International Journal of Business Analytics and Intelligence
11 (2) 2023, 01-06
http://publishingindia.com/ijbai/

Unraveling Emotions: Multimodal Deep Learning for Fine-Grained Emotion Recognition

Riktesh Srivastava*, Rajita Srivastava**

Abstract

In the landscape of natural language processing and artificial intelligence, sentiment analysis and emotion recognition hold crucial roles in deciphering human emotions across diverse communication channels. This study addresses a significant research gap by venturing into the promising domain of emotion identification through sentiment analysis, capitalising on the potential of multimodal deep learning. Through an exhaustive literature review, the study seeks to bridge the gap between conventional sentiment analysis methods and the intricate subtleties of human emotions, achieved by fusing data from various modalities. This integration, coupled with fine-grained emotion recognition, strives to heighten the precision of emotion comprehension. Anchored by a robust conceptual framework encompassing variables like information integration, information variability and model type, this research probes the interplay of these factors in shaping emotion consistency. The research methodology involves a two-tailed t-test, a potent statistical tool for hypothesis testing using SmartPLS. The outcomes shed light on the intricate interplay among these variables. While information integration may not exert a significant impact on information consistency, information variability and model type surface as critical factors, each distinctively contributing to the enhancement of information consistency. These insights offer a deeper comprehension of the complexities within this domain, charting a path towards refined insights into the examined relationships.

Keywords: Multimodal Deep Learning, Emotion Recognition, Sentiment Analysis, Fine-Grained Emotions, Information Integration, Model Type

Introduction

In recent years, sentiment analysis and emotion recognition have emerged as crucial areas in natural language processing and artificial intelligence, enabling machines to comprehend and respond to human emotions expressed through various modes of communication. The advent of social media, online reviews and user-generated content has accentuated the importance of understanding emotions in text, images, audio and videos. To address the complexities of emotions expressed across multiple modalities, this research aims to explore the promising domain of identifying the emotions through sentiment analysis.

The research offers several significant advantages. First, by combining the power of multimodal deep learning and fine-grained emotion recognition, the research endeavors to bridge the gap between traditional sentiment analysis methods and the intricate nuances of human emotions. This fusion of diverse modalities will enable more accurate and comprehensive emotion recognition, enhancing the depth of understanding in natural language processing systems. Secondly, this research opens avenues for a broader range of applications. Emotion recognition in a multimodal setting has tremendous potential in areas such as affective computing, customer feedback analysis, market research, mental health support and human-computer interaction. Accurately interpreting emotions from multiple sources could lead to more empathetic chatbots, better customer experience management and even contribute to mental health assessments through the analysis of text, audio and visual content. The research also holds implications for both academic and industrial contexts. Advancements

* College of Business, City University Ajman, UAE. Email: r.srivastava@cu.ac.ae; rajitariktesh@gmail.com
** PhD (Management), Banasthali Vidyapith, Rajasthan, India.

in multimodal deep learning and emotion recognition will contribute to the growing body of knowledge in the field of artificial intelligence. Moreover, industries can leverage the findings to build sophisticated sentiment analysis tools, sentiment-aware virtual assistants and personalised recommendation systems that cater to individual emotional preferences.

The urgency for this research stems from the critical challenges faced in current sentiment analysis and emotion recognition models. Existing approaches primarily focus on unimodal data, often overlooking the richness of emotions conveyed through various channels. Traditional methods fail to grasp the contextual intricacies that come from combining text, images, audio and videos, leading to limited accuracy and practicality in real-world applications.

Moreover, in the era of big data and social media, there is an explosion of multimodal content, making it more pressing to develop efficient methods capable of handling large-scale, diverse data sources. Users are increasingly expressing their emotions through mixed-media content, and an advanced multimodal deep learning approach is necessary to unlock the potential of this vast emotional data mine.

