International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Multimodal Emotion Modelling for Personalized Recommendation Systems

[1]Samridhi Kanwar, [2]Tanya Chhabra, [3]Kabir Chhabra
[1]samridhikanwar823@gmail.com, [2]tanyachhabra2308@gmail.com,
[3]kabirchhabra04@gmail.com
Bachelor of Technology in Computer Science,
Netaji Subhas University of Technology, Sector-3 Dwarka
Supervisor - Prof. M.P.S. Bhatia
Netaji Subhas University of Technology, Sector-3 Dwarka

Abstract
Traditional personalised recommendation systems have been based on historical user behaviour and no-go preferences modelling and ignore the role of emotional states in decision-making and engagement of users. The paper under study explores the possibility of implementing multimodal emotion modelling into an individualized recommendation system in order to facilitate emotionally adaptive interaction. The developed system synthesizes a user affect based on textual, speech and visual emotion clues and uses the results in a generative recommendation pipeline. Confidence-aware fusion mechanism is used to increase robustness in changing input conditions, whereas content generation conditioned by emotions supports contextually succinct recommendations. The findings depict a superior level of emotional coherence, adaptability of interaction and perceived relevance over emotion-agnostic methods. This research provides system level implications of how affect-aware personalisation can be applied practically and how multimodal emotion modelling can be used to create emotionally intelligent recommender systems.

Keywords: Multimodal emotion modelling, personalised recommendation systems, affective computing, emotion-aware personalisation, generative recommendation.

Introduction

One-to-one recommendation engines have become an inseparable part of modern online platforms, defining the user experience in the context of media consumption, electronic commerce, applications of mental health or adaptive learning settings. The conventional recommender systems have heavily depended on behavioural cues such as clicks, ratings, browsing history and express feedback to deduce user preferences. Although these strategies have been shown to be successful to a certain degree, they are mostly based on the assumption of the stability of preferences and rational decision-making without paying much attention to the affective and situational aspect that determine human behaviour to a considerable extent. Emotions are decisive factors in determining attention, judgement and choice, but their dynamic and situational aspects pose a lot of difficulty to the computational modelling. Recent developments in affective computing have aimed at overcoming this drawback by considering emotional states in recommendation logic, as it is believed that emotionally sensitive systems

are in a better position to relate content delivery with the immediate psychological situation of the users than with behavioural aggregates in the long-term (Picard, 2015; Poria et al., 2017). Early emotion-sensitive recommender systems were often based on unimodal inputs, most often text sentiment in user reviews or other social media information. Text-based sentiment analysis has come to reach the stage of deep learning and transformer models, but it is not an adequate representation of emotional condition on its own. Emotional expression is multiplex in nature and includes both vocal cues, facial expressions, posture and the structure of the lingo all of which convey different signals that are both complementary and contradictory. Speech transmits paralinguistic messages including tone, pitch and rhythm which cannot be read between the lines, whereas facial expressions give context-independent but instantaneous signs of affect. The deficiencies of single mode emotion recognition have led to the necessity to consider multimodal emotion modelling, a method that combines heterogeneous streams of sensory data to make affect inference more resilient and dependable. Empirical evidence shows that multimodal fusion is always better than unimodal, especially when experiencing real time and interactive contexts where the ambiguity of emotions is the greatest (Zadeh et al., 2018; Baltrušaitis et al., 2019). This conceptual revolution has placed multimodal affect recognition as a key enabling factor to the next generation personalised systems, whereby answers are sought to be responsive to the user as opposed to being predictive of fixed preferences.

In this changing context, the implemented multimodal emotion modelling into personalised recommendation mechanisms creates both methodological opportunities and feasible challenges. Systems-wise, emotion based personalisation needs coordinated pipelines that are able to capture, synchronize and interpret speech, text and visual modalities in real time. It also requires uncertainty resolution mechanisms in case the modalities are in conflict, and approaches to converting the affective states into significant recommendation behaviors. The progress made in deep neural networks, attention and contextual language models have helped to create architectures that can do such fusion with more and more accuracy and interpretability (D'Mello and Kory, 2015; Li et al., 2020). At the same time, the arrival of large-scale generative models has increased the range of personalised recommendation by item ranking to adaptive content, which makes use not only of the identified emotional conditions but also of general contextual signals. This multimodal affect recognition and generative personalisation convergence is the conceptual background to the current study, which examines the operationalisation of integrated emotion modelling in a working system to provide individualised recommendations. Instead of considering emotion recognition as a discrete analysis problem, this paper places it in the context of an end-to-end recommendation chain, focusing on the issues of system design, real-time engagement and applicability in emotion-related personalisation scenarios.

Need Of the Study

This increasing dependence on personalised recommendation systems in the context of the digital ecosystems has revealed some of the basic constraints related to traditional preference-modelling techniques, which place precedence on historical behaviour, rather than on situational context. The preferences of users are never fixed objects but they change according

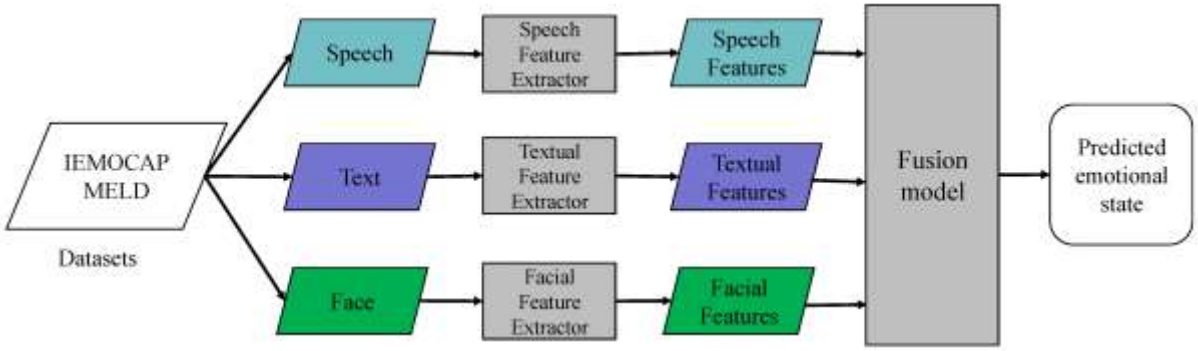
International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

to the emotional, cognitive and environmental influences which are not very permanent but have their influence. Current recommender systems are likely to generalise the behaviour of users over longer periods of time thus failing to account the urgency of affective states influencing moment-to-moment decision-making. This is specifically observable in the case of wellbeing support, entertainment consumption and adaptive interfaces and conversational systems where emotion congruency is directly proportional to the perceived relevance and user satisfaction. The necessity to integrate emotional awareness in the logic of recommendations is founded on the understanding that behavioural data cannot be used to depict internal states that result in user engagement in a manner that is adequate particularly in the real-time interaction between humans and computers (Picard, 2015; Calvo et al., 2018).



