

# Emotional Recognition and Classification Using Large Language Models

Clement Leung

*School of Science and Engineering and  
Guangdong Provincial Key Laboratory of  
Future Networks of Intelligence  
Chinese University of Hong Kong  
Shenzhen, China  
clementleung@cuhk.edu.cn*

Zhifei Xu

*School of Science and Engineering  
Chinese University of Hong Kong  
Shenzhen, China  
zhifeixu@link.cuhk.edu.cn*

**Abstract**—Many tasks, particularly safety-critical ones, require the associated human performers to be in the right emotional states. Correct emotion state recognition frequently becomes an important concern and mainstream methods often use Pre-trained Language Models (PLMs) as the backbone to incorporate emotional information. The latest Large Language Models (LLMs), such as ChatGPT have demonstrated strong capabilities in various natural language processing tasks. However, existing research on ChatGPT zero-shot has received insufficient evaluation of the performance of image emotion recognition and analysis. In this paper, we study emotion classification and prediction based on positive and negative emotional states and evaluate the emotion recognition capabilities of ChatGPT4 focusing primarily on images. We empirically analyze the impact of labeled emotion recognition and interpretability of different datasets. Experimental results show that, while ChatGPT4 can make some useful predictions of emotions based on images, there is still a substantial gap in prediction results and accuracy. Qualitative analysis shows its potential compared to state-of-the-art methods, but it also suffers from limitations in robustness and accurate inferences.

**Index Terms**—*image emotion recognition, large language model, zero-shot, ChatGPT4.*

## I. INTRODUCTION

Emotion recognition and prediction have been recognized to be a significant factor affecting human safety and have been widely studied [1][2][3][4][5][6]. There exist in general multiple ways for people to express their emotions or feelings naturally, such as voice, text, video, facial expressions, and physical behaviors. Moreover, since the ChatGPT [7] and Instruct-GPT [8] are currently believed to be a powerful and usable tool in different applications, we wish to investigate how they can be leveraged to assist in performing effective emotion recognition. Since emotional support is currently a key capability for many people in a wide variety of conversational scenarios, such as inter-social actions, mental health support, and customer chat services, we investigate the usefulness and competence of ChatGPT4 [9] to classify emotions based on facial expressions.

In fact, in today's society, people are under more and more pressure, such as being criticized by leaders, unfair

experiences, relationship break-ups, and so on. Once in a stressful situation, people may lose control of their emotions [10][11]. In this case, they may act irrationally to hurt themselves or others. Examples of incidents linked to emotional problems are rife: suicidal thoughts under the stress of school or work; frequent school shootings in the United States; and fatal crashes of vehicles involving angry drivers. In some special jobs, the emotion of the employee plays a particularly important role, such as a surgeon, pilot, truck driver, and so on (e.g., a recent incident of a pilot who attempted to cut off plane engines in mid-air was found to suffer from depression [12]). These highlight why emotion recognition is so important in our everyday life. So, what precisely is emotion recognition? Emotion recognition is a subfield of artificial intelligence that focuses on identifying and analyzing human emotions based on various inputs. The main goal of emotion recognition is to discern the emotional state of an individual or a group of individuals.

In recent years, since emerging large language model technologies and their rapid iterative development have produced many human-computer interaction robots, which have brought a new technological revolution to the field of dialogue, represented by ChatGPT4. At the same time, they demonstrate strong general language processing capabilities and also bring unprecedented semantic understanding and response generation capabilities to humans. Since their emergence has greatly improved the interactive experience with human users, the question of whether it shows emotion in the conversation has not yet been explored, and we are interested in the development of emotional dialogue technology in ChatGPT4. At the same time, we hope to explore the multi-modal tasks of ChatGPT4 in the field of emotion recognition and analyze its advantages and disadvantages [13][14][15].

In the next section, we discuss what others have done and why existing solutions are not enough. In the third section, we describe how to recognize and predict emotions. In the fourth section, experimental analysis is carried out for different categories of emotion. Finally, the conclusion of our study is drawn and summarized.

