Emotion Classification Using Large Language Models: A Comparison of Fine-Tuning and Prompting*

Ainhoa Guerrero-San Martín[1,2], Wenceslao González-Viñas[3] and César de Pablo-Sánchez[2]

[1]Institute of Data Science and Artificial Intelligence (DATAI), University of Navarra, Pamplona, España
[2]Banco Bilbao Vizcaya Argentaria, S.A., Bilbao, España
[3]Dept. of Physics and Applied Mathematics, University of Navarra, Pamplona, España

Abstract

As large language models (LLMs) improve in their interactions with humans, it is essential to evaluate whether they truly understand human behavior and can detect emotions. To explore this, we reviewed key emotion profiles and emotionally annotated datasets used for emotion detection. We then compared a model fine-tuned on such data in English and Spanish with a larger zero-shot model using prompting. Our goal was to assess whether a model with extensive parameters becomes so effective that domain-specific retraining is no longer necessary. Results show that although fine-tuned models remain more accurate, the performance gap is narrowing. This raises the question of whether the additional computational cost and time required for domain adaptation are justified, given the increasingly marginal gains.

Keywords

generative artificial intelligence, large language model, natural language processing, ChatGPT, fine-tuning, zero-shot, transfer learning, emotion analysis

1. Introduction

In the era of large language models (LLMs) and their growing use in human interaction, it seems pertinent to assess whether these LLMs can detect human emotion. As they become integrated into customer service, education, and virtual assistance, the ability to interpret emotional and social nuances is increasingly vital [1]. The correct detection of emotions not only enhances the precision and naturalness of interactions but also contributes to the development of systems that appropriately respect and respond to the complexities of human emotions and behaviors.

Emotions are essential adaptive mechanisms that enable humans to respond to various environmental challenges.

Regarding LLMs, our objective is to determine whether the additional computational cost of fine-tuning a model for a specific application or domain is justified, considering that these models have already been pre-trained on extensive datasets with a large number of parameters.

* SEPLN 2025: 41st International Conference of the Spanish Society for Natural Language Processing, Zaragoza, Spain, 23-26 September 2025

aguerrerosa@alumni.unav.es; ainhoa.guerrero@bbva.com (A. Guerrero); wens@unav.es (W. González-Viñas); cesar.depablo@bbva.com (C. de Pablo-Sánchez)

0009-0006-8344-6850 (A. Guerrero); 0000-0002-3738-6985 (W. González-Viñas); 0009-0000-1399-3281 (C. de Pablo-Sánchez)

In the following sections, we review the types of emotional profiles that exist, as well as the datasets with annotations used for emotion detection. This framework enables us to compare somewhat smaller language models (SLMs) fine-tuned on domain-specific datasets with large language models (LLMs) operating in zero-shot learning settings, all evaluated using the same efficiency metrics and dataset.

2. Emotional Profiles

The word "emotion" comes from the Latin "emovere" which means an impulse that induces action. Therefore, emotions can be defined as primary psychophysiological responses to internal or external stimuli. Emotions are immediate, intense, and generally short-lived.

To better understand emotions, it is important to distinguish and define what feelings are, as the two terms are often used interchangeably despite having distinct meanings. Feelings are subjective, conscious interpretations of emotions; they tend to be longer-lasting and less intense and are usually influenced by past experiences, beliefs, and thoughts. In essence, feelings are the sum of emotions and thoughts—that is, they are the result of emotions. An emotion transforms into a feeling as one becomes aware of it.

One of the characteristics of emotions is that they are evolutionary, and adaptive, and influence thinking, consequently, they affect decision-making and social interaction. Additionally, they serve as adaptation mechanisms that help us face different situations in our environment [2].
Throughout modern history, various interpretations of emotion have emerged, along with models or emotional profiles. These profiles can be divided into two types:

- Discrete or Categorical: separate classes of emotions with no inherent relationship among them.
- Dimensional: emotions are considered interconnected, with each emotion represented as a point influenced by different characteristics in a multidimensional space.

Within the first type is Ekman's emotional profile [3], based on facial expressions. He differentiates six types of basic emotions: anger, disgust, fear, joy, sadness and surprise.

Ortony, Clore, and Collins developed another categorical emotional model (OCC) [4] based on a cognitive approach to observing emotions. They define 22 categories of emotions (see **Figure 1**) that depend on how we interpret events, agents, or objects according to our goals, beliefs, and values.

