IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 September 2025, accepted 17 October 2025, date of publication 20 October 2025, date of current version 28 October 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3623877

RESEARCH ARTICLE

Unraveling Emotions With Pre-Trained Models

ALEJANDRO PAJÓN-SANMARTÍN[1], FRANCISCO DE ARRIBA-PÉREZ[1], SILVIA GARCÍA-MÉNDEZ[1], FÁTIMA LEAL[2], BENEDITA MALHEIRO[3], AND JUAN CARLOS BURGUILLO-RIAL[1]

[1]Information Technologies Group, atlanTTic, University of Vigo, 36310 Vigo, Spain
[2]Research on Economics, Management and Information Technologies, Universidade Portucalense, 4200-072 Porto, Portugal
[3]ISEP, Polytechnic of Porto, Institute for Systems and Computer Engineering, Technology and Science, 4200-465 Porto, Portugal

Corresponding author: Silvia GarcÍa-Méndez (sgarcia@gti.uvigo.es)

This work was supported in part by Portuguese National Funds through FCT–Fundaç ao para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) (https://doi.org/10.54499/UIDP/50014/2020) under Project UIDP/50014/2020.

ABSTRACT Transformer models have significantly advanced the field of emotion recognition. However, there are still open challenges when exploring open-ended queries for Large Language Models (LLMs). Although current models offer good results, automatic emotion analysis in open texts presents significant challenges, such as contextual ambiguity, linguistic variability, and difficulty interpreting complex emotional expressions. These limitations make the direct application of generalist models difficult. Accordingly, this work compares the effectiveness of fine-tuning and prompt engineering in emotion detection in three distinct scenarios: (*i*) performance of fine-tuned pre-trained models and general-purpose LLMs using simple prompts; (*ii*) effectiveness of different emotion prompt designs with LLMs; and (*iii*) impact of emotion grouping techniques on these models. Experimental tests attain metrics above 70 % with a fine-tuned pre-trained model for emotion recognition. Moreover, the findings highlight that LLMs require structured prompt engineering and emotion grouping to enhance their performance. These advancements improve sentiment analysis, human-computer interaction, and understanding of user behavior across various domains.

INDEX TERMS Emotion recognition, large language models, natural language processing, open-ended responses, prompt engineering, transformer models.

I. INTRODUCTION

Emotion recognition is a task for Natural Language Processing (NLP), which enables machines to understand and respond to human emotions embedded in text. Emotion recognition analyzes texts to detect and classify emotions such as sadness, joy, love, anger, fear, and surprise [1]. This ability is essential for various applications, including sentiment analysis, customer service, mental health monitoring, and human-computer interaction. The advent of transformer models has significantly advanced the field, offering good accuracy in capturing human emotions [2]. NLP models have gained significant importance in recent years, primarily due to advances in their ability to analyze and understand human language in an automated manner [3]. These improvements have been significantly enhanced, and the technology has been popularized by companies such as OpenAI and Google.

Transformer models, introduced by [4], are a type of deep learning architecture that leverages self-attention mechanisms to process data sequences. These models can capture long-range dependencies and contextual information more effectively than previous architectures, such as Recurrent Neural Networks (RNNs) and their subset, Long Short-term Memory Networks (LSTMs) [5]. Building on transformer architecture, Large Language Models (LLMs) are deep learning models trained on vast amounts of text data to understand and generate human language. LLMs, such as the Generative Pre-trained Transformer models (*e.g.*, GPT-3 and GPT-4), leverage the extensive knowledge gained during pre-training on diverse text corpora and can be fine-tuned for specific tasks, achieving high accuracy in various NLP applications. In contrast, traditional machine learning (ML) models, such as logistic regression, support vector machines,

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

and decision trees, typically rely on manually engineered features and are trained on specific tasks with limited data sets. While effective for certain applications, these models often lack the scalability and contextual understanding of transformer-based architectures and LLMs.

Transformer models, such as Bidirectional Encoder Representations from Transformers (BERT) GPT, have demonstrated superior performance in a wide range of NLP tasks due to their ability to capture long-range dependencies and contextual information effectively. These models can be employed for prompt engineering and fine-tuning. Prompt engineering formulates specific prompts to guide the model's responses without altering internal parameters. At the same time, fine-tuning adjusts the model's weights by training it on a task-specific dataset.