This research is organised into five sections to achieve the stated objectives. In Section 2, a comprehensive literature review will be conducted to identify existing methodologies, techniques and challenges in sentiment analysis, emotion recognition and multimodal deep learning. This review will help identify the research gap and set clear research objectives for the current study. Section 3 will elaborate on the research methodology adopted for this study. It will outline the techniques used for multimodal deep learning and fine-grained emotion recognition, as well as the frameworks employed to fuse and analyse data from diverse modalities effectively. Section 4 will delve into the data collection and analysis process. The selection and preparation of the multimodal dataset will be described, along with the pre-processing steps for each modality. Furthermore, the data analysis techniques and model evaluation metrics will be detailed to ensure transparency and replicability of the research findings. Finally, Section 5 will present the recommendations based on the research questions and hypothesis statements. This section will provide insights into the potential applications, limitations and future directions for multimodal emotion recognition and its impact on various industries and domains.

Literature Review

The recognition and understanding of human emotions have been central topics in the fields of psychology, cognitive science and artificial intelligence. The ability to accurately discern fine-grained emotions from various modalities, such as facial expressions, vocal tones and body language, is essential for business as well. In recent years, the advent of deep learning techniques has significantly advanced the field of emotion recognition, allowing for improved performance in discerning subtle emotional nuances. This segment of the literature review introduces an investigation into the realm of multimodal deep learning for fine-grained emotion recognition. The fusion and comprehensive analysis of disparate sensory inputs within multimodal deep learning serve to augment fine-grained emotion recognition, thereby capturing intricate emotional cues to facilitate more precise and nuanced emotion classification.

Multimodal Deep Learning for Emotion Recognition

Multimodal emotion recognition, characterised by the integration of information from multiple sources, has emerged as a promising avenue to enhance emotion recognition accuracy. This approach involves combining facial expressions, audio cues and textual data to achieve a more comprehensive understanding of emotional states. Pioneering work by (Zeng et al., 2015) introduced the concept of using convolutional neural networks (CNNs) for the joint analysis of facial expressions and speech signals, demonstrating improved emotion classification accuracy through the fusion of visual and auditory cues. This early integration of modalities marked a significant step toward the development of robust emotion recognition systems. Subsequently (Nguyen et al., 2017) proposed a novel framework that further extended multimodal emotion recognition by combining facial expressions, physiological signals and textual context. Their deep recurrent neural network (RNN) model effectively captured temporal dependencies and

contextual information, resulting in enhanced performance in recognising nuanced emotions in real-world scenarios. As multimodal deep learning gained momentum, recent studies explored the integration of visual, auditory, and textual cues to gain unique insights into an individual's emotional state (Li et al., 2018, Zhang et al., 2018). (Gao et al., 2019) introduced a multimodal attention-based model that effectively combined visual and textual information to enhance emotion recognition. By integrating textual descriptions of emotional contexts with visual cues, their deep learning architecture harnessed the synergy between modalities, leading to improved emotion recognition accuracy. Despite these successes, challenges persist in multimodal emotion recognition due to noise or inconsistency, necessitating careful preprocessing and feature extraction. (Xu et al., 2020) used a multi-stream fusion approach which employed convolutional and RNNs to jointly model visual and auditory information for emotion recognition.

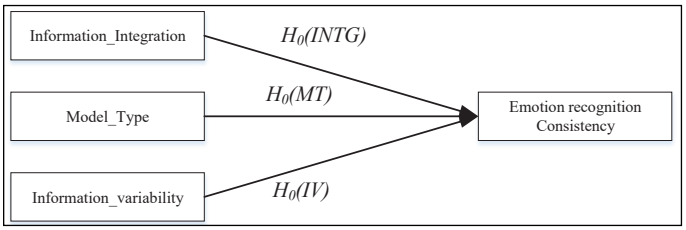
Use of Multimodal Deep Learning for Fine-Grained Emotions

