health condition, making mental health a significant priority. In recent research by Xu et al. [77], the evaluations are conducted at the user level using 10-fold cross validation. Thus, posts from the same user appear in either training or test sets but not both. The following classifiers are used: Decision Tree (DT, max depth= 4); Adaptive Boost (AB); Support Vector Machine with a linear kernel (SVMLinear). In addition, in the experiment, with a two-layer neural network classifier implemented using the TensorFlow library More specifically, a rectified linear unit (ReLU) is used as the activation function for the hidden layers, and a sigmoid function to constrain the output probability to be between zero and one. In another work, [78], they've got Implemented early and late fusion using machine learning to predict whether or not a person is stressed based on four modalities: computer interactions, body posture,

Table 4
Major issues related with the multimodal fusion based on late fusion.

Model Name	Merits	Demerits
Audio-Visual	• The model combines multiple measurements and can handle the absence of measurements by increasing the uncertainty of the predicted state • The uncertain measurements that are usually processed by the Kalman filter is replaced by multi-modal classifier outputs. • Filtering resulted in a reconstruction of missing classification outputs such that a class assignment is possible.	• A direct interpretation of parameter is difficult. For instance, the audio classifier is only able to provide outputs in case a signal is present, such that the classifier outputs of the video channel are much more frequent than the classifier outputs of the audio channel. As a result, the rejection rate and the process noise for audio and video cannot be compared with each other. • The ratio between these modalities is additionally modified by assessing the quality in terms of rejecting unreliable outputs
Ref	• The feature sets include audio-visual, lexical, POS, LIWC, emotional features and their combinations using majority voting. • Used predicted traits as features and designed a cascaded classification system	• Same feature set or architecture might not work for all traits. • Performance of the model using emotional feature set is very low compared to the other feature sets.
Multi CNN	• New CNN architecture that fully uses joint text-level and image-level representation • Complementary effect of the two representations as sentiment features improves performance	• Uses only two modalities of text and image.

Table 6
Major issues related with the multimodal fusion based on model-level fusion.

Model Name	Merits	Demerits
HMM-BLSTM	• Context-sensitive schemes for emotion recognition within a multimodal, hierarchical approach at utterance level • Sheds light into the flow of affective expressions revealing potentially useful patterns	• Less accuracy in context sensitive as compared to context free
Context-aware BLSTM	• System based on LSTM longrange temporal context modelling in order to discriminate between high and low levels of AROUSAL, EXPECTATION, POWER, and VALENCE using statistical functionals of a large set of acoustic low-level descriptors, linguistic information (including non-linguistic vocalizations), and facial movement feature	• Under performs and absolute values of the reported accuracies seem low in comparison to easier scenarios, such as the discrimination of acted, prototypical emotions.

facial expressions, and heart rate variability. They used transfer learning to forecast the NASA-TLX score, which predicts a person's stress level on a scale of 0 to 100 and provides a mechanism to store data from monitoring a person's mental state as the task load increases across a given timeline. In another research by Aloshban et al. [97], Bidirectional Long-Short Term Memory networks for sequence modelling and multimodal features which include language and acoustic features are employed in this work. It applied late fusion, joint representation and gated multimodal units. It produced better results as compared to unimodal.

7.2. Emotion recognition

Emotion is inextricably linked to humans; hence emotion comprehension is an important component of human-like artificial intelligence (AI). Due to its ability to mine opinions from a plethora of publicly available conversational data on platforms such as Facebook, YouTube, Reddit, Twitter, and others, emotion recognition in conversation (ERC) is becoming increasingly popular as a new research frontier in natural language processing (NLP). It could also be used in health-care systems (as a tool for psychological analysis), education (understanding student frustration), and other fields. Furthermore, ERC is critical for creating emotion-aware interactions that necessitate a grasp of the user's emotions. ERC presents a number of obstacles, including conversational context modelling, interlocutors' emotion shift, and others, all of which make the task more difficult to do [79]. In another work in ER, a context- and sentiment-aware framework, termed Sentic GAT is introduced. In Sentic GAT, commonsense knowledge is dynamically represented by the context- and sentiment-aware graph attention mechanism, and intra- and inter-dependency relationship of contextual utterances is obtained by the dialogue transformer based on hierarchical multi-head attention. The inter- and intra-dependency mean the dependency relationships of contextual information and its own key information on target utterance [80].