One of the most challenging approaches to emotion recognition is the joint exploration of open-ended queries and transformer models. Unlike structured data, open-ended questions are rich in context and detail, providing a comprehensive view of the inquirer's emotions and desires. This complexity presents significant challenges for emotion detection models, as they must accurately interpret diverse expressions and contexts of emotion. To address this challenge, this work compares the effectiveness of prompt engineering and fine-tuning in emotion detection with open-ended questions.

This paper analyses three distinct scenarios in the context of emotion recognition: (*i*) performance of fine-tuned pre-trained models and general-purpose LLMs using simple prompts; (*ii*) effectiveness of different emotion prompt designs; and (*iii*) impact of different emotion grouping techniques on the performance of LLMs. Regarding emotion recognition with open-ended responses, the goal is to identify the most effective methods to improve the accuracy and reliability of BERT, RoBERTa, Gemma, GPT-3.5, and LLaMA-3. The results with six emotion classes show that fine-tuned BERT and spaCy models are effective at emotion detection with at least 80 % accuracy, while general-purpose LLMs using prompt engineering achieve only around 50 % accuracy. The latter models reach 80 % accuracy when focusing solely on two emotion categories.

This document is structured as follows. Section II surveys multiple works based on traditional and transformer-based models. Section III details the proposed method, describing the explored scenarios. Moreover, Section IV presents and discusses the results compared to competing works from the literature. Finally, Section V makes the final remarks.

Despite recent advances, automatic emotion recognition in open-ended responses still presents significant research gaps. In particular, general-purpose LLMs show limitations in handling complex emotional expressions without specific tuning, and there is a critical dependence on prompt design. Furthermore, systematic comparisons between fine-tuning-based approaches and prompt engineering remain scarce in open-ended natural language contexts. Given these opportunities for contribution, the following objectives are proposed:

- To comparatively evaluate the performance of fine-tuned pre-trained models and general-purpose LLMs using prompt engineering in emotion recognition tasks.
- To analyze the impact of different types of prompts on model performance.
- To study how emotion grouping affects the emotion detection capabilities of LLMs.

II. RELATED WORK

Open-ended questions are a qualitative data collection method in which respondents can answer questions in their own words rather than selecting from predetermined options. Unlike closed-ended questions, which offer a fixed set of options (*e.g.*, yes/no, multiple-choice), open-ended queries enable respondents to express their thoughts, feelings, and experiences. Additionally, open-ended questions are employed across various domains, including psychology [6], sociology [7], marketing [8], and education [9]. Typically, they are employed in surveys, interviews, and focus groups to gather individual feedback and derive insights.

The analysis of open-ended responses presents several challenges. The unstructured nature of the data requires sophisticated methods of processing and interpretation. Traditionally, researchers have relied on manual classification, where themes and patterns are identified through an intensive reading process and categorization of responses [10]. However, recent advancements in NLP and ML have automated the analysis of open-ended data, enabling efficient and scalable evaluation [11]. NLP has grown significantly in recent years, primarily due to its flexibility to adapt to a wide range of applications and to generate complex responses with minimal instructions.

Additionally, in emotional evaluation, open-ended questions are a powerful source of information, as emotions are complex and multifaceted. In this context, open questions can help individuals understand how they feel and express emotions, which is fundamental to understanding consumer behavior [12], inferring mental health states [13], or improving human-computer interaction [14]. To address this challenge, this research explores the ability of NLP models (LLMs and traditional approaches) to perform text-based emotional evaluation, aiming to comprehend and analyze the emotions expressed in a text.

The literature on emotional evaluation includes methodologies for analyzing emotional content [15], ML classification methods [16], NLP techniques [17], and the implications of these findings for various domains [18]. Understanding human emotions in texts enables the analysis of human communication, leading to more informed decisions, better services, and improved outcomes across various fields [19]. The integration of emotional evaluation into the analysis of open-ended responses represents a critical advancement in harnessing the full potential of textual data.