## II. RELATED WORK

Two main models have emerged to represent and explain emotion recognition: categorical and continuous. Categorical models, also known as discrete emotion models, assume the existence of a certain number of primary or basic emotions that are universally recognized and experienced by humans, regardless of cultural or individual differences. The main approach of categories in emotion recognition which is one of the most well-known categorical models was proposed by Ekman, who identified six basic emotions: happiness, sadness, fear, anger, surprise, and disgust. These emotions are considered fundamental and universally recognizable. The other is Plutchik's extension [16] of Ekman's model [17] by including eight primary emotions arranged in a wheel. These include joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. On the other hand, continuous models represent emotions in a multidimensional space, usually visualized as a spectrum or continuum [18] [19]. These dimensions are used to represent a person's emotional state in a more granular manner than discrete categories. *Valence* refers to the positivity or negativity of an emotion. *Arousal* refers to the degree of excitement or calm. *Dominance* refers to the degree of control or influence a person feels in a situation.

Emotion recognition has been studied primarily in terms of single modes. However, people express emotions through voice, text, video, facial expressions, and physical behavior; therefore, it is difficult to accurately judge emotions through a single mode alone. Since multimodal emotion classification involves the integration of multiple information sources, such as facial expressions, intonation, and physiological signals, we shall make use of images of facial expressions and texts using multimodal data sets, and Convolutional Neural Networks (CNN) [20] or transform classification models to identify and classify emotions. In many application scenarios, in addition to the current classification, it is necessary to also predict the evolution of emotion states.

## III. EMOTION TRANSITION AND CLASSIFICATION

In the real world, people's emotions are usually continuous, transitioning from one emotional state to another [21]. In practical situations, it is often necessary to predict the emotional state of the relevant personnel, e.g., allocating work rosters and scheduling hospital operations. As indicated earlier, people may act out emotionally when they are unfair or tired. In safety-critical jobs, we want the person doing the work to be in a sound emotional state [12]. In other words, people who work in high-risk industries can endanger the safety of others if they have emotional problems.

We shall focus on the categorical models and make use of both Plutchik's model, as well as Ekman's model. For Plutchik's categories using the eight primary emotions of joy, trust, fear, surprise, sadness, disgust, anger, and anticipation, we group these into positive and negative emotions, so that an individual is regarded as emotionally competent if s/he is a positive emotion state, and incompetent otherwise. We group

the states of joy, trust, surprise, and anticipation as positive (+1) emotions and group fear, surprise, sadness, disgust, and anger as negative (-1) emotions. We consider an individual to be emotionally competent if s/he is in a positive emotional state.

For Ekman's categories using the six primary emotions of happiness, sadness, fear, anger, surprise, and disgust, we also group these into positive and negative emotions. We group the states happiness and surprise as positive (+1) emotions and group sadness, fear, anger, and disgust as negative (-1) emotions. As a variation, for safety-critical jobs, we may wish to be extra safe and more strict concerning positive emotion, and we may take surprise out from the positive emotion category, and place it in the negative emotion category. Here, though, we place surprise in the positive emotion category.

We represent the emotional state at time  $t$  by  $S(t)$ ;  $t$  is time, and  $S(t)$  is the person's emotional change with time. As indicated,  $S(t)$  can be take on the values  $S(t) = 1$  or  $S(t) = -1$ , which corresponds, respectively, to positive (+1) and negative (-1) emotions. Since humans are continually bombarded by various external happenings, mood changes are often caused by events outside their control, which may be due to a variety of factors. Such factors may be related to changing conditions of financial situation, relationships, health, work, stock market, and family, and the combination of these may cause a transition from a positive emotion state to a negative emotion state and vice versa.