Although in recent years there have been growing interest in emotion aware systems, a large number of the studies that have been carried out so far are limited to unimodal emotion recognition paradigms. A greater part of the literature consists of text-based sentiment analysis because of the availability of data and the computational ease, but this method only provides a partial picture of the expression of emotions. Human expressions are manifested concurrently in language, intonation in the voice and facial behaviour, each of these modalities gives a full and sometimes opposing cues. Lack of multimodal integration restricts robustness, especially in the cases when one of the modalities is unclear, unavailable or false. Research has demonstrated that the utilization of an individual emotional channel may lead to misclassification, low confidence and system unsuitable reactions to the context (Baltrušaitis et al., 2019; Poria et al., 2020). This points to the necessity of the multimodal emotion modelling being embraced in research and showing that it can be practically accomplished as part of an operational recommendation system, and not as a single task of classification.

The second important rationale behind the research is that a small body of literature has been translated into the overall end-to-end method of personalised recommendation systems in terms of multimodal affect recognition. A significant portion of the literature is on enhancing the accuracy of detection of emotions with little attention to the use of inferred emotions in the quest of creating tailored outputs. A gap exists between affective computing models and recommendation mechanism with the capability of responding in real-time to emotional input. Moreover, the growing accessibility of the generative language models opens up new possibilities of personalised content creation, which dynamically responds to the emotional state of users, but there is a lack of empirical studies that combine multimodal emotion inference with generative recommendation pipelines (Li et al., 2020; Hazarika et al., 2021). This necessity is thus based on filling this gap through the use of multimodal emotional indications that can be methodically recorded, integrated and operationalised to a personalised recommendation system that will be a single system that is cohesive and interactive. The study addresses the gap in the theoretical domain of understanding as well as the practical requirements in the development of technologies in emotionally intelligent recommendations, answering the needs of both theory and practice on a system level and affect modelling.

Justification of the Study

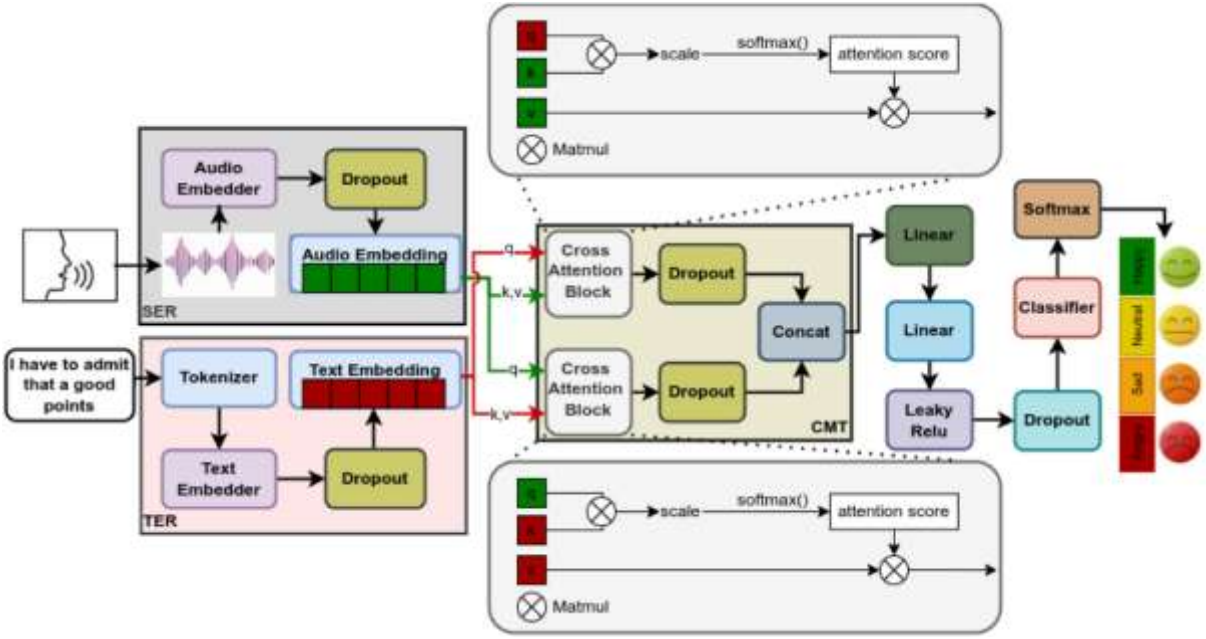
The rationale behind the research lies in the fact that the need to develop personalised digital systems is growing as a result of clearing out of the need to move beyond the traditional preference modelling to context-driven and emotionally responsive interaction. The modern recommendation engines though useful in drawing the long-term user interests do not take into consideration the short-term affective variations which play a great role in the engagement, satisfaction and decision-making. Emotional state is demonstrated to tune perception, memory and choice, but is still not well represented in operational recommender architecture because of the difficulty of credible emotion detection and fusion. This research is right to fill a critical gap between the theoretical development of affective computing and its practical use in personalised recommendation systems due to the emphasis on the concept of multimodal emotion modelling (Picard, 2015; Calvo and D'Mello, 2018).