7.3. Sarcasm detection

Sarcasm is the use of words or idioms to convey the opposite meaning of what is really said. People use sarcasm to make fun of others or to condemn them. Sarcasm is a concept used in multimodal sentiment analysis to describe attitudes in which people can communicate bad feelings using positive words or expressions and positive feelings using negative words or expressions. It functions as a polarity-flipping interfering object. Sarcasm sentiment analysis is an area of FULL (NLP) study that is fast expanding. It includes categorization at the word, phrase, and sentence levels as well as classification at the document and idea levels. Sarcastic sentiment detection can be divided into three categories based on the text features used: lexical, pragmatic, and hyperbolic. Due to the figurative nature of writing, which is accompanied by nuances and implicit meanings, detecting sarcasm is extremely difficult. This field of research has established itself as a major problem in NLP in recent years, with numerous papers providing various strategies to approach this task. The voice and text community has made the most significant contributions. Sarcasm is frequently presented without the use of linguistic indicators, instead relying on non-verbal and verbal clues. Changes in tone, overemphasis on words, and a straight face are

Table 7
Major issues related with the multimodal fusion based on tensor fusion.

Model Name	Merits	Demerits
TFN	• Learns intra-modality and inter-modality dynamics end-to-end • Tensor fusion layer used to generate 3-fold Cartesian product from modality embeddings	• Risk of Overfitting due to very high dimensionality of the produced tensor especially in case of small datasets like CMU-MOSI • Less specialized fusion process due to little consideration to acknowledging the variations across different portions of a feature vector • Exponential computational increase in number of parameters, cost and memory
MRFF	• Removes redundant information which is duplicated across modalities and in turn leads to fewer parameters with minimal information loss. • Work as a regularizer, in turn leads to less complicated model and reduces overfitting • Modality-based factorization approach helps to understand the differences in useful information between modalities for the task at hand • Tuckers tensor decomposition method is used which gives different compression rates for each modality.	• Small dataset • Requires more efficient training for large dataset
T2FN	• T2FN with rank regularization maintains good performance despite imperfections in data • Model's improvement is more significant on random drop settings, which results in a higher tensor rank as compared to • Structured drop settings	• Tensor rank increases in presence of imperfect data
MTFN—CMM	• Multi-tensor fusion network with the cross-modal modelling for multimodal sentiment analysis in this study, which can capture intramodal dynamics and intermodal interactions and can be used for multi-modal affective intensity prediction effectively • MTFN—CMM can perform better in regression and classification experiments	• Works on coarse grain modal fusion • High number of parameters used

Table 8
Major issues related with the multimodal fusion based on hierarchical fusion.

Model Name	Merits	Demerits
CHFusion	• Hierarchical Fusion, first fusing the modalities two in two and only then fusing all three modalities • Context-Aware Hierarchical Fusion (CHFusion)	• Performance difference between two datasets • Less accurate in unimodal specially in textual modalities
HFNN	• It achieves the highest F1 score. • HFFN learns local interactions at each local chunk and explores global interactions by conveying information across local interactions using ABS-LSTM that integrates two levels of attention mechanism	• The accuracy of HFFN is lower than that of BC-LSTM and CAT-LSTM

Table 9
Major issues related with the multimodal fusion based on bimodal fusion.

Model Name	Merits	Demerits
BBFN	• Bi-Bimodal Fusion Network- novel end-to- end network that performs fusion (relevance increment) and separation (difference increment) on pairwise modality representations • Modality-specific feature space separator and gated control mechanism	• Performance degrades when visual-acoustic input pairs are added. That is, even after including all modalities in the input, redundant network architecture can cause harmful effects bringing in malicious noise, which damages collected useful information and confuses the model.
BIMHA	• Information-augmented Multi-Head Attention using bimodal feature Inter-modal interaction and inter-bimodal interaction.	• BIMHA is a shallow model in terms of efficiency.

Table 10
Major issues related with the Multimodal Fusion based on Attention Based Fusion.

Model Name	Merits	Demerits
CATF-LSTM	• Attention-based long short-term memory (LSTM) network which not only models the contextual relationship amongst utterances, but also prioritizes more relevant utterances for classifying the target utterance • Attention capability to capture important • Employs attention mechanism in fusion	• Textual CAT-LSTM classifier performs better than trimodal CATF-LSTM for the presence of noise in the audio modality or when the speaker does not look directly at the camera while speaking
MMHA	• Learn the dependence of unimodal features, capture the internal structure of unimodal features • Learn the relationship between the multiple modalities, put more attention on the contributing features • Multi-head attention-based network to predict the sentiment of each utterance	• Due to the limitations of the text modality, bimodal and tri-modal accuracy scores are not improved well • The combination of audio and text accuracy score is lower than the state-of-the-art
Bi-LSTM with attention model	• Novel attention-based multimodal contextual fusion strategy, which extract the contextual information amongst the utterances before fusion	• Positive class obtains the maximum recognition rate than the negative class in all modality combinations. • It misclassified negative class modality

Table 11
Major issues related with the multimodal fusion based on quantum based fusion.