First, let  $S(0) = 1$  then, we represent the transition time points (from +1 to -1, or from -1 to +1) by a Poisson Process. Now,  $S(t) = 1$  if the number of transitions in the time interval  $(0, t)$  is even, and  $S(t) = -1$  if this number is odd. Therefore,

$$P[S(t) = 1 | S(0) = 1] = p_0 + p_2 + p_4 + \dots + \dots, \quad (1)$$

where  $p_k$  is the number of Poisson points in  $(0, t)$  with parameter  $\lambda$ . That is,

$$\begin{aligned} P[S(t) = 1 | S(0) = 1] &= e^{-\lambda t} \left[ 1 + \frac{(\lambda t)^2}{2!} + \frac{(\lambda t)^4}{4!} \dots + \dots \right] \\ &= e^{-\lambda t} \cosh \lambda t \end{aligned} \quad (2)$$

Now,  $S(t) = -1$  if the number of points in the time interval  $(0, t)$  is odd; that is,

$$\begin{aligned} P[S(t) = -1 | S(0) = 1] &= e^{-\lambda t} \left[ 1 + \frac{(\lambda t)^3}{3!} + \frac{(\lambda t)^5}{5!} \dots + \dots \right] \\ &= e^{-\lambda t} \sinh \lambda t \end{aligned} \quad (3)$$

Equation (2) represents the probability that the emotion is still positive at time  $t$  given that it was positive at time 0. Equation (3) gives the probability that the emotion is positive at time  $t$  given that it was negative at time 0. The parameter  $\lambda$  in both expressions represents a rate at which emotions change or decay over time. A larger value of  $\lambda$  would mean emotions

change more rapidly, while a smaller value would mean they change more slowly. Thus

$$E[S(t)|S(0) = 1] = e^{-\lambda t} [\cosh \lambda t - \sinh \lambda t] = e^{-2\lambda t} \quad (4)$$

#### IV. EXPERIMENTATION

##### A. Single Emotion Recognition

Emotion Recognition in Conversations [22][23] (ERC) is widely used in various conversation environments, including emotional analysis of comment areas on social media and supervision of various high-pressure industry personnel. At the same time, conversational emotion recognition can be implemented in chatbots to assess the user's emotional state and promote emotion-driven responses. As mentioned earlier, ChatGPT4 is a form of conversational bot, and we are interested in analyzing whether it can recognize emotions and sentiments.

1) *Dataset and Evaluation Graph:* We using three different datasets from Kaggle, Facial Expressions Training Data, Emotion Detection, and Natural Human Face Images for Emotion Recognition.

**Emotion Detection** This dataset consists of 35,685 examples of 48x48 pixel grayscale images, which contain two folders, one is trained, and the other one is tested. The folders contain different categories of emotional images. In addition, the images have been labeled by the authors for different types of emotions, including anger, disgust, fear, happiness, neutral, sad, and surprise.

**Facial Expressions Training Data** AffectNet [24] is a large database of faces marked with "impact" (the psychological term for facial expressions). In order to accommodate common memory limitations in this dataset, the authors reduce the resolution to 96x96 for the neural network processing, which indicates that all images are 96x96 pixels. Meanwhile, using Singular Value Decomposition, each image's Principal Component Analysis is calculated. The threshold for the Percentage of the First Component (index 0) in the principal components (in short the PFC%) was set to lower than 90%. This means that most if not all of the monochromatic images were filtered out. Finally, the dataset is based on Affectnet-HQ, using a state-of-the-art Facial Expression Recognition (FER) model that refines the AffectNet original label to re-label its dataset, which contains eight emotional categories - anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise.

**Natural Human Face Images for Emotion Recognition** Since facial expression recognition is usually performed using standard datasets, such as the Facial Expression Recognition dataset (FER), Extended Cohn-Kanade dataset (CK+) and Karolinska Directed Emotional Faces dataset (KDEF) for machine learning, however, this dataset was collected from the internet and manually annotated to provide additional data on real faces, with over 5,500 + images with 8 emotions categories: anger, contempt, disgust, fear, happiness, neutrality, sadness and surprise. All images contain grayscale human faces (or sketches). Each image is 224 x 224


pixel grayscale in Portable Network Graphics (PNG) format. Images are sourced from the internet where they are freely available for download e.g., Google, Unsplash, Flickr, etc.

