Exploring GPT-4 for Fine-Grained Emotion Classification

Ratchakrit Arreerard, Scott Piao

School of Computing and Communications, Lancaster University, Lancaster, UK

1 Introduction

Mental health has been receiving an increasing attention. To assist people with mental health issue, the digital health and natural language processing communities have been exploring methods and techniques for automatically detecting mental health issues from textual data. A key technique useful for mental health analysis is the identification of emotions people express in their social media messages. Mental state is closely related to people's emotions. In fact, emotions can be the second mostly used feature for detecting mental issues on social media, particularly depression and suicide risk [1].

There exist various theories and psychological frameworks for classifying emotion categories. Some of them are coarse-grained, which divide emotions into broad main categories. For example, Ekman's scheme consists of six emotion categories [2]. Other emotion classification schemes attempt to define fine-grained emotion categories. For example, the GoEmotions scheme [3] consists of 27+1 emotion categories.

To automatically recognise emotions from textual data, several emotion classification models have been proposed and tested, such as [4, 5]. Since ChatGPT was introduced in late 2022, which is capable of classifying text [6], generative AI models have been tested for emotion classification. An important issue in this regard is, how the granularity level of the different emotion schemes can affect the emotion classification performance of generative AI. In this study, we investigate this issue by comparing the performance of ChatGPT4 for emotion classification with two different emotion schemes including Ekman's and GoEmotions schemes.

2 Methods and Data

In our experiment, we chose ChatGPT based on GPT-4 [7] as an emotion classification tool and selected GoEmotions dataset [3] as our test data. GoEmotions dataset is a collection of English Reddit messages, where each message is manually tagged with one or more emotion categories. The annotation scheme of this dataset consists of a range of finely grained emotion categories, including a total 27 emotion types and neutral category. These 27 emotion types can be grouped under the Ekman's broader six basic emotion categories.

We selected GPT-4 as the classifier because it was the latest generative AI model when we started this study. We wrote a Python script to access the OpenAI API of GPT-4, and used prompts to request GPT-4 to complete the emotion classification task. Figure 1 shows a prompt

template used in our experiment. The purpose of the prompt is to ask GPT-4 to select only one emotion category from a provided list of emotions conveyed by the given *Text*.

Figure 1: A GPT-4 prompt template.

We tested GPT-4 based on Ekman and GoEmotions schemes separately. For each scheme, we evaluated the performance of GPT-4 using an accuracy metric. First, we calculated an accuracy for each emotion category. Then we averaged the accuracy of all emotion categories, obtaining an overall accuracy for each scheme. Finally, we compared the overall accuracy between the two schemes.

3 Results

In our experiment, GPT-4 achieved an overall accuracy of 46.5% with the Ekman's scheme. It showed the best performance in identifying *fear* followed by *disgust* and *joy*, with accuracy of 70.1%, 67.1%, and 52.4% respectively. With GoEmotion scheme, GPT-4 obtained an overall accuracy of 35.6%, which is 10.9% lower than that with Ekman's scheme. GPT-4 obtained the best performance in identifying *amusement* with 80.1% accuracy, followed by *nervousness* (75%) and *disapproval* (67.2%).

We found that GPT-4 sometimes classifies the messages into classes beyond the range of provided candidate categories. With GoEmotion scheme, the number of emotions classified by GPT-4 reached 46 categories, while it produced 10 categories with Ekman's scheme. While some of them are incorrect, some others indeed capture correct emotions linked to the manual gold-standard annotation. In a couple of cases, GPT-4 suggestions appear to be even more appropriate than manual annotation. Occasionally, GPT-4 lacks understanding of context and refused to provide an answer when it determines a message as offensive.

4 Conclusion

Our experiment shows that, overall, a higher granularity level of emotion scheme negatively affects the performance of GPT-4. However, as shown in the previous section, some emotion categories of GoEmotion received higher accuracy than those of Ekman scheme. This implies that generative AI can perform better on more specifically defined narrower emotion categories than on broader basic categories. Another interesting finding is, the GPT-4 suggested 18 and 3 new categories to the GoEmotion and Ekman respectively, some of which even make more sense than human classification. This result implies a possibility that generative AI can potentially assist in designing a robust emotion classification scheme. Finally, it also raises an issue of how to accurately evaluate classification of generative AI where human produced gold-standard may not be completely reliable. For future work, we will extend our study for more emotion classification schemes and multiple emotion classification cases.

5 Study Context

This study used a subset (4,590 Reddit posts) of GoEmotions dataset [3] which is accessible at https://huggingface.co/datasets/go_emotions) and is publicly freely available. For detecting emotions, we used GPT-4 API that was accessed using a Python program. There is no ethical issue in our work.

References

[1] Malhotra A, Jindal R. Deep learning techniques for suicide and depression detection from online social media: A scoping review. Applied Soft Computing. 2022;130:109713. Available from: https://www.sciencedirect.com/science/article/pii/S1568494622007621.

[2] Ekman P. Basic emotions. Handbook of cognition and emotion. 1999;98(45-60):16.

[3] Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: A Dataset of Fine-Grained Emotions. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 4040-54. Available from: https://aclanthology.org/2020.acl-main.372.

[4] Hasan M, Rundensteiner E, Agu E. Automatic emotion detection in text streams by analyzing twitter data. International Journal of Data Science and Analytics. 2019;7:35-51.

[5] Akhtar MS, Ekbal A, Cambria E. How Intense Are You? Predicting Intensities of Emotions and Sentiments using Stacked Ensemble [Application Notes]. IEEE Computational Intelligence Magazine. 2020;15(1):64-75.

[6] Kocoń J, Cichecki I, Kaszyca O, Kochanek M, Szydło D, Baran J, et al. ChatGPT: Jack of all trades, master of none. Information Fusion. 2023;99:101861. Available from: https://www.sciencedirect.com/science/article/pii/S156625352300177X.

[7] OpenAI. GPT-4 Technical Report; 2023. https://arxiv.org/abs/2303.08774.