A. ML MODELS FOR EMOTION RECOGNITION

ML

models have revolutionized the field of emotional evaluation, offering tools for extracting and interpreting the emotional content embedded in textual data [20]. These models leverage advanced algorithms to analyze vast amounts of text efficiently and accurately, identifying patterns and emotional cues that may be imperceptible to traditional manual analysis.

Supervised learning involves training models with labeled data sets where the emotional categories of text samples are predefined. This approach allows the model to learn the relationship between linguistic features and specific emotions. Standard supervised learning algorithms used in emotional evaluation include: (*i*) Naive Bayes (NB); (*ii*) Support Vector Machines (SVM); (*iii*) Neural Networks (NN); (*iv*) Logistic Regression (LR); (*v*) K-Nearest Neighbours (KNN); and/or (*vi*) Boosting and Gradient ensemble techniques, *e.g.*, Random Forest. These ML algorithms have been successfully applied to sentiment analysis and emotion recognition [21].

Elaborating on the characteristics of these models for our work, NB is a probabilistic classifier based on Bayes' theorem that estimates the probability of a given class based on a series of observed features. This model has proven effective in text classification tasks, where vector representations (such as bags of words or TF-IDF) generate high-dimensional structures. Thanks to its low computational cost, ease of implementation, and competitive results, it can be used as a baseline. Moreover, SVM is especially effective for high-dimensional data such as text, as it optimizes class separation through the superposition of hyperplanes. This enables robust classification even with small data sets. Its ability to handle multi-class problems makes it suitable for emotional analysis extended to a vast number of categories. NNs enable the capture of nonlinear and complex relationships in data. In text processing, this approach can learn hierarchical representations, detecting emotional nuances that elude simpler methods. Furthermore, LR is a linear classifier that is widely used for binary and multi-class classification. Its simplicity makes it a solid starting point, especially when combined with text representation techniques such as TF-IDF or word embeddings. On the contrary, KNN is a nonparametric method that classifies based on proximity in feature space. Although time-consuming to perform, it is helpful in exploratory phases and as a benchmark against more sophisticated models. Finally, boosting and gradient ensemble techniques combine multiple classifiers to build more accurate and robust models. They are instrumental when working with unbalanced distributions.

Sentiment analysis enables automated and efficient processing of textual data to discern and categorize sentiment patterns. Typically, it focuses on determining the polarity of a text – whether it expresses positive, negative, or neutral sentiment – and is often used to measure public opinion, customer feedback, or overall sentiment towards a particular topic, product, or event [22]. In this line, [23] applied SVM to train a sentiment classifier with reference data sets. Moreover, [24] employed Convolutional Neural Networks (CNNs) for sentiment analysis. Unlike traditional language-dependent methods that rely on word-level, this language-agnostic model processes raw text at the character level. The resulting robust sentiment classifier works across multiple languages without extensive pre-processing or language-specific resources. Later, [25] performed a sentiment analysis on Twitter data using the NB model. The probabilistic model categorized sentiment, providing valuable information on public opinion and trends on social media platforms.

Emotion recognition identifies the emotions contained within a text. Unlike sentiment analysis, which typically categorizes text into broad sentiment categories (positive, negative, and neutral), emotion recognition aims to identify and categorize emotions such as sadness, joy, love, and anger. This more sophisticated analysis involves complex modeling to detect emotional cues [22]. Accordingly, [26] used an ML-based ensemble technique to classify six primary textual emotions. Specifically, it compares eight standard ensemble techniques to conclude that the ensemble with Term Frequency-Inverse Document Frequency (TF-IDF) achieves the best results. Moreover, [27] proposed a Multi-label KNN classifier to enable iterative adjustments in multi-label emotion recognition. This method was applied to enhance the accuracy and efficiency of emotion recognition in short Twitter texts. Among more recent works are the studies [28] and [29], which focused on emotion recognition from audio data. The study [28] uses deep neural networks while [29] exploits CNNs.

B. TRANSFORMER MODELS

LLMs represent a significant evolution in deep learning applied to natural language. Based on transformer-like architectures, these models are capable of processing text sequences considering the full context of a sentence or paragraph, allowing for a richer and more accurate understanding of meaning. An LLM is characterized by having been trained with large volumes of textual data, giving it a generalist capability to tackle multiple linguistic tasks. Typical applications include text generation, machine translation, sentiment analysis, and, more recently, emotion recognition [30].