In recent years, there has been a growing interest in leveraging multimodal deep learning approaches for enhancing fine-grained emotion recognition. Multimodal deep learning techniques have emerged as a powerful tool for capturing intricate emotional nuances by integrating information from multiple sources such as facial expressions, voice tonality and physiological signals. Researchers have demonstrated the effectiveness of these approaches in deciphering subtle variations in emotions that traditional unimodal methods often struggle to capture (Akhtar & Madani, 2019). For instance, the fusion of visual and auditory modalities has been shown to yield superior results in distinguishing closely related emotions such as "happy" and "excited" (Zhong & Zhang, 2020). Moreover, the utilisation of deep neural networks, such as CNNs and RNNs, in conjunction with attention mechanisms has facilitated the extraction of intricate features from multimodal data, leading to improved fine-grained emotion classification (Chen & Ji, 2017; Huang et al., 2019). Additionally, the incorporation of physiological signals, such as electroencephalography and heart rate, has further enhanced the discriminative power of multimodal models by capturing subtle physiological responses associated with specific emotional states (Lin et al., 2018). Overall, the integration of multimodal deep learning techniques has proven to be instrumental in advancing the field of fine-grained emotion recognition, enabling a more nuanced understanding of human affective states.

The thorough literature evaluation reveals that multimodal deep learning has great potential for precise emotion recognition, but it is unclear how these techniques can be used to address practical issues and fulfill business demands. Thus, the primary research objective is to explore and develop practical applications of multimodal deep learning techniques using sentiment analysis thus bridging the gap between theoretical and business requirements.

Research Methodology

The primary data for this study was collected using a survey approach. The survey included a questionnaire that utilised a five-point Likert scale and consisted of five items. These items were categorised into three distinct variables. Among these variables, three were independent variables: Integration of Different Types of Information (INTG) with two items, Model Type (MT) with one item and Information Variability (IV) with one item. The dependent variable was Emotion Recognition Consistency (ERC), represented by one item. The conceptual framework depicting these variables is presented in Fig. 1.



Source: Author.

Fig. 1: Conceptual Framework

The conceptual framework proposed for the research is presented in Table 1, which outlines the research questions and hypothesis statements. This framework serves as a comprehensive guide for investigating the relationship between INTG, MT and IV with ERC. By formulating these research questions and hypothesis

statements, the study aims to explore the extent to which multimodal deep learning can be used to identify the fine-grained emotions using different mediums. The null hypotheses associated with each research question will be tested to ascertain the presence or absence of statistically significant relationships.

Table 1: Research Questions and Hypothesis Statements for the Proposed Conceptual Framework

Variable Name	*Research Questions*	*Hypothesis Statements*
Integration of Different Types of Information (INTG)	INTG1: Using a combination of text, images, and sounds helps me understand emotions more accurately.	*H_0(INTG): There is significant relationship between integration of different types of information and Information consistency.*
	INTG2: Integrating text, images, and sounds into emotion recognition models is a valuable approach to capture a wide range of emotional cues.	
Model Type (MT)	MT1: Models using multiple types of information have a better grasp of the diverse ways emotions can be expressed.	*H_0(MT): There is significant relationship between different models of social media and Information consistency.*
Information Variability (IV)	IV1: Taking into account different types of information enhances the ability of emotion recognition models to work well in a wide range of contexts.	*H_0(IV): There is significant relationship between different information variability and Information consistency.*

Sample Size

To gather data, stratified sampling method is employed. This method is valuable because it helps us focus on a specific and accurate group of people for the research. The aim is to conduct a survey among customers who've been using social media for their purchase decisions. We're particularly interested in those who have used various social media tools to read reviews before making their purchases. Through stratified sampling, we'll divide the customers into smaller groups based on how they've been using these digital tools. Our research involves a sample size of 100 people who consistently engage with reviews from different social media tools, whether it's reading text, watching videos, or any other form of content.