2) *Task Definition of Single Emotion:* We are given the three data sets and select six types of emotions in the data set: anger, disgust, happiness, neutral, sadness, and surprise. Table I shows some examples of comparison between annotation and ChatGPT4's prediction, where red highlights the discrepancy. In each data set, 50 images of 6 types of emotions are randomly selected and put into ChatGPT4 for judgment. At the same time, since ChatGPT4 was released in 2023, the above experiments are all conducted using ChatGPT4. We use supervised learning and evaluate ChatGPT4's performance in zero-shot prompting settings for the above task. After the judgment of ChatGPT4, if the result is the same as our cognitive result, it will be recorded as 1, if the result is different, it will be recorded as 0, and the emotion will be recorded as positive, negative or neutral according to the description of ChatGPT4. Additionally, a Receiver Operating Characteristic (ROC) [25] curve is generated based on our recorded results. In the ROC curve, if it is a positive emotion, such as happiness, neutral, or surprise; we mark the fact result as 1. On the contrary, if it is a negative emotion, such as anger, disgust, or sadness; we mark the fact result as 0. The prediction result of ChatGPT4, in the positive emotion, is recorded as 1 if it is consistent with the actual result, otherwise, it is recorded as 0. In the same way, if it is a negative emotion if it is consistent with the fact, it will be recorded as 0, and if it is opposite, it will be recorded as 1. The evaluation index is divided into 1-3 points, 1 point means low confidence, 2 points means moderate confidence, and 3 points means high confidence.

3) *Result of Single emotion:* For tabulated data, TPR is True Positive Rate also known as Sensitivity, which measures the proportion of actual positives that are correctly identified by the model. FPR is False Positive Rate also known as 1-Specificity, it is the ratio of negative instances that are incorrectly classified as positive. Observed Operating Points are points on the ROC curve that correspond to specific thresholds used in the classifier. Each point represents the balance between TPR and FPR for a specific threshold. For example, a high threshold may result in low FPR but also low TPR, while a low threshold may increase both TPR and FPR. These points help evaluate the performance of the model and select the best threshold for the classification task. They demonstrate the trade-off between capturing as many positive results as possible (higher TPR) and avoiding false positives (lower FPR).

Table II shows the results of ChatGPT4's prediction based on a single emotion. For the surprise positive emotion, we see that the accuracy of ChatGPT4's prediction results is around 70%; for the happiness positive emotion, we see that the corresponding accuracy is around 78%, indicating highly discriminative discerning of positive emotions. Grouping the two positive emotions, a good degree of accuracy is obtained.

TABLE I. EXAMPLE OF CHATGPT4'S PREDICTION ON ERC TASK WITH IMAGES.

Image Content	Question	Annotation	Prediction
	What is the emotion of this person?	anger	surprise/shock/fear
	What is the emotion of this person?	happiness	happiness
	What is the emotion of this person?	happiness	happiness/joy
	What is the emotion of this person?	anger	frustration/concern/disapproval
	What is the emotion of this person?	sadness	sadness/crying
	What is the emotion of this person?	surprise	surprise

For negative emotions, the accuracy of ChatGPT4’s predictions from low to high is disgust, anger, and sadness. In the actual test, we find that zero-shot ChatGPT4 can predict negative emotions, but it cannot accurately determine whether it is disgust or anger. At the same time, because the individual expressions of disgust emotions are inconsistent, the prediction results are the lowest. We can find that GPT4 has six types of emotion recognition accuracy from high to low: happiness, surprise, neutral, fear, anger, and disgust. For surprise images, although GPT4 can identify most of the images as surprise or astonishment, it cannot accurately judge whether surprise is a positive emotion or a negative emotion, so it thinks that the emotion of surprise is mainly neutral. This is why the result is very similar to the neutral result.