Linguistic feature extraction, used by initiatives such as the Semantic Orientation Calculator [31], assigns polarities to different words, creates a dictionary, and applies several algorithms to calculate emotional scores for each entry, resulting in a final classification. Recent advancements in NLP rely on developing transformer models and LLMs [32]. While a transformer model provides the underlying deep learning architecture, an LLM applies the same architecture to complex linguistic tasks on a considerably larger scale. This is the case of well-known LLMs like ChatGPT and Gemini, which have become the basis for many state-of-the-art NLP solutions. These models are dedicated to understanding and generating human language, including billions of parameters, and are trained with large amounts of text data in parallel.

The mathematical operation of an LLM begins with the conversion of the text into numeric tokens using a Byte Pair Encoding (BPE) segmentation scheme [33]. These tokens are transformed into fixed-dimensional vectors through an embedding layer, which assigns each token a continuous representation in a vector space. These vectors are processed sequentially by a stack of decoder blocks, each of which is responsible for progressively refining the internal representation of the sequence. The central component of each block is the Multi-Head Attention (MHA) mechanism [34], which enables the model to determine the relevance of each token concerning the others within the sequence context. Mathematically, this mechanism projects the input vectors into three matrices: Query (Q), Key (K), and Value (V). The attention weights are then calculated using the scaled dot-product attention formula, which adjusts the magnitude of the similarities between Q and transposed K by dividing them by the square root of the dimension of K, thereby stabilizing the gradients and improving training efficiency. The final step is to apply the softmax function, which takes a vector of real values as input and converts them into a probability distribution.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The result of this operation is a weighted combination of the vectors V, in which the weights assigned to each position depend on its contextual relevance in the sequence. Now that each token has a contextual vector, it is passed to a Feed-Forward Network (FFN) model[1] [34], which is applied independently to each position in the sequence. This network is composed of two linear transformations separated by a non-linear ReLU activation function ($\max(0, z)$). This component introduces nonlinearity and more complex transformation capability, allowing the model to represent highly expressive functions.

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

where x is the input vector for a token, already enriched by the attention function; W_1 and W_2 are trainable weight matrices, whose values are adjusted during training using an optimization algorithm; and b_1 and b_2 are bias vectors that allow shifting the output and increasing the flexibility of the model.

Currently, the most advanced and sophisticated solutions are based on transformers, which capture representations of words and contexts in a general and flexible way, adding significant value. Prominent models in emotion recognition include BERT [35] and GPT [36]. Consideration should also be given to proprietary solutions, such as Anthropic, the basis for the enterprise conversational assistant Claude, or Inflection, a model used to create a personal intelligence assistant, as well as open source offers, such as Vicuna, a modified version of the LLM by Meta, LLaMA.

Transformer technology has been employed in various emotion recognition tasks, demonstrating superior performance in accurately identifying and classifying emotions across diverse data sets and languages [36], [37], [38]. This is true for both multimodal environments – including multiple types of data like audio, video, and text [39], [40], [41] – and unimodal environments – which rely on a single data type [30], [42], [43], [44], [45]. Consequently, transformer models can perform audio-based, video-based, and text-based emotion recognition.

Reference [39] proposed Emotion-LLaMA, a large multimodal language model. It incorporates HuBERT for audio processing and visual encoders to gather facial details, dynamics, and context. Emotion-LLaMA significantly enhances emotional recognition and reasoning capabilities by integrating multiple descriptive elements like the audio tone, lexical subtitle, visual objective, visual expression, classification label, and modality. Conversely, [41] applied model adaptation techniques – deep prompt tuning and low-rank adaptation – to customize the Chat General Language Model (ChatGLM), an open-source pre-trained language model, for emotion recognition tasks. The adapted versions outperform state-of-the-art models, tested on six audio, video, and text datasets.

More recent solutions using transformers are [46], [47], and [48]. Firstly, [46] proposed MobileBERT for emotion recognition from textual and video data. A similar solution is proposed in [47], which utilizes KoELECTRA and HuBERT in a multimodal scenario involving both textual and audio data. Finally, in [48], a textual emotion recognition system based on LLaMA2 is presented.