Data Analysis

To examine the validity of the hypothesis statements, a two-tailed *t*-test was employed for the data analysis. Before conducting the *t*-test, the model's factor loadings were evaluated using SmartPLS. The findings indicated that all factors related to the independent variable showed significance, with factor loadings surpassing 0.70 (except for INTG2, which registered a factor loading of 0.653). Notwithstanding this, INTG2 was retained within the model due to the fact that removing it did not lead to a substantial enhancement in the performance of the other factors. The details of factor loadings are mentioned in Fig. 2.

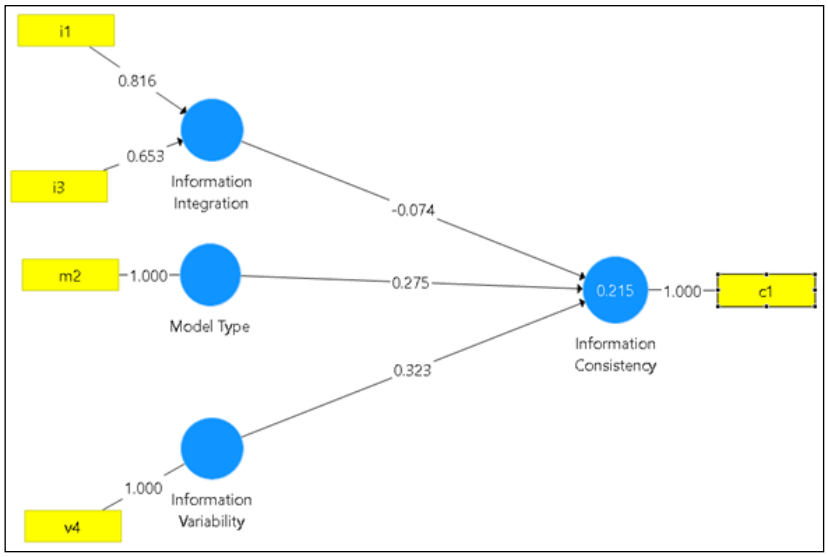


Fig. 2: Factor Loading Values for the Proposed Conceptual Framework

The two-tailed *t*-test is a statistical technique that evaluates whether notable differences exist between two groups within a dataset. This method considers variations in both directions, accounting for instances where one group is either larger or smaller than the other. In the realm of emotion recognition, the two-tailed *t*-test proves valuable in appraising the credibility of hypothesis statements linked to the factors that impact the consistency of emotion recognition. It probes for significant dissimilarities in consistency scores across different groups that have experienced varying levels of INTG, MT and IV. By doing so, this test aids in discerning if observed variations in ERC are the result of mere randomness or if they genuinely stem from the specific factors under investigation. The application of this statistical technique enriches the analysis with quantitative insights, empowering researchers to draw more robust conclusions about how distinct variables influence emotion recognition. This, in turn, advances our understanding of the intricate interplay between emotion recognition and the variables of interest.

To evaluate the hypotheses, the internal model's outcomes were examined, encompassing data such as the r-squared value, parameter coefficients, and t-statistics. With a predetermined significance level of 0.05 (5%), a t-statistic exceeding 1.96 and a positive beta coefficient were considered statistically significant indicators. The results of this hypothesis assessment are detailed in Table 2.

Table 2: Results of Hypothesis Test

Relationship	*P-Value*	*Interpretation*
Information Integration -> Information Consistency	0.588	Statistically insignificant, There is no impact of Information Integration and Information Consistency.
Information Variability -> Information Consistency	0.000	Highly statistically significant, positive impact of Information Variability on Information Consistency.
Model Type -> Information Consistency	0.002	Highly statistically significant, positive impact of Model Type on Information Consistency.

In summation, these findings illuminate the complex interplay between these factors. While Information Integration might not wield a substantial influence on Information Consistency, Information Variability and MT emerge as pivotal contributors to enhancing information consistency, each in their distinct manner. These insights shed light on the intricate dynamics within this domain, paving the way for a more nuanced understanding of the relationships under examination.