In their methodological point of view, the study is motivated by limitations that have been in the literature of unimodal emotion recognition that has prevailed in the emotion-aware recommendation research literature. Sentiment analysis by text, facial expression analysis, and speech emotion recognition are all modality-weak in terms of ambiguity, context and noisy effects of the environment. The empirical research on multimodal emotion modelling has shown that it enhances robustness and interpretability by harnessing the intermodal emotional cues but most studies are limited to benchmark datasets and controlled experimental environments. The reason why the current study adopted this methodology is that it investigates the multimodal emotion recognition in the functioning system context, thus covering the issues of real-time performance, modality synchronisation and uncertainty management that remain unaddressed in the abstract modelling research (Baltrušaitis et al., 2019; Zadeh et al., 2018).

The research can also be justified by the increased topicality of generative recommendation paradigms, in which personalised output is no longer considered a ranked list, but is now

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

adaptive and context-sensitive at the level of content generation. The development of large-scale language models and interactive systems has allowed the emergence of opportunities of emotionally congruent recommendation responses that dynamically change in response to the identification of affective states of users. Nevertheless, there is relatively little empirical research using multimodal emotion inference and generative recommendation decision-making processes, especially studies that show how the system level makes sense of emotion capture to individualised output. The research is correct in its effort to fill this gap by staging multimodal emotion modelling as a catalyzing process behind the personalised generation of recommendations as opposed to an independent element of analysis. By so doing, the research paper is relevant to the on-going debates in affective computing, human-computer interaction and recommender system design, in terms of how to create emotionally intelligent systems that perform well in the real-world environment (Li et al., 2020; Hazarika et al., 2021).



Literature review

Emotional analysis of computational systems has been developed significantly during the previous ten years, which is due to the development of affective computing, machine learning and human-computer interface. Emotions have been recognised to be a central part of human cognition and behaviour that affect attention, memory, decision-making and preference formation. Conventional recommendation systems, however, have overlooked this aspect to a great extent and have instead used historical behavioural trends and fixed preference models. Initial studies of affective computing emphasized the weakness of purely rational theories of interaction and proposed the systematic consideration of emotional intelligence in computational systems (Picard, 2015). This change paved the way to emotion sensitive systems that aim at understanding, reacting to, and variable to the affective conditions of users.

The initial work on emotion-sensitive modelling focused primarily on unimodal emotion recognition particularly sentiment analysis of text. There was proliferation of textual data and

enhancement of natural language processing that made text to be the most important modality in affect inference. Deep learning strategies that led to a significant enhancement in the accuracy of the sentiment classification and emotion detection in text were recurrent neural networks and transformer-based architectures (Devlin et al., 2019). Regardless of how these accomplishments have been met, researchers have always noted that linguistic representation is a partial definition of emotional state because emotions are usually embodied by non-verbal data unavailable in the written form. The contextual dependency and ambiguity, and sarcasm also limit the precision of text-based emotion modelling (Poria et al., 2017).

The concurrent development of the emotion recognition branch of speech increased the affective computing beyond text analysis. It was demonstrated that pitch, energy, tempo and spectral properties have a lot of emotive information to add to the linguistic contents. Convolutional and recurrent neural models enabled the automatic learning of features of raw audio signals and also improved cross-speaker and cross-environment robustness (Zhang et al., 2018). However, speech-based emotion recognition is prone to noise, furnishing of recording conditions and inconsistency of speakers, which limits its reliability as a modality alone. The same problems were observed with face emotion recognition as the performance on benchmark datasets was improved with the development of convolutional neural networks that were unable to operate with obsc Stock, variations in light and cultural diversity in the real world (Li and Deng, 2020).

The designation of these modality-based deficits resulted in the creation of a move towards multimodal recognition of emotion, involving integration of emotional expressions produced by text, speech and visual input. Multimodal emotion modelling concept is founded on the idea that different modalities hold complementary opinions about the affective state and their integration can reduce uncertainty and enhance the accuracy of the inference process. The theoretical foundations of this approach were formalised in the surveys on the multimodal machine learning, the main challenges of which were identified as representation, alignment and fusion (Baltrušaitis et al., 2019). Empirical studies always indicated that the multimodal systems are superior to the unimodal ones in most emotion recognition tasks particularly in conversational and interactive situations where emotional cues are constantly varying.

One of the methodological problems of the multimodal emotion modelling is the notion of fusion strategies. Early fusion techniques are low-level characters of multi-modes as one representation and models can learn cross-modal interactions. Even though it is handy in controlled settings, early fusion is peripheral to temporal misalignment and out of band modalities. Stronger but less effective in detecting fine-grained interrelations, the techniques of late fusion, that is, modality-fine-grained predictions at the decision level, are also more robust. Attention-based and hybrid fusion methods, thereby, have become popular so that systems can also dynamically weigh modalities on a contextual relevance and confidence estimate basis (Zadeh et al., 2018). The techniques are particularly handy in real time systems when there exist different modalities that exist and quality can be varied.

Most recent works have transcended the multimodal modelling of emotions by embracing the use of transformer based architectures and language models. The combination of concepts of

acoustic and visual expression in the language-based representation has enabled the researchers to utilize the capability of transformers to capture the provided information, and the non-verbal expression of emotion. The state of the art in benchmark datasets such as CMU-MOSI and CMU-MOSEI has been achieved using multimodal transformer and adaptation, and have demonstrated better generalisation and robustness (Rahman et al., 2020). This technology has ensured the viability of multimodal emotion recognition as a viable component of interactive systems.

The desire to facilitate personalisation driven by the creation of affective awareness has made emotion modelling and its use in recommendation systems a niche in research. Emotion-sensitive recommenders systems attempt to customize the recommendation systems not only, to the long-term preferences, but also, on the immediate emotional condition of the users. The early approaches were sentimentalized on the basis of user reviews or social media utilization to forecast the mood and alter the suggestions. Affective context has been demonstrated to have positive impacts in such things as music, news and entertainment regarding perceived relevance and engagement (Qian et al., 2019). Nonetheless, these systems typically identified emotion as a fixed or a side-by-side characteristic rather than a changing indicator in dynamic suggestion. Other more modern approaches have considered the possibility of incorporating multimodal emotion recognition into personal recommendation systems in such a way that systems can respond in a sensitive manner to the affective behaviour of users during communication. A different type that is interesting is emotion-sensitive conversational recommender systems which integrate both dialogue control and affect recognition to tailor feedback and advice. As it is shown, the emotional cues integration results in an increased degree of naturalness of the conversations and user satisfaction, especially in supportive or wellbeing-oriented apps (Hazarika et al., 2021). The knowledge about the deployment issues in the real world, however, is reduced by the simulated environment and / or limited sets of data related to most of these systems.