Model Name	Merits	Demerits
QMR	• Quantum-inspired Multimodal Sentiment Analysis • Fill the 'semantic gap' and model the correlations between different modalities via density matrix • Quantum Interference inspired Multimodal decision Fusion (QIMF) • QIMF adds an interference term	• Influence of different cos θ on the classification results • The computation time used for training and classification is longer than the use of other baselines
QMN	• It uses quantum theory (QT) mathematical formalism and a long short-term memory (LSTM) network • Intra- and inter-utterance interaction dynamic is considered • Decision correlations between different modalities is generated using quantum interference • Strong-weak influence model to make better inferences about social influence amongst speakers is generate using quantum measurement	• It is largely dependant on the density matrix so how to take a further step towards accurately capturing the interactions amongst speakers and naturally incorporating them into an end-to-end framework is difficult. • Because it used an emotion recognition dataset, QMN was tested on emotion recognition tasks rather than sentiment analysis.
QMF	• Superposition and entanglement are used respectively at different stages are used to formulate word interaction within a single modality and the interaction across modalities. • Quantum-theoretic multimodal fusion framework	• Quality of the extracted visual and acoustic features is not high • Inconsistency with quantum theory

all signs of sarcasm. There have been very few works that adopt multimodal strategies to determine sarcasm [76]. Some of the works include the work of Castro et al. [81], in which a new dataset, MUStARD is created for multimodal sarcasm research with high-quality annotations, including both mutlimodal and conversational context features. It also provides preceding turns in the dialogue which act as context information. Consequently, it summarises that this property of MUStARD leads to a new sub-task for future work: sarcasm detection in conversational context. In another study by Du et al. [100], they proposed a dual-channel convolutional neural network which analyses the semantics of the target text along with its sentimental context. They used SenticNet to add common sense to the long short-term memory (LSTM) model. Then attention mechanism is applied to take the user's expression habits into consideration.

7.4. Fake news detection

Fake news on social media and other platforms is widespread, and it is a reason for great concern because of its potential to inflict significant social and national harm with negative consequences. Detection is already the topic of a lot of study. In the year 2020, there was widespread bogus news about health, putting world health at risk. Early in February 2020, the WHO issued a warning that the COVID-19 epidemic had resulted in a large 'infodemic,' or a burst of real and fake news, which contained a lot of misinformation. This is a region where virtually little work is done. In research by Patwa et al. [82], they describe and release a fake news detection dataset containing 10,700 fake and real news related to COVID-19. Posts from various social media and fact checking web-sites are collected, and manually verified the veracity of each post. The data is class-wise balanced and can be used to develop automatic fake news and rumour detection algorithms. The developed dataset is benchmarked using machine learning algorithm and project

Table 12
Major issues related with the multimodal fusion based on word level fusion.

Model Name	Merits	Demerits
MFN	• MFN shows a consistent trend for both classification and regression • DMAN can model asynchronous cross-view interactions because it attends to the memories in the System of LSTMs which can carry information about the observed inputs across different timestamps.	• Doesn't support cross view dynamics • Underperforms on some of the datasets than baseline models
DFG	• A new neural-based component called the Dynamic Fusion Graph which replaces DMFN in MFN	• Efficiency decreases when visual modalities are contradictory
RMFN	• Decomposes the fusion problem into multiple stages, each of them focused on a subset of multimodal signals for specialized, effective fusion • Crossmodal interactions are modelled using this multistage fusion approach which builds upon intermediate representations of previous stages • Temporal and intra-modal interactions are modelled by integrating our proposed fusion approach with a system of recurrent neural networks	• Works well for CMU-MOSI dataset than all other datasets.
RAVEN	• Recurrent Attended Variation Embedding Network (RAVEN) • Nonverbal Sub-networks • Gated Modality mixing Network • Multimodal Shifting that models the fine-grained structure of nonverbal subword sequences and dynamically shifts word representations based on nonverbal cues. considering subword structure of nonverbal behaviours • Learning multimodal-shifted word representations conditioned on the occurring nonverbal behaviours.	• Need for large dataset • Underperforms in cross domain

them as the potential baselines. amongst the machine learning models, SVM-based classifier performs the best with 93.32%F1-score on the test set.

7.5. Hate speech detection

Hate speech is used to convey disdain for a specific group of people. It can also be used in any social media to embarrass or degrade members of the group. Hate speech in black-and-white on social media that disparages and can injure or put the victim in a risky situation. It is a biased, unreceptive, and cruel speech directed at a person or a group of people because of some of their worst traits. Hate speech on the internet, particularly on microblogging sites like Twitter, has become possibly the most serious issue of the last decade. Hate crimes have increased dramatically in several nations as a result of malevolent hate campaigns. While hate speech detection is an emerging study area, the genesis and dissemination of topic-dependant hatred in the information network has yet to be investigated. In research by Masud et al. [83], they predict the initiation and spread of hate speech on Twitter. Analysing a large Twitter dataset that we crawled and manually annotated for hate speech, we identified multiple key factors (exogenous information, topic-affinity of the user, etc.) that govern the dissemination of hate. In another research by Araque and Iglesias [99], they used machine learning framework using an ensemble of different features. They also studied the effects of different feature extraction methods such as TF-IDF and SIMilarity-based sentiment projectiON (SIMON). They used five different datasets cover radicalization and hate speech detection tasks.