Conclusion

In conclusion, the research delves into the critical domain of emotion recognition through sentiment analysis, leveraging the potential of multimodal deep learning to bridge the gap between traditional sentiment analysis methods and the intricate subtleties of human emotions. The integration of diverse modalities and fine-grained emotion recognition aims to enhance the precision of emotion comprehension, offering a more comprehensive understanding of human affective states. The literature review demonstrates the evolution of emotion recognition, emphasising the significance of multimodal deep learning in capturing intricate emotional nuances. The research methodology employs a survey approach, and the data analysis reveals nuanced relationships between variables such as information integration, IV, MT and ERC. While information integration may not significantly impact consistency, IV and MT emerge as critical factors contributing to enhanced ERC.

Recommendations

Based on the outcomes and conclusion, following are the recommendations for businesses to follow:

- **Refinement of Multimodal Models**

Given the pivotal role of IV and MT, companies should focus on refining multimodal models to incorporate

a broader range of emotional cues. This may involve exploring new modalities or enhancing existing ones to capture the richness of emotional expressions.

• Real-World Application Development

Companies can leverage the insights from this research to enhance real-world applications of emotion recognition. The findings suggest that diverse sources of information contribute to consistency, guiding the development of sentiment-aware virtual assistants, customer feedback analysis tools and personalised recommendation systems.

• Further Exploration of Information Integration

While Information Integration did not show a significant impact, further research could explore nuanced approaches to integrating different types of information. This may involve investigating the role of specific combinations of modalities or refining the methods of integration to uncover potential impacts on ERC.

• Continued Advancements in Model Types

As model type emerged as a highly significant factor, companies should focus on advancing and diversifying model architectures. Exploring innovative deep learning models tailored for emotion recognition could contribute to improved consistency in identifying fine-grained emotions across diverse modalities.

In essence, this research contributes valuable insights into the complex dynamics of multimodal deep learning for fine-grained emotion recognition. By embracing the recommendations and furthering exploration in the identified areas, the field can advance towards more accurate, versatile and ethically sound emotion recognition systems that cater to the diverse and evolving nature of human emotions.

References

Gao, Y., Li, W., Zhang, X., & Ji, Q. (2019). Emotion recognition from photo-realistic facial images and audio cues based on multimodal deep learning. *IEEE Transactions on Affective Computing, 10*(4), 469-482.

Li, X., Zhao, G., & Pietikäinen, M. (2018). Deep learning of facial expression and physiological responses to affective multimedia. *IEEE Transactions on Affective Computing, 9*(4), 578-592.

Nguyen, M., Proulx, J., & Vo, B. N. (2017). Multimodal emotion recognition using deep learning architectures. *IEEE Access, 5,* 25158-25166.

Xu, Y., Song, R., Li, Y., & Tao, D. (2020). Deep multimodal learning for emotion recognition from speech and facial expression data. *IEEE Transactions on Cybernetics, 50*(4), 1436-1449.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2015). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(1), 147-170.

Zhang, Z., Zhan, X., Barbu, A., & Yang, Y. (2018). Multimodal deep learning for facial expression recognition. *IEEE Transactions on Image Processing, 27*(5), 2235-2245.

Akhtar, M. S., & Madani, S. A. (2019). A survey of deep learning-based methods for multimodal emotion recognition. *Pattern Recognition Letters, 119,* 3-11.

Chen, L., & Ji, Q. (2017). Emotion recognition from facial expressions using multilevel HOG-LBP features and multiple kernel learning. *Pattern Recognition, 61,* 15-26.

Huang et al. (2019). Learning deep representation from big and noisy multimodal data with shared representation regularization. *Pattern Recognition, 87*, 164-174.

Lin et al. (2018). A multimodal deep learning approach for emotion recognition in the wild. *IEEE Transactions on Affective Computing, 11*(2), 159-172.

Zhong, Z., & Zhang, X. (2020). Multimodal emotion recognition using 3D convolutional neural networks. *Information Sciences, 520,* 48-59.