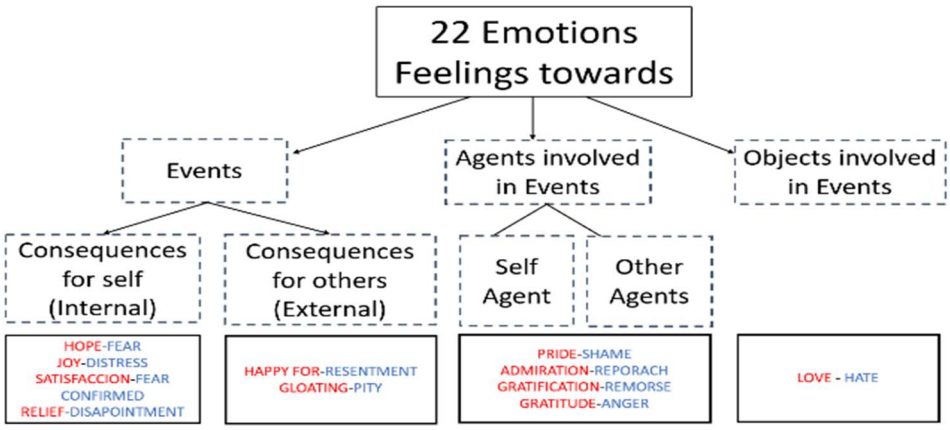

Figure 1: Representation of the OCC discrete emotional model.

Among multidimensional models, one type of emotional profiling is that of Plutchik [5], who notes that emotional states exhibit certain similarities, allowing them to be combined in varying degrees of intensity to give rise to secondary or tertiary emotions. He defines emotions through a wheel of eight primary emotions, which are as follows: anger, aversion (disgust in Ekman's model), fear, joy, sadness, surprise, anticipation, and trust.

Another multidimensional model is Russell's model [6], which measures human emotional states using two variables: valence and activation. Valence indicates whether an emotional experience is positive or negative on a continuous scale. Activation describes the intensity with which an emotion is experienced, ranging from inactive to highly active. Russell places these variables within a coordinate system, with valence as the horizontal axis and activation as the vertical. This results in a circular distribution of emotions, where "prototypical" states are located at the center of the graph. This approach is known as the circumplex model of emotions (**Figure 2**).

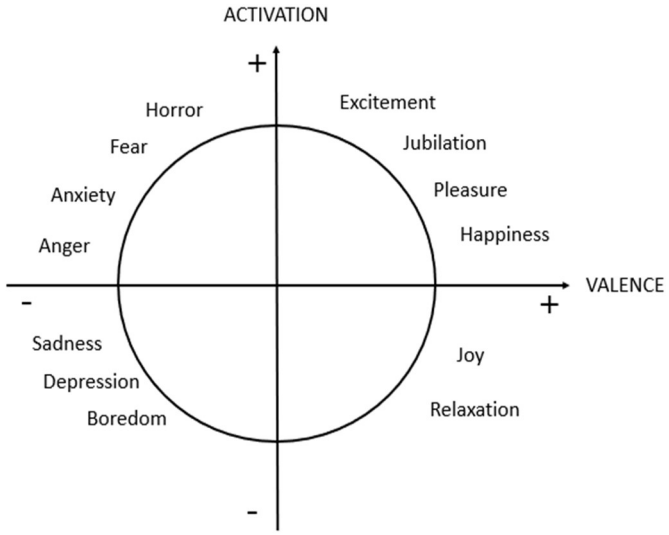

Figure 2: Representation of Russell's Multidimensional Emotional Model.

3. Emotional Datasets

We reviewed public datasets with emotional annotations to identify those that serve our goal of detecting emotions.

The first dataset examined is SemEval-2018 by Mohammad et al. (2018) [7], created with 22,000 tweets obtained from the X app, formerly known as Twitter, in three languages: English (49% of the tweets), Arabic (20% of the tweets), and Spanish (31% of the tweets). The tweets were classified using multi-class labels with 11 emotions plus a neutral label. The classification task was manually performed on 5% of each emotion to obtain a golden dataset or validation base. The rest of the classification was carried out by collaborative crowdsourcing teams who used different methods of automatic classification and natural language processing methodologies, such as embeddings trained from Spanish tweets (for example, those provided by Rothe et al., 2016 [8]) which were used as a basis to train deep learning models, including convolutional neural networks and recurrent neural networks with gated units such as LSTM and GRU. Traditional machine learning models were also used, such as Support Vector Machine (SVM), along with Spanish affective lexicons to construct the

feature space, like the Spanish emotion lexicon by Sidorov et al. (2012) [9] and ML-SentiCon by Cruz et al. (2014) [10].