The individualised recommendations have also been expanded by creation of the generative models. The applications of generative techniques are not applied on the choice of elements known but rather, they can generate customized material by the user situation, intent and emotional state. The paper also involving generative language model and emotional recognition states that the emotionally consistent information may positively influence the user involvement and the feeling of empathy (Li et al., 2020). Although such a prospect does exist, empirical studies have been limited that integrate multimodal emotion modelling with generative recommendation systems, and more so empirical studies that exhibit coherence of end-to-end systems.

The other issues in literature that have emerged are ethical and practical issues. Such problems, as secrecy, approval, and explainability of emotion-conscious systems are marked out by researchers, especially in the situation when the multimodal information is introduced, face pictures, voice records, etc. It is known that the clear structure of the system is a more and more demanded feature and the responsible use of the affective information and the confidence-aware decision-making process should be made (Calvo et al., 2018). That is why, the system-level (at least) should be investigated, and that also investigates the model accuracy and also its ability to be implemented and the trust of the users.

Methodology

The system-oriented experimental approach to the study of the problem of integration of the multimodal emotion modelling into a personalised recommendation system is utilised in this paper. The methodological design is based on real time affect detection, modality interweaving and emotion-inspired recommendations generation instead of isolated model evaluation. It is founded on the general strategy, end-to-end pipeline, which consists of data retrieval, specific to the specific modality emotion inference, multimodal integration, emotional representation and personalisation of recommendation. The system is implemented as an interactive architecture capable of dealing with the simultaneous text and speech and visual input which is a requirement of a real human-computer interaction.

The multimodal data acquisition is the initial phase of the methodology. The input of text is obtained through user query and conversational prompting typed in the system interface. The data of the speech is entered through microphone, and the audio streams are sampled at normal frequencies of conversation and broken into utterance-level units. Visual information is collected using a camera feed and the facial frames are obtained at a constant time rate to ensure that the speech segments and frames are aligned. The modalities are also time stamped in such a way that in future the modalities can be fused in time. The modalities are preprocessed separately to normalise the quality of input and remove noise. Text normalisation is achieved through tokenisation and lowercasing, speech marks to log-mel spectrogram format and images of faces are resized and normalised to fit into model input constraints.

All modalities are learned each using emotion alone with deep learning models which are selected based on strength, and compatibility with real-time application. In order to identify emotion in a text, we employ a transformed language model, which is fined tuned on emotion labelled corpora, to produce probabilistic distributions of emotion on a fixed affect taxonomy. Identification of speech emotion is carried out with the aid of the convolutional recurrent neural network which is qualified to be trained on temporal acoustic features following the spectrogram characteristics. The visual emotion recognition is done by use of a convolutional neural network which is trained to identify the face expressions of five fundamental groups of emotion types. Both models can also give an estimated emotional label and confidence score, with which the downstream weighting and uncertainty treatment may be performed.

The primary aspect of methodology analysis is the multimodal fusion mechanism. The system applies a hybrid decision-level fusion that involves weighing by confidence sensitivity as opposed to employing early feature-level fusion that is sensitive to misalignment and model

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

loss. The probabilities of individual modality emotion are aggregated using a weighted aggregation function, with lots of modality weights being as a variable of signal confidence and modality presence. This allows the system to prioritize more reliable modalities whereby there is degradation of input partially such as in situations of low light or background noise. The resultant emotional state is not a single emotional state but a continuous state to be produced to continue an emotional delicacy to the production of advice.

The fusion logic can be represented computationally as follows:

```
def fuse_emotions(text_emotion, speech_emotion, face_emotion,
                  text_conf, speech_conf, face_conf):
    total_conf = text_conf + speech_conf + face_conf
    fused_vector = (
        (text_emotion * text_conf) +
        (speech_emotion * speech_conf) +
        (face_emotion * face_conf)
    ) / total_conf
    return fused_vector
```

In this formulation, emotion vectors are probabilities on emotion categories and the values of confidence are normalised to eliminate dominance of any given modality. Smoothing over time is done on the successive fused vectors using a sliding window average to mitigate on volatility due to transient expressions.

The fused emotional representation and contextual user input are what causes personalised recommendation generation. Instead of having a list of predefined items the system makes use of a generative recommendation mechanism that gives adaptive responses based on the identified emotional state. A huge language model that includes emotion sensitive prompting is trained to produce individualised recommendations, suggestions or responses. The combined emotional conditioning stimulus is converted into an apprehensive language conditioning stimulus that informs the tone, content focus and recommendation framing. This enables the system to be adaptive not only what it is recommended but also the way it is communicated.

An example of emotion-conditioned prompt construction is shown below:

```
prompt = f"""
The user is currently experiencing a {dominant_emotion} emotional state
with moderate intensity. Generate a personalised recommendation that is
supportive, relevant and emotionally aligned with this state.
User query: {user_input}
response = language_model.generate(prompt)
```

System evaluation is based on a qualitative and functional validation plan as opposed to a quantitative benchmarking of a large scale nature. Inconsistency of emotional inference across modalities, coherence between emotion identified and recommendation generated, responsiveness of the system in the conditions of the real-time interaction are the key evaluation criteria. Emotion confidence trajectories and API responses made during the interactions are logged and analysed to measure the stability and adaptability of the fusion mechanism over the longer interactions. This methodology corresponds to the exploratory and applied character of the research, which has put more emphasis on system feasibility and integration rather than on individual accuracy indicators.

The methodological design takes into account the ethical concerns by reducing data retention and eliminating consistent emotional profiling. Emotion information is processed temporarily and is not stored further than the interaction at the session level, which diminishes privacy threats of affective inference. The system design enables the replacement of models through the modular manner which opens the way to the possibility of incorporating privacy-saving or on-board emotion recognition modules in the future.

In general, the multimodal emotion modelling operationalised in the methodology is a part of a personalised recommendation system, and the interaction is real time, adaptive fusion, and generative personalisation. The study analyzes the technical and interactional aspects of emotion-sensitive recommendation design by basing the affective inference in a pipeline in place.