Moreover, Emotion Recognition in Conversation (ERC) focuses on detecting emotions during dialogues. It aims to identify the emotional category of each utterance in a conversation, whether text-based or audio-based. In this line, [42] introduced InstructERC, a framework that combines the strengths of retrieval-augmented mechanisms and LLM solutions like GPT-3 and T5 to access external knowledge and contextual information, thereby addressing the limitations of traditional ERC models. By employing these models, InstructERC improves the accuracy of emotion classification in dialogues. Conversely, [30] explored the text-generating capabilities of LLMs to enrich intelligent conversational agents with the ability to recognize and adapt to the emotions of the partner speaker during textual dialogues. Similarly, [43] proposed a text-based ERC considering contents and contextual factors like dialogue history, speaker roles, and the interplay between different conversational turns. In the end, [44] presented DialogueLLM, an emotion and context knowledge enhanced language model designed explicitly for ERC, based on open-source base models, namely LLaMA2. Similarly, [45] and [49] presented new advancements in the field of ERC. More in detail, [45] worked with ambiguous emotions in zero-shot and few-shot settings, while [49] focused on zero-shot conditions but performed experiments

[1] Available at https://arxiv.org/pdf/2406.08413, September 2025

with real and synthetic data in both text and speech modalities.

Speech Emotion Recognition (SER) identifies and classifies the speaker's emotional state based on vocal expressions. This implies analyzing various speech signal features, such as pitch, tone, intensity, rhythm, and prosody, to detect emotions like happiness, sadness, anger, and fear. In this line, [40] explored the integration of speech analysis, text generation, and speech synthesis. The data2vec pre-trained model performs speech analysis to capture nuanced vocal features; GPT-4 generates text to provide contextual understanding and augment emotion detection, and Azure Text-to-Speech implements emotional speech synthesis to create a more expressive and accurate SER system.

Finally, recognizing emotions through transformer technology enables asking open-ended questions using an old human strategy, *i.e.*. An open-ended query expands the range of potential responses and increases the model's uncertainty. The model generates an elongated response to traverse the spectrum of possible interpretations to curtail this ambiguity. In this regard, [50] evaluated the capabilities of LLMs to understand human intentions, emotions, and reasoning processes when addressing open-ended questions. The study compares human and LLM responses using Zephyr-7B, LLaMA2, and GPT-4. The results show the effectiveness of incorporating mental states, such as human intentions and emotions, into prompt tuning to improve the quality of LLM reasoning. Moreover, our prior work by [51] combined contextual information with prompt engineering and a general-purpose LLM to enhance emotion recognition. We adopted a prompt template integrating the head, emotions, polarities, objective, structure, question, and optional conversation. In addition, [41] recommended and implemented a two-sentence prompt template. The first sentence provides the emotion recognition instructions: " `Classify the sentence of the sentence to Emotion`$_1$`, Emotion`$_2$`, ..., Emotion`$_k$". The second sentence holds the contents submitted for emotion recognition: " `<a single sentence from the test set>`". The value of k corresponds to the number of sentiment/emotion categories specific to the dataset.

In summary, most of the reviewed research follows code rather than prompt-oriented strategies, which significantly restricts reuse by other researchers and professionals. Standard code-oriented strategies can be used with specialized and general-purpose pre-trained models, whereas prompt-oriented strategies only work with general-purpose pre-trained models. The primary advantages of the current prompt-based proposal over existing methods lie in its simplicity and seamless adaptability to diverse fields and general-purpose pre-trained models.

C. RESEARCH CONTRIBUTIONS

Emotion recognition has improved human-computer interaction, enhanced customer service, and supported mental health interventions. It involves identifying and classifying emotional states from textual, auditory, or visual data, making it essential for creating empathetic and responsive intelligent systems. This work presents a novel contribution to the field of emotion recognition from open text through the combined use of specialized pre-trained models and general-purpose LLMs. Unlike previous studies that focus exclusively on a single technique (either fine-tuning or prompt engineering), our proposal systematically compares both approaches in different scenarios, also incorporating different prompt design and emotion grouping strategies. The main contributions are the following:

- A hybrid approach is proposed that integrates fine-tuned models and general models with prompt engineering techniques.
- Five types of prompts and three different ways of grouping emotions are analyzed, evaluating their impact on model performance.
- The approach's replicability and extension are facilitated by providing reusable prompt structures that are adaptable to different models.