As mentioned above, in order to avoid the harm caused by negative emotions to people in high-risk industries or high-risk groups, we mainly look at the three categories of emotions: anger, disgust, and sadness. We observe that the FPR of sadness is 0.3267, the FPR of anger is 0.4800, and the FPR of disgust is 0.6467. According to the above explanation of the FPR index, it means that the emotion of disgust is the least accurate to identify, and the emotion of the disgust category is the most difficult to judge among the six categories of emotion. In addition, the accuracy of negative emotion recognition is too low, and more prompt words may be needed to help GPT4 make judgments because according to the current zero-shot, GPT4 can determine that people have negative emotions, but cannot accurately identify disgust, contempt, or anger.

TABLE II. RESULT OF CHATGPT4’S PREDICTION ON SINGLE EMOTION RECOGNITION TASK WITH IMAGES

Emotion	Accuracy
anger	30%
disgust	19.30%
happiness	78%
neutral	69.34%
sadness	44.30%
surprise	70%

4) *Analysis and Discussion:* During the training process, it is inevitable that the images in some data sets are inconsistent with our cognition in real life. Since people have different feelings about images, there may be biases in partial image emotion recognition. For this part of the image, we use our cognition as the final judgment and compare it with the results of GPT4.

Additionally, we discover another issue: an inconsistency between ChatGPT4 and the dataset guide. Examining these actual prediction samples shows that the main challenge of ChatGPT4 is the bias between its norm and the norm of the dataset. Although dataset annotations may follow specific guidelines for determining corresponding sentiments, for specific cases, ChatGPT4 has its own interpretations and standards. For example, the dataset annotation classifies emotions when the person in the image is described as angry, while ChatGPT4 considers it as sad or lost. The difference

cannot be attributed to one being right and the other wrong, but rather emphasizes the use of different criteria, both of which are negative emotions. Upon further discussion, this misaligned criterion may not be due to the functionality of ChatGPT4 but may be attributed to under-posting tips. As prompt word guides become more complex, it becomes unreasonable to cover them with only a small amount of content. This insight can speculate on possible future directions: if the goal is not to strictly adhere to a specific guideline, then enhancements based on a few prompt settings (e.g., describing people in images) are feasible. However, evaluation using dataset labels may not be appropriate and may require extensive manual evaluation. Conversely, if the goal is to strictly adhere to specific guidelines, then several prompt settings may not be the best option, and supervised fine-tuning of the model is still a better option.

*B. Different Categories Emotion Recognition in Different Dataset*

1) *Task Definition of Emotion Dataset:* First, we use three data sets: emotion detection, facial expressions training data, and natural human faces. Since each dataset has different label classifications, each dataset randomly selects 50 images from 6 images of the same category (anger, disgust, happiness, neutral, sadness, surprise), for a total of 300 images. Next, we put them into GPT4 for inspection and record the results, which are shown in Table III.

2) *Result of Emotion Dataset:* Since the concept of the partial definition has already been explained previously, here we only discuss the Fitted ROC Area and Empiric ROC Area. Fitted ROC Area refers to the area under the ROC curve that uses some form of parametric or semi-parametric model to fit the data. Here we use the maximum likelihood fit of a binormal model to calculate and draw the ROC curve. Empirical ROC Area, often referred to simply as Area Under the Curve (AUC), is a measure based on an empirical ROC curve constructed directly from data. The curve is created by plotting the True Positive Rate (TPR), versus the False Positive Rate (FPR), at different threshold settings. The AUC of an empirical ROC curve provides a measure of a model’s ability to differentiate between two classes (positive and negative) at all possible thresholds. The larger the AUC, the better the model performance. An AUC of 0.5 indicates no discrimination (equivalent to a random guess), whereas an AUC of 1.0 indicates perfect discrimination.