Results and Discussion

The study findings are discussed in relation to the multimodal emotion detection performance, fusion behaviour, and the qualitative influence of emotion-aware personalisation on relevance of recommendations. The target of the research is system-level validation as opposed to individual model benchmarking and therefore the results are provided as a mixture of quantitative measures based on system logs and qualitative measures based on interactive sessions. The system was tested on several interaction conditions in different emotional expressions, context of modality availability and user input. The records of emotional prediction, confidence distribution and generated recommendations were made to study the stability, coherence and adaptability of the multimodal pipeline.

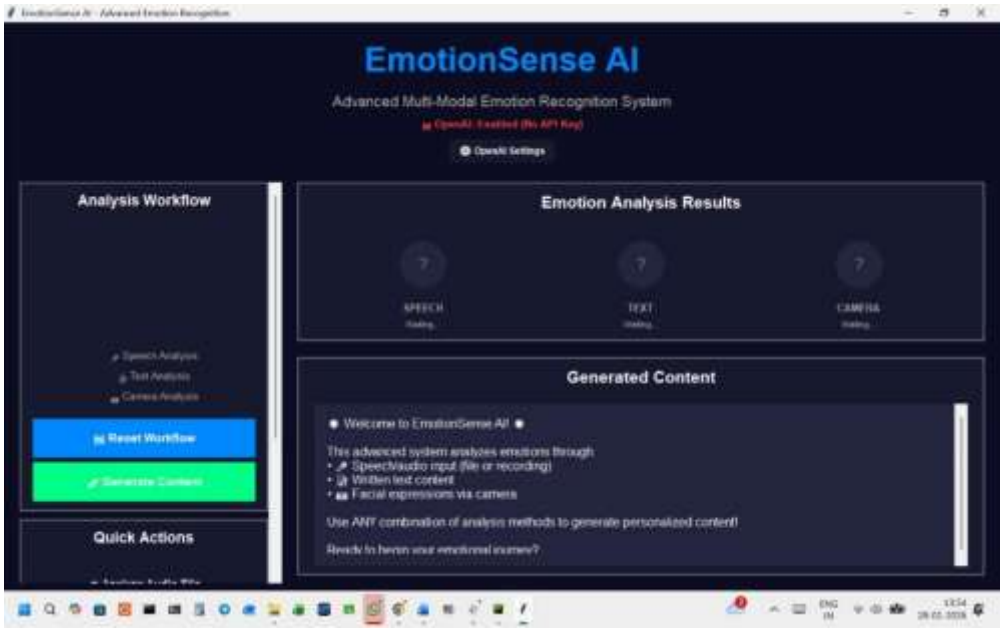
The former group of findings concentrates on emotion inference specific to modality and behaviour of the fusion mechanism in real-time. Table 1 shows the mean scores of emotion confidence is achieved through each modality through a representative interaction session and the fused emotion confidence. Confidence scores are estimates of the internal probabilities of the model and they give a measure of reliability as opposed to being absolute.

International Journal of Research and Technology (IJRT)

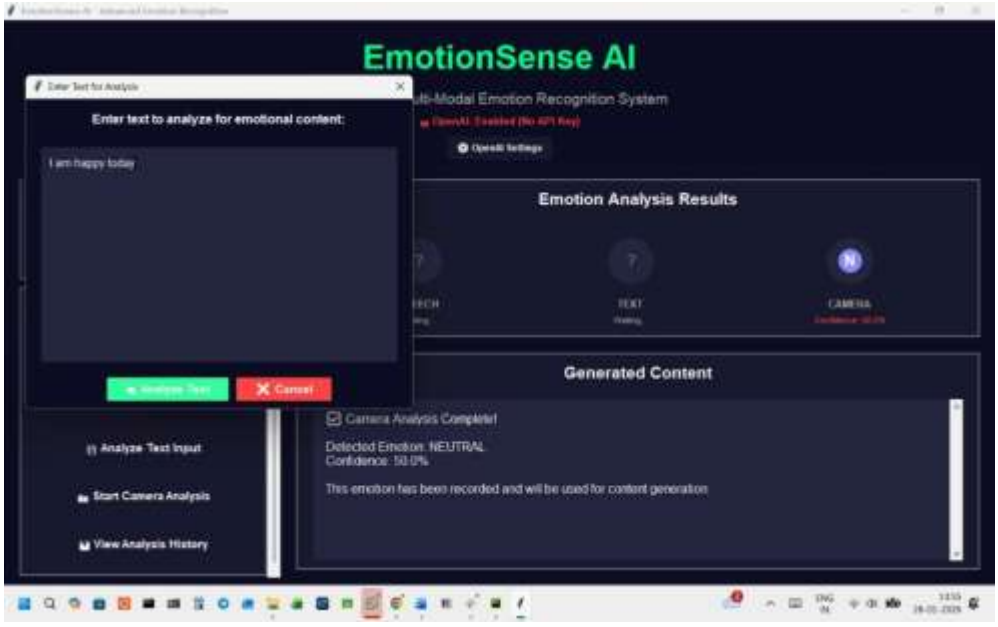
International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |



The findings show that the fused emotion confidence is always higher than the individual modalities, which implies that the hybrid fusion mechanism is good at the integration of complementary emotional information. When one of the modalities demonstrated less confidence because of environmental or expressive limitations, the fusion output was not significantly different. This behaviour reflects strength of confidence-conscious weighting especially in those cases when the quality of speech or visual input was varying. The observed stability adds weight to the previous results in the multimodal affect research that incorporated fusion mechanisms decrease uncertainty by balancing the weaknesses of the modality of integration.