Other datasets have been constructed as derivative work from the SemEval-2018 dataset: TweetEval, by Barbieri et al. (2020) [11] which exclusively include English tweets from SemEval-2018 labeled with a single class for four basic emotions: anger, sadness, optimism, and joy. Since the number of tweets with single labels was limited, only those emotions with a minimum frequency of 300 examples in the training set were included.

MELD was developed from multi-participant dialogues from the TV series Friends, containing 13K utterances from 1.4K dialogues in English. It was classified manually using 6 of Ekman's emotions plus a neutral category considering all available modalities like text, audio, and video as described by Poria et al. (2018) [12].

GoEmotions by Demszky et al. (2020) [13] contains 58K messages downloaded from Reddit from 2005 to 2019 in English, with a taxonomy based on 27 emotions plus a neutral label, obtained through an initial classification based on emotions identified by Cowen and Keltner (2017) [14]. The data set was classified manually with three annotators.

ExTES is an example of a synthetically generated dataset, described by Zheng et al. (2023) [15]. For its development, comprehensive emotional support scenarios that integrate response strategies are used to simulate chat-based dialogues. Then, an expanded set of dialogues is generated using a GPT-3.5-turbo model, followed by a manual correction process. This iterative process generates a large dataset while reducing the amount of human labor required.

4. Experimental setting

For our experiment, we selected the SemEval-2018 dataset, which is widely adopted by the research community for emotion detection tasks. This dataset includes tweets in both Spanish and English and employs a classification based on a simplified discrete profile, thereby avoiding interdependencies among emotions. Additionally, only four basic single-class emotions were considered, providing an adequate representation of the human emotional profile. For English messages, the TweetEval subset—derived from SemEval-2018—was used, while the Spanish dataset was constructed by filtering under the same premises.

To address this task, we have adopted the approach described in Zhang et al. (2024) [16], which explores whether generative large language models (LLMs) using zero-shot prompting deliver classification results comparable to that of smaller language models (SLMs) fine-tuned on the same datasets across various sentiment analysis tasks. Fine-tuning is a form of transfer learning in which a pre-trained model is adjusted to better adapt to new data [17].

Their findings indicate that zero-shot prompted LLMs yield results comparable to fine-tuned SLMs in simpler tasks, such as sentiment classification. However, for more complex tasks—such as opinion classification, which involves extracting both polarity and topic—there remains significant room for improvement compared to fine-tuned models on domain-specific data. Additionally, they conducted a comparative study on emotion detection using the TweetEval dataset, as described in Section 2.2 (Barbieri et al., 2020) [10]. This dataset consists of 5,052 English tweets, divided into training, testing, and validation sets (64%, 28%, and 8%, respectively).

In this study, we first aim to replicate the emotion detection findings using the same TweetEval dataset and evaluate their transferability to another language, specifically Spanish. We employed two language models: a smaller model T5-Base (hereafter T5), fine-tuned with the dataset using the Adam optimizer with a learning rate of 1e-4, and a fixed batch size of 4 for all tasks with 3 epochs for the full training setting and a larger model GPT 3.5-Turbo, version 0125 (hereafter 3.5), in a zero-shot

configuration [18]. The evaluation metric was the average macro F1 score over three runs with different random seeds.

Similarly, we explored more advanced versions of the GPT model: the 4o-mini version, released on July 18, 2024 (hereafter 4o) [19], and a preliminary version of GPT-4.5 (hereafter 4.5) [20]. Additionally, we conducted experiments using two Spanish datasets: TweetEval, which was translated into Spanish using the 4o model, and a subset of SemEval-2018 containing Spanish tweets. Aware of the potential biases introduced by literal translation, we included the latter to provide a more robust evaluation on native content. This subset consists of 7,000 tweets, divided into training, testing, and validation sets (50%, 40%, and 10%, respectively). For both datasets, only the unique classes corresponding to the four emotions represented in TweetEval were selected, and a comparative analysis across the four models was conducted.

5. Evaluation

The average results obtained from three runs with a random seed in our experiment were consistent with those reported in the reference article. Although some differences are observed, they are likely due to the use of different random seeds. For T5, the difference is -0.58%, and for 3.5, it is -0.55% (see **Table 1**).