7.6. Deception detection

Deception is defined as persuading or persuading someone to believe something that is not true. It is described as a message sent with the intent of instilling in the receiver a false belief. Most users are concerned about distinguishing true from fake messages in this digital era of computer-facilitated interactions. This is especially important for activities with a high-risk element. Online banking, shopping, and information searching in critical areas such as health or financial guidance are just a few examples. In a work by Chebbi and Jebara [84], they use audio, video, and text modalities to automatically discriminate between deception and truth, and they're looking into merging them to identify deception more precisely. Each modality was examined separately first, and then a feature and decision-level fusion strategy to integrate the modalities was presented. The proposed feature level fusion approach investigates a variety of feature selection techniques in order to select the most relevant ones from the entire used feature set, whereas the decision level fusion approach is based on belief theory and takes into account information about each modality's certainty degree. To accomplish so, we used a real-life video dataset from public American court hearings of people interacting truthfully or falsely.

7.7. Stress detection

Multiple types of stress exist in today's culture, all of which have an impact on our mental and physical health. Stress is defined as a

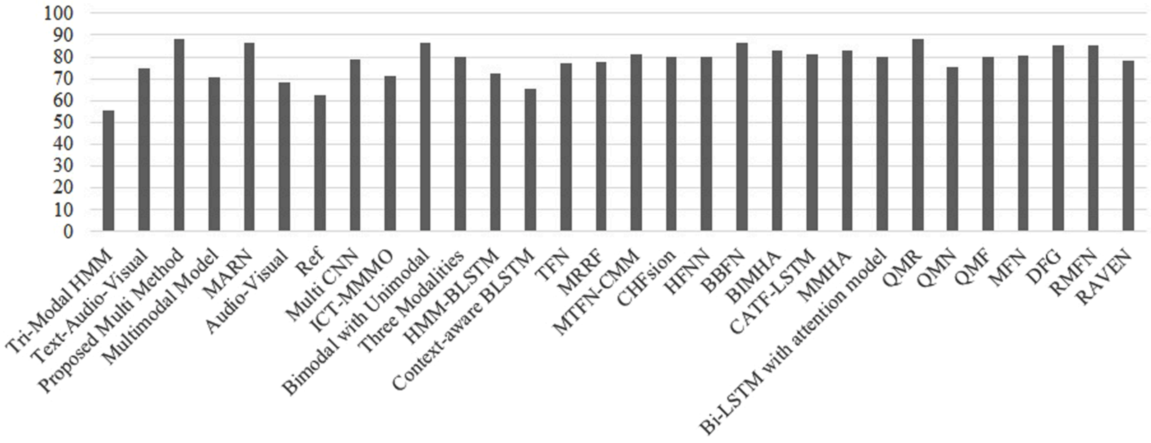

Fig. 11. Binary accuracy of each architecture.

transition from a peaceful to an enthusiastic state with the goal of maintaining a person's integrity. Stress can have a number of detrimental effects on a person's work/life productivity, including reduced decision-making ability, lower situational awareness, and poor performance. In a study by Sander et al. [85], use a repeated measures experimental design was used in which the same individuals work under two noise conditions which are carefully manipulated to simulate typical open plan and private office noise levels. In another work by Mou et al. [86], a framework of driver stress detection through multimodal fusion using attention based deep learning techniques is proposed. Specifically, an attention based convolutional neural networks (CNN) and long short-term memory (LSTM) model is proposed to fuse non-invasive data, including eye data, vehicle data, and environmental data. Following this, the proposed model can automatically extract features separately from each modality and give different levels of attention to features from different modalities through a self-attention mechanism.

7.8. Multimodal robust systems

Creating multimodal robust systems is a difficult task. Some of the benchmark studies have attempted to improve baseline results and make the system more robust. AI still struggles with complex tasks that require commonsense reasoning, such as natural language understanding. In the study by Cambria et al. [94], a commonsense-based neurosymbolic framework is created that aims to overcome these issues in the context of sentiment analysis. In particular, they employ unsupervised and reproducible sub symbolic techniques such as auto-regressive language models and kernel methods to build trustworthy symbolic representations that convert natural language to a sort of protolanguage and, hence, extract polarity from text in a completely interpretable and explainable manner. In another study by Zou et al. [95], a more robust system is created by applying multiple theories. A novel multimodal fusion architecture is developed from an information theory perspective, and its practical utility is demonstrated using Light Detection and Ranging (LiDAR) camera fusion networks. For the first time, a multimodal fusion network as a joint coding model is created, where each single node, layer, and pipeline is represented as a channel. In another study [96], the majority of research in multimodal sentiment analysis is performed on a dataset with speaker overlap in train-and-test splits. As each individual is unique in how they express emotions and sentiments, it is imperative to find generic, person independent features for sentimental analysis. However, given this overlap, where the model has already observed the behaviour of a certain individual, the results do not scale to true generalization. In real-world applications, the model should be robust to individual variance.