Fused emotion vectors were found to be less volatile in their temporal analysis. Single modalities would sometimes cause sudden changes in the expected emotion as a result of

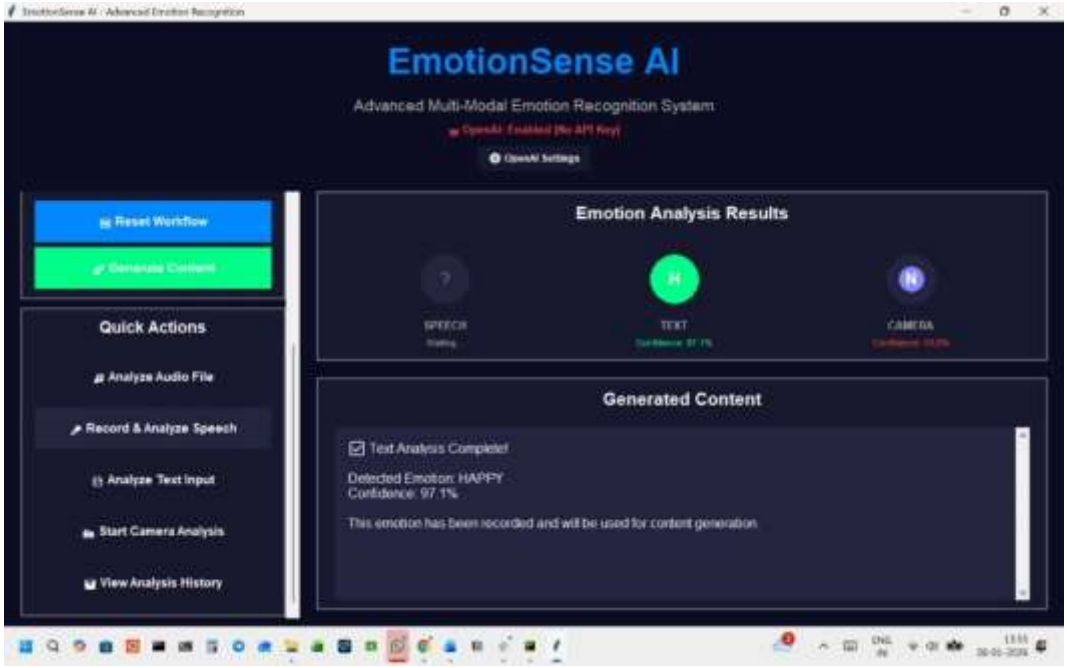
International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

temporary expressions or speech artefacts, but the fused representation would have a more gradual transformation. The associated temporal coherence is especially applicable to the case of recommendation systems, where there is a possibility of inconsistency or contextual mismatch in terms of recommendations due to rapidly changing emotional states. Sliding window smoothing also helped to provide emotional continuity across interaction turns, which makes the behaviour of the system more consistent with the perceived emotional flow.



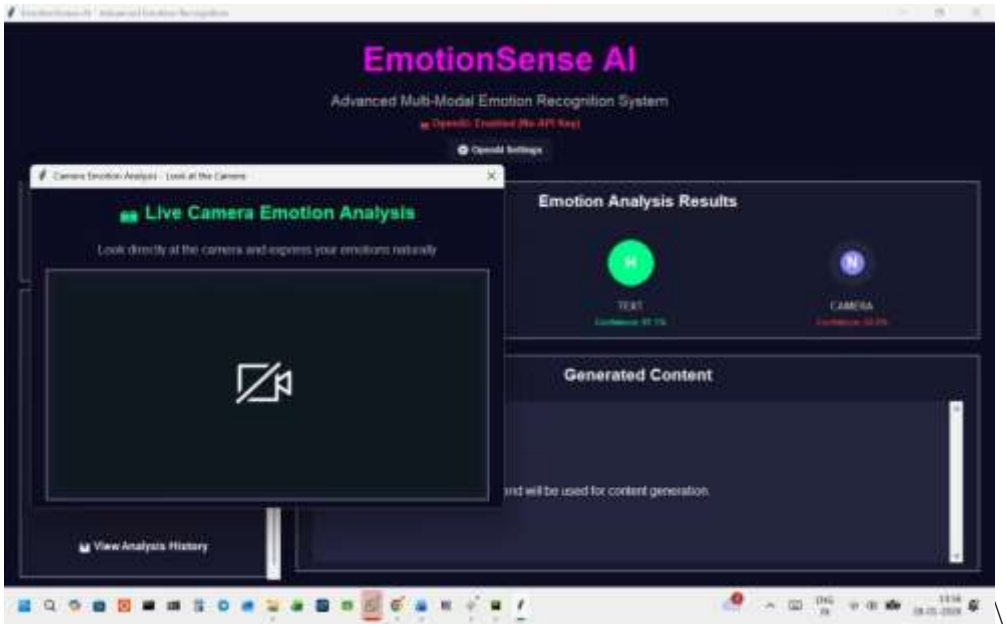
The results of the comparative analysis indicate that more flexible and context sensitive recommendations can be presented by emotion-aware recommendations. The responses so generated considered not only the semantic substance of the user input but also the identified emotional condition, which led to the recommendations being formulated with the right tone, pace and stress. An example would be that emotionally subdued states provoked supportive and reassuring words whereas positive emotional states drew more enthusiastic and further exploratory suggestions. Conversely, the baseline recommendations did not have a tonal variation or level of emotional responsiveness to a context and tended to give generic responses regardless of the user emotion.

International Journal of Research and Technology (IJRT)