In the Emotion Detection data set, because the two Observed Operating Points of TPR are both 0.8133, the Fitted ROC Area is Degenerate. In addition, we can find that the prediction results of Emotion Detection are the best regardless of the accuracy or empirical ROC Area, which shows that using the Emotion Detection data set for ChatGPT has the highest zero-shot prediction. Next is Natural Human, Facial Expression. Facial Expression is an RGB image data set, and the other two are black-and-white image data sets. Therefore, we find that the accuracy of RGB images is not necessarily higher

TABLE III. COMPARISON DIFFERENT DATASET OF CHATGPT4'S PREDICTION FOR EMOTION RECOGNITION TASK WITH IMAGES

Dataset	Accuracy	Sensitivity	Specificity	Fitted ROC Area	Empiric ROC Area
<b>Emotion Detection</b>	75.30%	81.30%	69.30%	Degenerate	0.74
<b>Facial Expression</b>	66.00%	61.30%	70.70%	0.665	0.634
<b>Natural Human</b>	70.30%	74.70%	66.00%	0.752	0.681

than that of black and white images, which means that color has little impact on the prediction process in emotional image recognition.

The ordinate of the ROC curve represents sensitivity. The higher the index, the higher the diagnostic accuracy. The abscissa represents 1-specificity. The lower the index, the lower the false positive rate. So in general, the closer the point is to the upper left corner of the ROC space, the better the diagnostic effect is. This means that the closer the sensitivity is to 1, the higher the prediction accuracy of the model. We can find that the sensitivities of the three data sets are 81.3% (emotion detection), 61.3% (facial expression), and 74.7% (natural human), respectively. From the specificity, it would appear that GPT4's emotion detection may be considered to be acceptable, especially for the first and last datasets. From the Accuracy and Specificity columns of Table III, the figures are somewhat comparable to the sensitivity, although marginally less acceptable.

V. CONCLUSION AND FUTURE WORK

In this paper, we study the zero-shot ability of ChatGPT4 in emotional reasoning and judgment based on images. The experimental results show that ChatGPT's predictive ability is limited, but it has the potential to improve via mental health analysis and some humanistic inputs. We target the analysis for limitations, such as unstable predictions and inaccurate inferences. Overall, our study shows that subjective tasks, such as mental health analysis and image conversational emotion reasoning remain challenging for ChatGPT. With more refined prompt engineering and contextual example selection, we believe greater future efforts are needed to improve the performance of ChatGPT and address its limitations in order to enable it to be practically applied to real-world mental health and related situations.

REFERENCES

[1] C. H. C. Leung, J. J. Deng, and Y. Li, "Enhanced Human-Machine Interactive Learning for Multimodal Emotion Recognition in Dialogue System," Proceedings of the 5th International Conference on Algorithms, Computing and Artificial Intelligence, pp. 1-7, 2022.

[2] J. J. Deng, and C. H. C. Leung, "Towards Learning a Joint Representation from Transformer in Multimodal Emotion Recognition," Brain Informatics: 14th International Conference, BI 2021, Virtual Event, September 17-19, 2021, Proceedings 14, pp. 179-188, 2021, Springer.

[3] J. J. Deng, C. H. C. Leung, and Y. Li, "Multimodal emotion recognition using transfer learning on audio and text data," Computational Science and Its Applications-ICCSA 2021: 21st

International Conference, Cagliari, Italy, September 13-16, 2021, Proceedings, Part III 21, pp. 552-563, 2021, Springer.

[4] J. J. Deng, and C. H. C. Leung, "Deep Convolutional and Recurrent Neural Networks for Emotion Recognition from Human Behaviors," Computational Science and Its Applications-ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1-4, 2020, Proceedings, Part II 20, pp. 550-561, 2020, Springer.