It can also be observed that in the Spanish dataset, 4o and 4.5 models outperform the adjusted T5, suggesting that the effort dedicated to dataset-specific training may not be justified. In contrast, for the English dataset, the fine-tuned model is even more efficient (+2.25%). In this case, it is necessary to assess whether the efficiency gains sufficiently offset the additional training time required.

Table 1
Results of the average of three runs with a random seed of a Macro F1 obtained in our experiment. For models T5-Base (T5), GPT 3.5-turbo (3.5), GPT 4o-mini (4o) and GPT 4.5-Preview (4.5) and those obtained in the reference article

F1-Macro Score	T5	3.5	4o	4.5
Zhang et al., 2024[16]	**80.35**	72.83	N/A	N/A
TweetEval in English	**79.77**	72.28	75.89	78.10
TweetEval in Spanish	63.96	72.40	75.86	**78.37**
Native Spanish SemEval-2018	51.61	63.11	**71.92**	70.51

Table 2
Comparative analysis of emotion classification accuracy between T5-Base (T5), GPT 3.5-turbo (3.5), GPT 4o-mini (4o) and GPT 4.5-Preview (4.5) models

TweetEval in English	T5	3.5	4o	4.5	TOTAL
Anger	86.3	76.9	**88.5**	85.1	558
Joy	**84.8**	82.7	**78.4**	81.2	358
Optimism	**68.6**	44.2	**54.7**	66.7	123
Sadness	78.5	**83.3**	**77.5**	81.1	382
TweetEval in Spanish	T5	3.5	4o	4.5	TOTAL
Anger	84.6	70.7	**85.0**	**86.1**	558
Joy	64.8	78.3	**83.9**	**80.8**	358
Optimism	32.2	61.8	**51.8**	**70.7**	123
Sadness	67.9	**84.7**	**76.9**	79.0	382
Native Spanish SemEval-2018	T5	3.5	4o	4.5	TOTAL
Anger	77.3	59.5	**77.8**	63.3	914
Joy	78.2	78.0	**85.9**	82.7	799
Optimism	1.0	41.9	35.4	**57.7**	97
Sadness	52.7	**91.5**	87.8	88.8	396

Table 2 presents a comparative analysis between some versions of the GPT models and the reference T5 model for detecting the four emotions separately across two datasets—SemEval and TweetEval (the latter in both English and Spanish). For the English dataset, the results indicate that the smallest fine-tuned model more effectively detects the emotions of joy and optimism, while the GPT models demonstrate greater efficiency in identifying sadness and anger—although the latest version shows a decrease in efficiency for anger.

Regarding the detection of emotion types in the Spanish datasets, the GPT models outperform the T5 model in classifying all emotions. Furthermore, compared to previous versions, the latest GPT model exhibits improved performance in detecting optimism, while its ability to detect joy has diminished.

We analyzed tweets with label mismatches (Table 3), where discrepancies often stem from emotional ambiguity and the lack of contextual information.

Table 3
Errors Analysis with Native Spanish SemEval-2018

Tweets	True Label	Prediction
Que risa con las personas que dicen que unos hacen pedos cuando todos los pedos los hacen ellos. #Ironico	Anger	Joy
El partido de Isco fue una mierda y el de Bale, Modric, Kroos, Casemiro y CR7 fue la hostia en verso JAJAJAJA #ironia	Anger	Joy
Nunca hay que creer que los artistas se suicidan porque son cool. Son cool muy a pesar de ese destino fatal.	Optimism	Sadness
No estoy llorando, se me metió un Hybrid Theory en el ojo.	Sadness	Joy

6. Conclusion and Future Work

In this study, we reviewed emotional profiles and corpus for emotion detection and compared two tweet datasets—TweetEval (in English and Spanish) and SemEval-2018—using a fine-tuned T5-Base model versus advanced GPT versions with zero-shot prompting.

The results indicate that using prompting with large generative models is increasingly accurate for the problem of emotion classification.

The difference between them and a domain-specific fine-tuned model is diminishing. For the Spanish datasets, the large models demonstrate better accuracy than that offered by a model fine-tuned on the Spanish dataset.

Additionally, regarding the specific detection of emotions, the latest version of GPT shows improved detection of all emotions with the Spanish dataset compared to the smaller fine-tuned model. However, with the English dataset, the detection of joy still has room for further refinement.

Based on these results, it is observed that for Spanish datasets, the time spent fine-tuning the model appears unnecessary; however, for English datasets, further improvements in detecting joy are needed to enhance overall efficiency.

As future work, we plan to extend this analysis to the domain of turn-based conversation, with the goal of evaluating model performance in more dynamic and context-rich interaction settings.