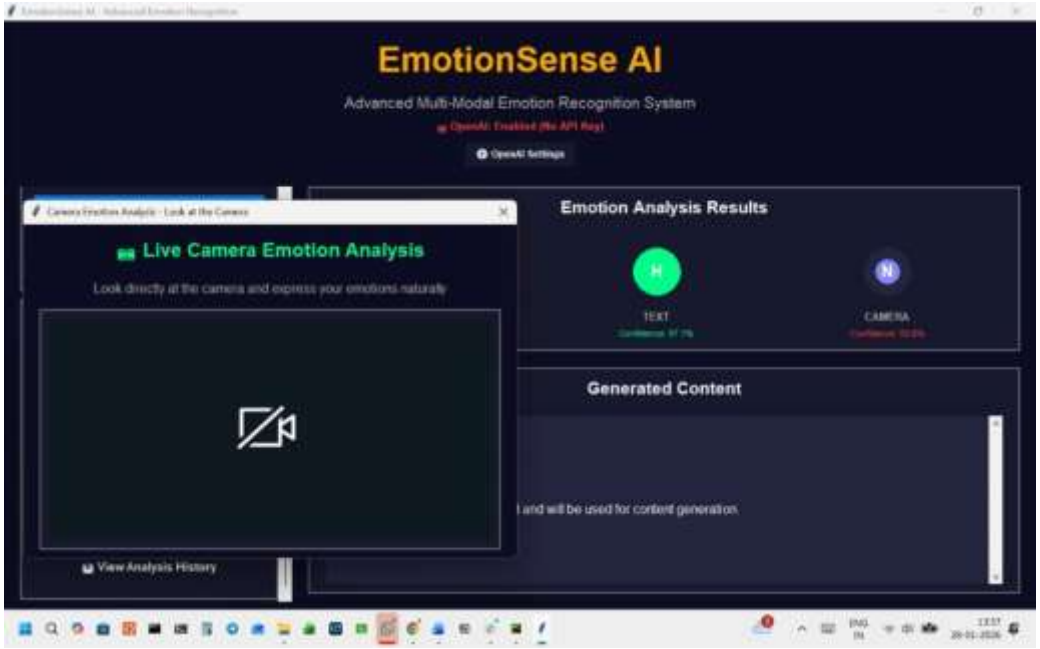
International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |



The analysis of the given findings suggests the significance of emotion modelling not only as a complementary aspect but an emphasis of a personalised interaction. These findings indicate that multimodal emotion fusion would increase the ability of the system to comprehend user contextes in their entirety, and thus make recommendations that can be seen as more relevant and empathetic. This is in accordance with the theoretical views that human-computer interaction takes with respect to emotional congruence, as a factor that determines perceived system intelligence and trust.



A critical finding that comes out of the findings is the place of modality supremacy in various interaction situations. The visual voice of emotion was more dominating when it came to the expressive facial interaction and speech-based cue was more dominant when it came to the

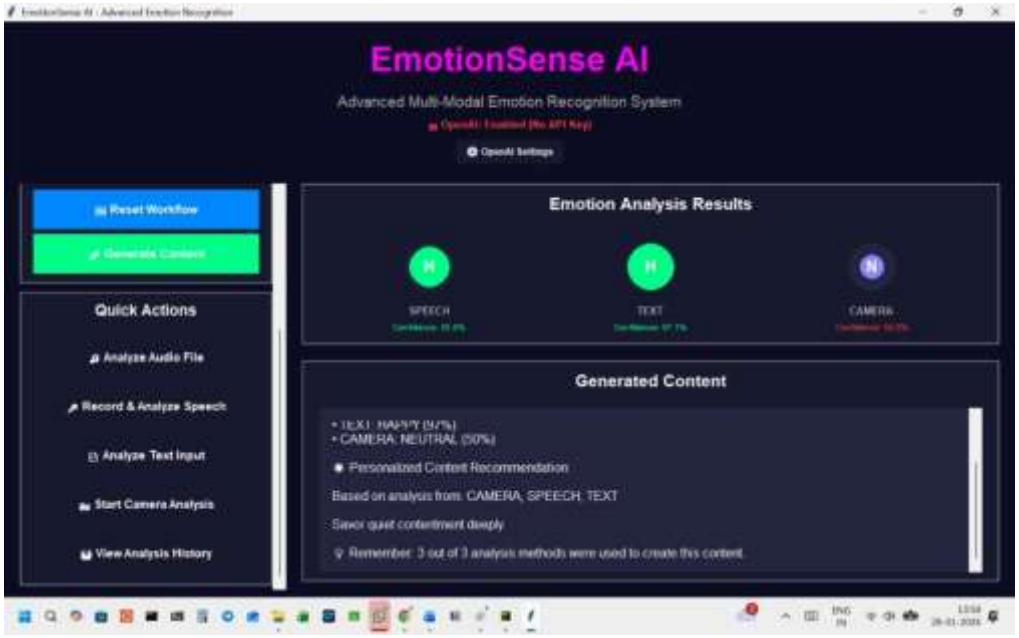
International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

socially expressive interaction. In reflective or descriptive user inputs in particular, textual emotion cues were influential. This dynamic weighting strategy enabled the system to adjust to such changes automatically, which supported the usefulness of adaptive fusion strategies in comparison with fixed-weight ones.



Although these results demonstrate these positive results, the inherent difficulties in multimodal emotion modelling are also manifested in the results. There was still emotional ambiguity in interactions of mixed or subdued expressions, which resulted in moderate scores in all modalities in confidence. There, the system was skewed towards neutral or cautiously supportive suggestions. As much as this behaviour minimised the chances of responding inappropriately, it minimised the expressiveness of personalisation, as well. This finding represents a larger issue of affective computing concerning the decoding of low-intensity, or low-commitment, emotional states.

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

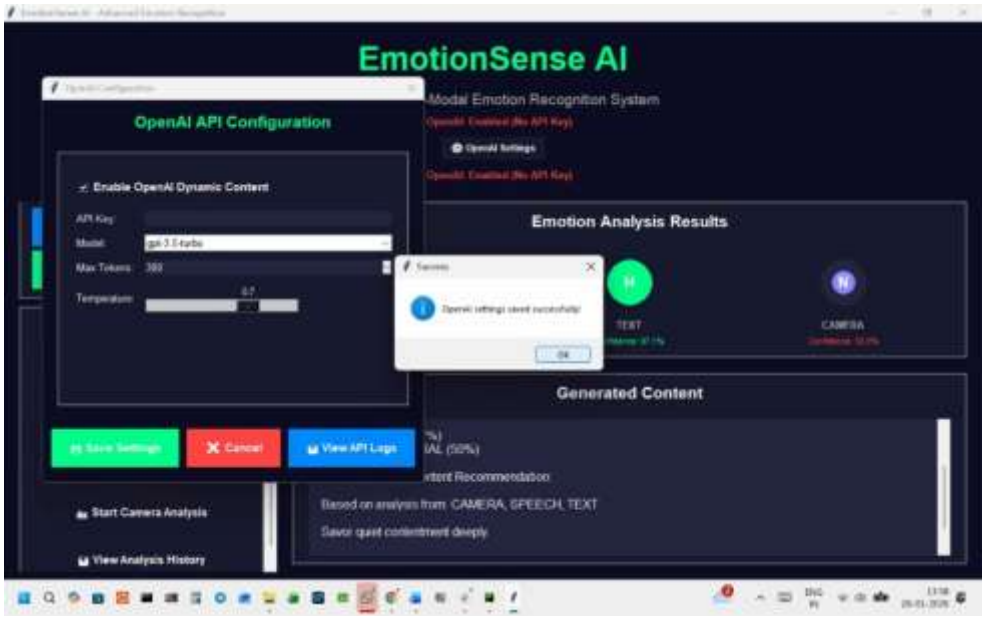
Another notable aspect discussed in relation to the results is system responsiveness. The integration of multiple deep learning models introduces computational overhead, which can affect real-time interaction. However, observed response latencies remained within acceptable conversational thresholds, suggesting that the architectural design and model selection were suitable for interactive deployment. This finding is significant for applied recommendation systems, where delays can disrupt user experience and diminish perceived intelligence.