[5] J. J. Deng, and C. H. C. Leung, "Dynamic time warping for music retrieval using time series modeling of musical emotions," IEEE transactions on affective computing, vol. 6, no. 2, pp. 137-151, 2015.

[6] J. J. Deng, C. H. C. Leung, M. Alfredo, and L. Chen, "Emotional states associated with music: Classification, prediction of changes, and consideration in recommendation," ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 5, no. 1, pp. 1-36, 2015, ACM New York, NY, USA.

[7] T. Brown et al., "Language models are few-shot learners," Advances in neural information processing systems, vol. 33, pp. 1877-1901, 2020.

[8] L. Ouyang et al., "Training language models to follow instructions with human feedback," Advances in Neural Information Processing Systems, vol. 35, pp. 27730-27744, 2022.

[9] Open AI, ChatGPT-4, "https://openai.com/gpt-4".

[10] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: a narrative review," NPJ digital medicine, vol. 5, no. 1, pp. 46, 2022, Nature Publishing Group UK London.

[11] D. Ciraolo, A. Celesti, M. Fazio, M. Bonanno, M. Villari, and R. S. Calabrò, "Emotional Artificial Intelligence Enabled Facial Expression Recognition for Tele-Rehabilitation: A Preliminary Study," 2023 IEEE Symposium on Computers and Communications (ISCC), pp. 1-6, 2023.

[12] R. Lewis, and J. Rose, 'I'm not okay,' off-duty Alaska pilot allegedly said before trying to cut the engines, 'https://www.npr.org/2023/10/24/1208244311/alaska-airlines-off-duty-pilot-switch-off-engines', Oct. 2023.

[13] K. Yang, S. Ji, T. Zhang, Q. Xie, and S. Ananiadou, "On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis," arXiv preprint arXiv:2304.03347, 2023.

[14] W. Zhao, Y. Zhao, X. Lu, S. Wang, Y. Tong, and B. Qin, "Is ChatGPT Equipped with Emotional Dialogue Capabilities?" arXiv preprint arXiv:2304.09582, 2023.

[15] H. D. Le, G. S. Lee, S. H. Kim, S. Kim, and H. J. Yang, "Multi-Label Multimodal Emotion Recognition With Transformer-Based Fusion and Emotion-Level Representation Learning," IEEE Access, vol. 11, pp. 14742-14751, 2023.

[16] P. Robert, "Emotion: Theory, research, and experience. vol. 1:

- Theories of emotion,” 1980, Academic Press: Cambridge, MA, USA.
- [17] P. Ekman, ”Facial expressions of emotion: New findings, new questions,” 1992, SAGE Publications Sage CA: Los Angeles, CA.
- [18] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, ”Emotion recognition in context,” Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1667–1675, 2017.
- [19] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, ”Context based emotion recognition using emotic dataset,” IEEE transactions on pattern analysis and machine intelligence, vol. 42, no. 11, pp. 2755–2766, 2019.
- [20] W. Zhang, X. He, and W. Lu, ”Exploring discriminative representations for image emotion recognition with CNNs,” IEEE Transactions on Multimedia, vol. 22, no. 2, pp. 515–523, 2019.
- [21] A. Metallinou, and S. Narayanan, ”Annotation and processing of continuous emotional attributes: Challenges and opportunities,” 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), pp. 1–8, 2013.
- [22] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, ”Emotion recognition in conversation: Research challenges, datasets, and recent advances,” IEEE Access, vol. 7, pp. 100943–100953, 2019.
- [23] S. Poria et al., ”Recognizing emotion cause in conversations,” Cognitive Computation, vol. 13, pp. 1317–1332, 2021, Springer.
- [24] A. Mollahosseini, B. Hasani, and M. H. Mahoor, ”Affectnet: A database for facial expression, valence, and arousal computing in the wild,” IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18–31, 2017.
- [25] T. Fawcett, ”An introduction to ROC analysis,” Pattern recognition letters, vol. 27, no. 8, pp. 861–874, 2006, Elsevier.