8. Conclusion

Research into MSA has gained traction over the past decade. In particular, studies have demonstrated its effectiveness for sentiment prediction and emotion recognition. A number of significant contributions have been identified, such as adaptation of MSA fusion methods to improve efficiency and performance. Other categories in which MSA advancements have been reported include: variation in the number of modalities that are bimodal or trimodal; context-aware and speaker-independent humour and sarcasm detection; fusion techniques; application-specific modifications in architectures and development of various learning algorithms and recommendation systems. This timely review summarizes recent developments in MSA architectures. Ten fundamental MSA architectural advancements are identified based on fusion categories, specifically: early, late, hybrid, model-level, tensor, hierarchical, bi-modal, attention-based, quantum-based and word-level fusion respectively. Several MSA architectural variations were examined with the word level architecture identified as the most efficient, which is defined by classifying target utterance using contextual information from neighbouring utterances in a video. The architecture has two components, the order of which varies based on the model. The first module is the context extraction module, which is used to model the contextual link between neighbouring utterances in the video and highlights relevant contextual utterances that are more important to predict the target emotion. In most recent models, a bidirectional recurrent neural network-based module is utilised. The second module is the attention-based module, which is responsible for merging the three modalities (text, audio, and video) and contextually selecting the most useful ones. Finally, a number of key MSA applications and future research challenges and opportunities were identified.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgement

The authors would like to thank the anonymous reviewers for their insightful comments and suggestions. Amir Hussain acknowledges the support of the UK Engineering and Physical Sciences Research Council (EPSRC) - Grants Ref. EP/M026981/1, EP/T021063/1, EP/T024917/1. Soujanya Poria acknowledges the support of Singapore's Ministry of Education for its AcRF Tier-2 grant (Project no. T2MOE2008, and Grantor reference no. MOE-T2EP20220–0017).

References

[1] J. A. &. V.J.D. Balazs, Opinion mining and information fusion: a survey Inform. Fusion 27 (2016) 95–110.

[2] S. L. C. &. C.J. Sun, A review of natural language processing techniques for opinion mining systems Inform. Fusion 36 (2017) 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 *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 2007.

[5] S Mohammad, P Turney, "Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon," 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] L.P. Morency, R. Mihalcea, P. Doshi, 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] Ver 'onica Perez Rosas, Rada Mihalcea, Louis-Philippe Morency, Multimodal sentiment analysis of Spanish online videos, IEEE Intell. Syst. (2011) 38–45.

[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), 2236–2246*, Melbourne, Australia, 2018.

[12] Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas P.Y.K.L., Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, Bjorn Gamback, "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, S. Bjorn, The multimodal sentiment analysis in car reviews (MuSe-CaR) dataset: collection, insights and improvements, IEEE Trans. Affect. Comput. (2021) (Early Access)1-1.

[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,* 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 *Proceedings of the 13th international conference on multimodal interfaces*, 2011.

[20] V.P. Rosas, R. Mihalcea, L.P. Morency, Multimodal sentiment analysis of spanish online videos, IEEE Intell. Syst. 28 (3) (2013) 38–45.

[21] S. Poria, E. Cambria, A. Hussain, G.B. Huang, Towards an intelligent framework for multimodal affective data analysis, Neural Netw. 63 (2015) 104–116.

[22] S. Park, H.S. Shim, M. Chatterjee, K. Sagae, L.P. Morency, Multimodal analysis and prediction of persuasiveness in online social multimedia, ACM Trans. Interact. Intell. Syst. (TiiS) 6 (3) (2016) 1–25.

[23] Zadeh, A., Liang, P.P., Poria, S., Vij, P., Cambria, E., & Morency, L.P., " Multi-attention recurrent network for human communication comprehension.," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 5642-5649, 2018.

[24] Glodek, M., Reuter, S., Schels, M., Dietmayer, K., & Schwenker, F., "Kalman filter based classifier fusion for affective state recognition," *International Workshop On Multiple Classifier Systems*, Berlin, Heidelberg., 2013.

[25] Alam, F., & Riccardi, G., "Predicting personality traits using multimodal information," in *Proceedings of the 2014 ACM multi media on workshop on computational personality recognition*, 2014.

[26] G.&.X.B. Cai, Convolutional neural networks for multimedia sentiment analysis," in *Natural Language Processin*, Chinese Computing (2015).

[27] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, L.P. Morency, Youtube movie reviews: sentiment analysis in an audio-visual context, IEEE Intell. Syst. 28 (3) (2013) 46–53.