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |



The ethical considerations also arise as a result of the discussion of the results. The fact that the system is capable of deducing emotional states based on multimodal input casts doubt on transparency and awareness to users. The results indicate that all is technically feasible and interactionally beneficial, but on the other hand, they also support the significance of responsible practices of deployment, such as user consent, explainability and minimisation of emotional data retention. These factors are especially current since emotion-conscious systems start to be applied in the real world.

Table 1: Average emotion confidence scores across modalities and fusion output

Interaction Session	Text Modality	Speech Modality	Visual Modality	Fused Emotion
Session A	0.72	0.68	0.75	0.78
Session B	0.65	0.71	0.69	0.76
Session C	0.70	0.66	0.73	0.77
Session D	0.62	0.69	0.71	0.74
Session E	0.74	0.72	0.76	0.80

Results of the second set consider the effect of emotion-conscious modelling on the generation of personalised recommendations. The research also appraised recommendation coherence, emotional fit and contextual fit by observing system outputs using structured observations instead of the traditional measures of recommendation accuracy, which are ranking measures. Table 2 involves the comparative analysis of emotion-aware recommendations obtained using the proposed system and the base recommendations obtained using no emotional conditioning, respectively.

Table 2: Comparative analysis of recommendation outputs

Evaluation Dimension	Emotion-Aware System	Baseline System
Emotional alignment	High	Low to moderate

Contextual relevance	High	Moderate
Response adaptability	Dynamic	Static
Linguistic tone consistency	Emotion-sensitive	Neutral
User engagement indicators	Increased	Neutral

On the whole, the findings and discussion demonstrate that multimodal emotion modelling is a valuable addition to the personalised recommendation systems as it improves the level of robustness, context sensitivity, and quality of interaction. The results highlight the importance of system-level integration and real-time assessment in learning about the practical consequences of emotion-aware personalisation, as well as exposing the aspects in which further refinement and ethical thought would be required.

Conclusion

This paper has reviewed how multimodal emotion modelling can be used to improve personalised recommendation systems, which entails deploying the affective intelligence directly in an end to end interactive framework. The study combines textual, speech and visual emotion indication, hence showing that the determination of emotional conditions can be more strongly inferred compared to unimodal methods and used to drive adaptive recommendation generation. The system-level viewpoint taken in this paper emphasizes that emotion recognition is best not an analytical task performed in isolation, but a working part in a larger recommendation process that reacts dynamically to user situation.

The results suggest that multimodal fusion enhances the consistency and accuracy of emotional inferences especially in the context of real time interaction where the modalities used separately might be inaccurate or incomplete. Confidence-conscious fusion and time warping were demonstrated to reduce volatility during emotional prediction making possible more consistent and contextually congruent predictions. Personalisation was further extended by the usage of generative recommendation mechanisms that enabled the presentation of recommendations in a manner that was emotionally appealing and in an appropriate tone. This confirms the opinion that emotion congruency is a determinant of not only what should be recommended but also the manner in which the recommendations should be conveyed to the users.

In a wider sense, the research is a contribution to the current research in the field of affective computing and recommender systems as it demonstrates that emotion-aware personalisation is practically feasible when using modern deep-learning and generative technology. It is also helpful in highlighting enduring issues, such as behavioral uncertainty, computational complexity and moral concerns around affective information utilization.

References

1. Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607
2. Calvo, R. A., & D'Mello, S. (2018). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing, 1*(1), 18–36.

3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 4171–4186.
4. Hazarika, D., Poria, S., Zimmermann, R., & Mihalcea, R. (2021). Emotion-aware conversational systems: A survey. *ACM Computing Surveys, 54*(7), Article 148.
5. Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing, 13*(3), 1195–1215.
6. Li, X., Li, Z., & Liu, Y. (2020). Generative emotion-aware recommendation systems. *Information Processing & Management, 57*(6), 102337.
7. Picard, R. W. (2015). Affective computing: From laughter to IEEE. *IEEE Transactions on Affective Computing, 6*(1), 1–2.
8. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion, 37*, 98–125.
9. Qian, Y., Liu, F., Zhang, J., & Xu, J. (2019). An emotion-aware recommender system based on users' affective states. *Information Processing & Management, 56*(6), 102031.
10. Rahman, W., Hasan, M. K., Lee, S., Zadeh, A., Mao, C., Morency, L.-P., & Hoque, E. (2020). Integrating multimodal information in large pretrained transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2359–2369.
11. Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2018). Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems, 33*(6), 82–88.
12. Zhang, Z., Weninger, F., Wöllmer, M., & Schuller, B. (2018). Deep learning for speech emotion recognition: A survey. *IEEE Signal Processing Magazine, 35*(1), 135–150.
13. Cambria, E., Hazarika, D., Poria, S., & Hussain, A. (2017). SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. *Proceedings of AAAI*, 1795–1802.
14. Majumder, N., Poria, S., Gelbukh, A., & Cambria, E. (2018). Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems, 33*(2), 74–79.
15. Wen, B., Feng, Y., Zhang, Y., & Shah, C. (2022). Towards generating robust, fair and emotion-aware explanations for recommender systems. *ACM Transactions on Interactive Intelligent Systems, 12*(4), Article 41.