Table 1 offers a comparative analysis of the emotion recognition works referred to above considering the task (emotion classification: EC, emotion recognition: ER, ERC, sentiment analysis: SA and SER), the data modality (audio: A, text: T, video: V), the category of the technique (traditional ML or transformer-based) and the usage of prompt-based strategies.

While earlier works, such as those by [23], [24], and [25], focused on sentiment analysis using ML-based techniques, more recent studies adopt transformer-based methods for text-based [30], [36], [38], [42], [43], [44], [45], [50], [51] and multimodal [39], [40], [41], [49] emotion detection. Only [41], [50], and [51] have explored prompt-based strategies for emotion recognition with transformer models. [50] used prompts to perform text-based emotion recognition. Specifically, they compare, for a given topic, the reasoning quality and the emotional contents of open-ended responses produced by humans and LLM models. Reference [51] used prompt engineering with GPT-3.5 for contextual emotion recognition within interactive conversations and extensive texts. The extensive texts are dynamically divided into fragments and submitted sequentially as a conversation. In the case of a conversation, the prompt holds the entire conversation as context and indicates which part to analyze in each iteration; in the case of an extensive text, the prompt holds the emotions and polarities identified so far as context and specifies the text fragment to analyze in each iteration. This decision to provide the maximum possible context enables the model to achieve more profound results, as past events are essential. Reference [41] adapted the pre-trained ChatGLM language model for emotion recognition and then tested the resulting models using a sentence-level emotion recognition prompt.

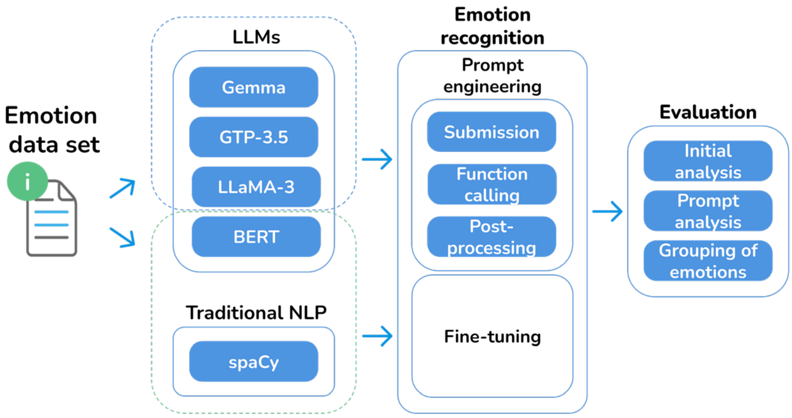


FIGURE 1. System diagram.

TABLE 1. Comparison of the surveyed works concerning emotion analysis with ML and transformer models.

Authorship	Task	Data modality	Technique	Prompt engineering
[23] [24] [25]	SA	T	ML	✗
[40]	SER	A,V	Transformer	✗
[26] [27]	EC	T	ML	✗
[28] [29]	ER	A	ML	✗
[36]	ER, SA	T	Transformer	✗
[38]	ER	T	Transformer	✗
[46]	ER	T,V	Transformer	✗
[47]	ER	A,T	Transformer	✗
[48]	ER	T	Transformer	✓
[50]	ER	T	Transformer	✓
[39]	ER	A,V	Transformer	✗
[51]	ER	T	Transformer	✓
[41]	ER	A,T,V	Transformer	✓
[30]	ER	T	Transformer	✗
[43]	ER	T	Transformer	✗
[42]	ERC	T	Transformer	✗
[44]	ERC	T	Transformer	✗
[45]	ERC	T	Transformer	✓
[49]	ERC	A,T	Transformer	✓
Proposed solution	ER	T	ML Transformer	✓

In contrast, the proposed solution integrates ML- and transformer-based techniques with multiple prompt strategies and emotion groupings. The design and refinement of open-ended questions are innovative features that enhance the model's ability to adapt and respond to contextual variations, significantly advancing emotion recognition. Furthermore, the current work leverages the power of fine-tuned specialized and general-purpose pre-trained models to enhance the performance of text-based emotion recognition.