[28] S. Poria, E. Cambria, A. Gelbukh, " Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," *Proceedings of the 2015 conference on empirical methods in naturl language processing,* 2539–2544, 2015.

[29] S. Poria, E. Cambria, N. Howard, G.B. Huang, A. Hussain, Fusing audio, visual and textual clues for sentiment analysis from multimodal content, Neurocomputing 174 (2016) 50–59.

[30] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, S. Narayanan, Context-sensitive learning for enhanced audiovisual emotion classification, IEEE Trans. Affect. Comput. 3 (2) (2012) 184–198.

[31] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, G. Rigoll, LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework, Image Vis. Comput. 31 (2) (2013) 153–163.

[32] Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.P., "Tensor fusion network for multimodal sentiment analysis," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 1103–1114, 2017.

[33] E. J. &. F.P. Barezi, "Modality-based factorization for multimodal fusion," in *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, Florence, Italy, 2019.

[34] Liang, P.P., Liu, Z., Tsai, Y.H.H., Zhao, Q., Salakhutdinov, R., & Morency, L.P. Learning representations from imperfect time series data via tensor rank regularization, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019.

[35] X. Yan, H. Xue, S. Jiang, Z. Liu, Multimodal sentiment analysis using multi-tensor fusion network with cross-modal modeling, Appl. Artif. Intell. (2021) 1–16.

[36] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, S. Poria, Multimodal sentiment analysis using hierarchical fusion with context modeling, Knowl. Based Syst. 161 (2018) 124–133.

[37] Mai, S., Hu, H., & Xing, S., " Divide, conquer and combine: hierarchical feature fusion network with local and global perspectives for multimodal affective computing," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[38] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.P. Morency, S. Poria, Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis, in: Proceedings of the 2021 International Conference on Multimodal Interaction, IEEE, 2021.

[39] Wu, T., Peng, J., Zhang, W., Zhang, H., Tan, S., Yi, F., ... & Huang, Y., " Video sentiment analysis with bimodal information-augmented multi-head attention," *Knowl. Based Syst.,* 235, 2022.

[40] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, L.P. Morency, Multi-level multiple attentions for contextual multimodal sentiment analysis, in: International Conference on Data Mining, IEEE, 2017, pp. 1033–1038.

[41] C. L. G. &. Y.J. Xi, Multimodal sentiment analysis based on multi-head attention mechanism, in: Proceedings of the 4th international conference on machine learning and soft computing, IEEE, 2020.

[42] M.G. Huddar, S.S. Sannakki, V.S. Rajpurohit, Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM, Multimed. Tools Appl. 80 (9) (2021) 13059–13076.

[43] Y. Zhang, D. Song, P. Zhang, P. Wang, J. Li, X. Li, B. Wang, A quantum-inspired multimodal sentiment analysis framework, Theor. Comput. Sci. 752 (2018) 21–40.

[44] Y. Zhang, D. Song, X. Li, P. Zhang, P. Wang, L. Rong, B. Wang, A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis, Inform. Fusion 62 (2020) 14–31.

[45] Q. Li, D. Gkoumas, C. Lioma, M. Melucci, Quantum-inspired multimodal fusion for video sentiment analysis, Inform. Fusion 65 (2021) 58–71.

[46] Zadeh, A., Liang, P.P., Mazumder, N., Poria, S., Cambria, E., & Morency, L.P., Memory fusion network for multi-view sequential learning, In *Proceedings of the AAAI Conference on Artificial Intelligence,* 5634–5641, 2018.

[47] A. Zadeh, P. Liang, S. Poria, E. Cambria, and L.P. 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*, 2236–2246, 2018.

[48] Liang,P.P., Liu, Z., Zadeh, A., & Morency, L.P., "Multimodal language analysis with recurrent multistage fusion," 2018.

[49] Y. Wang, Y. Shen, Z. Liu, P.P. Liang, A. Zadeh, L.P. Morency, Words can shift: dynamically adjusting word representations using nonverbal behaviors, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, 2019.

[50] J. He, H. Yanga, C. Zhang, H. Chen, Y. Xua, Dynamic Invariant-Specific Representation Fusion Network for Multimodal Sentiment Analysis, Comput. Intell. Neurosci. (2022).

[51] X. Wang, J. He, Z. Jin, M. Yang, Y. Wang, H. Qu, M2Lens: visualizing and explaining multimodal models for sentiment analysis, IEEE Trans. Vis. Comput. Graph. (2021) 802–812.

[52] K. Dashtipour, M. Gogate, E. Cambria, A. Hussain, A novel context-aware multimodal framework for persian sentiment analysis, Neurocomputing (2021) 377–388.

[53] Y. Li, K. Zhang, J. Wang, X. Gao, A cognitive brain model for multimodal sentiment analysis based on attention neural networks, Neurocomputing (2021) 159–173.