Acknowledgements

The authors would like to thank BBVA and Universidad de Navarra for their valuable support throughout this research and financial support.

Declaration on Generative AI

(by using the activity taxonomy in ceur-ws.org/genai-tax.html):
During the preparation of this work, the authors used GPT-4 in order to: Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G., Emotion recognition in human–computer interaction, IEEE Signal Processing Magazine, (2001), pp. 32–80. DOI: 10.1109/79.911197

[2] Keltner, D., & Haidt, J., Social functions of emotions at four levels of analysis, Cognition & Emotion, (1999), pp. 505–521.

[3] Ekman, P., Facial expressions of emotion: New findings, new questions, Psychological Science, (1992), pp. 34–38.

[4] Ortony, A., Clore, G. L., & Collins, A., The Cognitive Structure of Emotions [La estructura cognitiva de las emociones], Siglo XXI de España Editores, Madrid, 1996.

[5] Plutchik, R., Emotion: A Psychoevolutionary Synthesis, Harper & Row, New York, 1980.

[6] Russell, J. A., & Barrett, L. F., Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant, Journal of Personality and Social Psychology, (1999), pp. 805–819.

[7] Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S., Semeval-2018 task 1: Affect in tweets, in: Proceedings of the 12th International Workshop on Semantic Evaluation, (2018), pp. 1–17.

[8] Rothe, S., Ebert, S., & Schütze, H., Ultradense word embeddings by orthogonal transformation, arXiv preprint arXiv:1602.07572, (2016). URL: https://arxiv.org/abs/1602.07572

[9] Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., et al., Empirical study of machine learning based approach for opinion mining in tweets, in: Lecture Notes in Computer Science, (2013), pp. 1–14. DOI: 10.1007/978-3-642-37807-2_1

[10] Cruz, F. L., Troyano, J. A., Pontes, B., & Ortega, F. J., ML-SentiCon: A Multilingual Lexicon of Semantic Polarities at the Lemma Level, Procesamiento del Lenguaje Natural, (2014), pp. 113–120.

[11] Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L., TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification, ArXiv, abs/2010.12421, (2020). URL: https://arxiv.org/abs/2010.12421

[12] Poria, S., Hazarika, D., Majumder, N., et al., MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations, ArXiv, abs/1810.02508, (2018). URL: https://arxiv.org/abs/1810.02508

[13] Demszky, D., Movshovitz-Attias, D., Ko, J., et al., GoEmotions: A Dataset of Fine-Grained Emotions, in: Annual Meeting of the Association for Computational Linguistics, (2020).

[14] Cowen, A. S., & Keltner, D., Self-report captures 27 distinct categories of emotion bridged by continuous gradients, Proceedings of the National Academy of Sciences, (2017), pp. E7900–E7909.

[15] Zheng, Z., Liao, L., Deng, Y., & Nie, L., Building emotional support chatbots in the era of LLMs, arXiv preprint arXiv:2308.11584, (2023). URL: https://arxiv.org/abs/2308.11584

[16] Zhang, W., Deng, Y., Liu, B., Pan, S., & Bing, L., Sentiment analysis in the era of large language models: A reality check, Findings of the Association for Computational Linguistics: NAACL 2024, (2024).
[17] Pan, S. J., & Yang, Q., A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering, (2010), pp. 1345–1359.
[18] OpenAI, ChatGPT: Optimizing language models for dialogue, 2022. URL: https://platform.openai.com/docs/models/gpt-3.5-turbo
[19] OpenAI, GPT-4o-mini: Experimental version, 18 July 2024. URL: https://platform.openai.com/docs/models/gpt-4o-mini
[20] OpenAI, GPT-4.5 preliminary release: Preliminary technical report, 2024. URL: https://platform.openai.com/docs/models/gpt-4.5-preview

A. Appendix. Prompts for Emotion Classification

English Prompt Zero-Shot: "Please perform Emotion Classification task. Given the sentence, assign an emotion label from ['0' to anger, '1' to joy, '2' to optimism, '3' to sadness]. Return label only without any other text."

Spanish Prompt Zero-Shot: "Por favor, realiza la tarea de Clasificación de Emociones. Dada la oración, asigna una etiqueta de emoción de entre ['0' para enojo, '1' para alegría, '2' para optimismo, '3' para tristeza]. Devuelve únicamente la etiqueta sin ningún otro texto."

Prompt to translate TweetEval: "You are a translator. Translate the following text into Spanish. Provide only the translation without any extra text."