III. METHOD

Figure 1 presents the modules of the proposed architecture. This work distinguishes between traditional and LLM approaches, dividing LLM into general- and specific-purpose models. To this end, a general-purpose LLM provides an already pre-trained model to evaluate the performance of prompt engineering.

A. LARGE LANGUAGE MODELS

This work explores the following LLM implementations: Gemma, GPT-3.5, LLaMA-3, and BERT.

1) GEMMA

Gemma is a model developed by Google and introduced in February 2024. The adopted Gemma 1.1 model (gemma1.1-7b-it[2]) has 7 billion parameters and offers excellent versatility in a wide range of areas. This version (1.1) has undergone substantial modifications by being trained using a novel method of Reinforcement Learning from Human Feedback (RLHF). This resulted in significant improvements in quality, coding capabilities, veracity, instruction following, and quality of multi-turn conversations.

The great advantage of this model lies in the balance between computing capacity and the resources required. Although it has a limited number of parameters compared to others, it exhibits very high performance in various applications.

2) GPT-3.5

GPT-3.5,[3] based on GPT-3, represents a significant evolution in natural language generation technology. This model, created by OpenAI, demonstrates an enhanced ability to comprehend and generate text with a deeper and more coherent context, thanks to its use of 175 billion parameters and extensive training on diverse datasets.

Compared to other models, the distinctive features of GPT-3.5 are its scale and complexity, which translate into high fluency and a deep understanding of complex linguistic features. This last feature allows this version to explore broader contexts than its predecessors.

Thanks to these improvements, the GPT-3.5 model can be utilized in complex domains and real-time applications due to its enhanced inference capability, which enables faster responses. This makes it ideal for complex process automation tasks, such as the current work.

3) LLAMA-3

LLaMA-3 is a model developed by Meta and introduced in April 2024. The third and most recent version includes powerful NLP capabilities similar to those of the models mentioned above.

The current system uses the 8 billion parameters version oriented to instructions[4] to provide optimized outputs for feature extraction processes. Unlike other models, LLaMA-3 needs to incorporate different prompts delimited by keywords indicated in the documentation to exemplify the outputs.

4) BERT

BERT is a language model developed by Google in 2018. Unlike traditional NLP models, which process text sequentially and in only one direction (left-to-right or right-to-left), BERT employs a transformer architecture that enables bidirectional encoding of context. This means that BERT can simultaneously consider information from both preceding and following words in a sentence, providing a richer and more accurate understanding of the contextual meaning of each word.

One of the most prominent features of BERT is its ability to pre-train large corpora of unlabelled text using two main tasks: masked language modeling (MLM) and next sentence prediction (NSP). With MLM, BERT learns to predict hidden words in a sentence based on context, while with NSP, the model learns the relationship between two consecutive sentences.

The versatility of BERT lies in its ability to adjust to specific NLP tasks with little additional labeled data. This feature allowed us to fine-tune BERT for emotion recognition. The process ensures that the system consistently produces an adequate output, regardless of the quality of a specific prompt.

5) ROBERTA

RoBERTa is an optimized variant of BERT, introduced by Facebook AI in 2019. Unlike the original BERT, RoBERTa removes the next-sentence prediction objective, applies dynamic masking during pretraining, and is trained on a significantly larger corpus. These modifications lead to richer contextual representations and better generalization capabilities, which are crucial for fine-grained emotion recognition. Consequently, RoBERTa has been widely adopted as one of the reference transformer models in multiple NLP tasks, including sentiment and emotion analysis [52]. To avoid redundancy in diagrams and figures, only BERT is explicitly depicted since RoBERTa is a direct improvement, sharing the same underlying architecture with modifications in the pretraining process (e.g., dynamic masking and removal of the next-sentence prediction objective). Therefore, whenever BERT is referenced in methodological diagrams, it should be understood that RoBERTa is also encompassed in the evaluation.