[54] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognit. Lett. 125 (2019) 264–270.

[55] H.N. Tran, E. Cambria, Ensemble application of ELM and GPU for real-time multimodal sentiment analysis, Memetic Comput. 10 (1) (2018) 3–13.

[56] J. Xu, Z. Li, F. Huang, C. Li, S.Y. Philip, Social image sentiment analysis by exploiting multimodal content and heterogeneous relations, IEEE Trans. Ind. Inf. 17 (4) (2020) 2974–2982.

[57] S. Mai, S. Xing, H. Hu, Analyzing multimodal sentiment via acoustic-and visual-LSTM with channel-aware temporal convolution network, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2021) 1424–1437.

[58] Han, W., Chen, H., & Poria, S., Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis, in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,* Online and Punta Cana, Dominican Republic, 2021.

[59] A.S. Alqahtani, S. Pandiaraj, M. Murali, S. Alshmrany, H. Alsarrayrih, Hybrid grass bee optimization-multikernal extreme learning classifier: multimodular fusion strategy and optimal feature selection for multimodal sentiment analysis in social media videos, Concurr. Comput. Pract. Exper. 33 (16) (2021).

[60] Wu, W., Wang, Y., Xu, S., & Yan, K., "SFNN: Semantic Features Fusion Neural Network for multimodal sentiment analysis.," in *5th International Conference on Automation, Control and Robotics Engineering (CACRE)*, 2020.

[61] M. &. L. X. Chen, SWAFN: sentimental words aware fusion network for multimodal sentiment analysis, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020.

[62] J. Sun, H. Yin, Y. Tian, J. Wu, L. Shen, L. Chen, Two-Level Multimodal Fusion For Sentiment Analysis in Public Security, Security and Communication Networks, 2021.

[63] Lai, H., & Yan, X., " Multimodal sentiment analysis with asymmetric window multi-attentions," *Multimedia Tools and Applications,* pp. 1–14, 2021.

[64] V. Lopes, A. Gaspar, L.A. Alexandre, J. Cordeiro, An AutoML-based approach to multimodal image sentiment analysis, in: International Joint Conference on Neural Networks (IJCNN), IEEE, 2021.

[65] Jongchan Park, Min-Hyun Kim and Dong-Geol Choi, "Correspondence learning for deep multi-modal recognition and fraud detection," *Electronics (Basel)*, p. 800, 2021.

[66] S.-H.H. Yu-Fu Chen a, "Sentiment-influenced trading system based on multimodal deep reinforcement learning," ELSEVIER*,* 2021.

[67] P. Basu, S. Tiwari, J. Mohanty and S. Karmakar, "Multimodal Sentiment Analysis of #MeToo Tweets," in *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, New Delhi, 2020.

[68] M.P. Agnieszka Rozanska, "Multimodal sentiment analysis applied to interaction between patients and a humanoid robot Pepper," ELSEVIER, pp. 411–414, 2019.

[69] Adnan Muhammad Shah, Xiangbin Yan,Syed Asad Ali Shah,Gulnara Mamirkulova, “Mining patient opinion to evaluate the service quality in healthcare: a deep-learning approach,” Springer*,* p. 2925–2942, 2019.

[70] Zhongmei Han, Jiyi Wu, Changqin Huang, Qionghao Huang, Meihua Zhao, A review on sentiment discovery and analysis of educational big-data, WIREs Data Mining Knowl. Discov. (2019).

[71] Yang Li, Suhang Wang, Quan Pan, Haiyun Peng, Tao Yang, Erik Cambria, “Learning binary codes with neural collaborative filtering for efficient recommendation systems,” Elsevier*,* pp. 64–75, 2019.

[72] Z. Hussain, Z. Sheikh, A. Tahir, K. Dashtipour, M. Gogate, A. Sheikh, A. Hussain, Artificial Intelligence-Enabled Social Media Analysis For Pharmacovigilance of COVID-19 Vaccinations in the United Kingdom: Observational Study, Springer, Cham, 2022, pp. 13–43. JMIR Public.

[73] Sharma, R., Bhattacharyya, P., Dandapat, S., and Bhatt, H.S., “Identifying transferable information across domains for cross-domain sentiment classification,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL* 2018, Melbourne, Australia, 2018.

[74] Wilson, T., J. Wiebe and P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing,* Vancouver, British Columbia, Canada, 2005.

[75] Y.K. Lal, V. Kumar, M. Dhar, M. Shrivastava, P. Koehn, De-mixing sentiment from code-mixed text, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, ACL, 2019.

[76] and S. Poria, D. Hazarika, N. Majumder, R. Mihalcea, Beneath the tip of the iceberg: current challenges and new directions in sentiment analysis research, IEEE Trans. Affect. Comput. (2020) *(Early Access)*.

[77] Z. Xu, V. Pérez-Rosas, R. Mihalcea, Inferring social media users’ mental health status from multimodal information, in: Proceedings of the 12th Language Resources and Evaluation Conference, IEEE, 2020.

[78] R. Walambe, P. Nayak, A. Bhardwaj, K. Kotecha, Employing multimodal machine learning for stress detection, J. Healthc. Eng. (2021).

[79] S. Poria, N. Majumder, R. Mihalcea, E. Hovy, Emotion recognition in conversation: research challenges, datasets, and recent advances, IEEE Access 7 (2019) 100943–100953.

[80] G. Tu, J. Wen, C. Liu, D. Jiang, E. Cambria, Context-and sentiment-aware networks for emotion recognition in conversation, IEEE Trans. Artif. Intell. 3 (5) (2022) 699–708.

[81] Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., & Poria, S., Towards multimodal sarcasm detection (an _obviously_ perfect paper), in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019.

[82] Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, Tanmoy Chakraborty, “Fighting an infodemic: covid-19 fake news dataset,” in *International Workshop on Combating Online Hostile Posts in Regional Languages During Emergency Situation*, 2021.

[83] Sarah Masud, Subhabrata Dutta, Sakshi Makkar, Chhavi Jain, Vikram Goyal, Amitava Das, Tanmoy Chakraborty, “Hate is the new infodemic: a topic-aware modeling of hate speech diffusion on Twitter,” in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, Chania, Greece, 2021.

[84] Chebbi, S., & Jebara, S.B., “Deception detection using multimodal fusion approaches.,” *Multimedia Tools and Applications,* pp. 1–30., 2021.

[85] E.L.J. Sander, C. Marques, J. Birt, M. Stead, O. Baumann, Open-plan office noise is stressful: multimodal stress detection in a simulated work environment, J. Manage. Organiz. (2021) 1–17.

[86] L. Mou, C. Zhou, P. Zhao, B. Nakisa, M.N. Rastgoo, R. Jain, W. Gao, Driver stress detection via multimodal fusion using attention-based CNN-LSTM, Expert Syst. Appl. 173 (2021).

[87] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.F. Chang, M. Pantic, A survey of multimodal sentiment analysis, Image Vis. Comput. 65 (2017) 3–14.

[88] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: from unimodal analysis to multimodal fusion, Inform. Fusion 37 (2017) 98–125.

[89] L.C. De Silva, T. Miyasato, R. Nakatsu, Facial emotion recognition using multi-modal information, in: Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Informal Engineering and Wireless Multimedia Communications, IEEE, 1997 (Cat. (Vol. 1).

[90] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). Meld: a multimodal multi-party dataset for emotion recognition in conversations, In Proceedings of ACL, 527-536, 2019.

[91] and Devamanyu Hazarika, Roger Zimmermann, Soujanya Poria, MISA: modality-invariant and -specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1122–1131.

[92] and W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, AAAI 35 (12) (2021) 10790–10797. May.

[93] W. Rahman, M.K. Hasan, S. Lee, A. Zadeh, C. Mao, L.P. Morency, E. Hoque, Integrating multimodal information in large pretrained transformers, in: Proceedings of the conference. Association for Computational Linguistics. Meeting 2020, NIH Public Access, 2020, p. 2359.

[94] E. Cambria, Q. Liu, S. Decherchi, F. Xing, K. Kwok, SenticNet 7: a commonsense-based neurosymbolic AI framework for explainable sentiment analysis, in: Proceedings of LREC, ELRA, 2022, pp. 3829–3839.

[95] Z. Zou, X. Zhang, H. Liu, Z. Li, A. Hussain, J. Li, A novel multimodal fusion network based on a joint coding model for lane line segmentation, Inform. Fusion 80 (2022) 167–178.

[96] E. Cambria, S. Poria, A. Hussain, Speaker-independent multimodal sentiment analysis for big data. Multimodal Analytics for Next-Generation Big Data Technologies and Applications, Springer, Cham, 2019, pp. 13–43.

[97] N. Aloshban, A. Esposito, A. Vinciarelli, What you say or how you say it? depression detection through joint modeling of linguistic and acoustic aspects of speech, Cognit. Comput. (2021) 1–14.

[98] Z. Jin, M. Tao, X. Zhao, Y. Hu, Social media sentiment analysis based on dependency graph and co-occurrence graph, Cognit. Comput. (2022) 1–16.

[99] O. Araque, C.A. Iglesias, An ensemble method for radicalization and hate speech detection online empowered by sentic computing, Cognit. Comput. 14 (1) (2022) 48–61.

[100] Y. Du, T. Li, M.S. Pathan, H.K. Teklehaimanot, Z. Yang, An effective sarcasm detection approach based on sentimental context and individual expression habits, Cognit. Comput. 14 (1) (2022) 78–90.