B. TRADITIONAL MODELS

spaCy is an NLP library written in Python, which stands out for speed, efficiency, and accuracy in a wide range of NLP tasks. Spacy allows the training of small models focused on a specific classification function, like detecting six emotion classes.[5]

Spacy is a powerful tool for building a text classification model for polarity or entity detection. These features, combined with the extensive existing community, make SpaCy a versatile tool for text-based emotion detection.

C. EMOTION RECOGNITION

1) PROMPT ENGINEERING

The current prompt engineering approach is applied to Gemma, GPT-3.5, and LLaMA-3. The prompts used with Gemma (see Listing 1 for an illustrative example) and GPT-3.5 (see Listing 2) are similar because the system understands the instructions perfectly without needing special

[2]Available at https://huggingface.co/docs/transformers/model_doc/gemma, September 2025

[3]Available at https://platform.openai.com/docs/models/gpt-3-5-turbo, September 2025

[4]Available at https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct, September 2025

[5]Available at https://spacy.io/api/textcategorizer, September 2025

```
*Contxt:* Imagine that you are doing a study about the emotions present in a text.
*Instru:* Only detect the following emotions in the study: sadness, joy, love, anger, fear, surprise.
*Instru:* Objective: Detect the key emotion present in the text.
*Instru:* The output will be a JSON list with the key emotion delimited by % like %{value:x}%.
*Instru:* Perform the study of this text fragment following strictly the structure indicated above,
    without introducing any of the given text and only with the emotions indicated ->
*Sentnc:* sentence
*Answer:* %{value:x}%
```

LISTING 1. Gemma prompt and answer template.

```
*Contxt:* You are doing an emotional study on text input.
*Instru:* You will organize the emotions in 3 independent groups focusing on the emotion you want to
    express: the positive emotion group will be (love), the negative emotion group will be (fear),
    and the neutral emotion group will be (surprise).
*Instru:* The output will be a JSON list with a single key with the format -> emotion: positive,
    negative, or neutral.
*Instru:* The input text will be enclosed in three quotes.
*Sentnc:* '''sentence'''
*Answer:* emotion: value
```

LISTING 2. GPT-3.5 prompt and answer template.

```
*Instru:* <|begin_of_text|><|start_header_id|>system<|end_header_id|>You are a system that always
    detects the emotions [sadness, joy, love, anger, fear, surprise]. The answer format will only
    include the detected emotion. Never mix two emotions; only make single detections.
*Contxt:* All texts are for academic study.<|eot_id|>
*Sentnc:* <|start_header_id|> user <|end_header_id|>sentence<|eot_id|>
*Answer:* <|start_header_id|>assistant<|end_header_id|>value<|eot_id|>
```

LISTING 3. LLaMA-3 prompt and answer template.

commands. The LLaMA-3 prompt requires, as referred to in Section III-A3, specific keywords to obtain optimal results (see Listing 3). These prompts provide context (`Contxt`), instructions (`Instru`), and the sentence (`Sentnc`) for the general-purpose LLM to analyze.

Prompt execution comprises prompt submission, with the support of function calling in the case of GPT-3.5, and response post-processing. The goal is to identify the specified emotions within the submitted text.

- Submission. The interface with Gemma was provided via a Python server with Flask[6] as front-end and the gemma1.1-7b-it[7] model as a back-end. The LLaMA-3 model was deployed using the Text Generation Interface[8] system provided by Hugginface to optimize model generation runtime. However, it is not yet fully compatible with Gemma. Both models have been deployed using Kubernetes, allowing a fast and resource-efficient deployment. The interface with GPT-3.5 was via the API offered by OpenAI. The calls to the API require using an API-key. Being GPT-3.5, a private model, OpenAI implements a pay-as-you-go charging model.
- Function calling is used with GPT-3.5. This functionality allows the specification of the expected outputs. This provides greater precision when making requests, avoiding results with additional text or wrong ones. However, it is still necessary to incorporate post-processing functions to obtain a clean output according to the desired format.
- Post-processing is applied to LLM outputs to filter out unwanted words and incorrect formatting. In some cases, emotions with similar connotations can be confused, *e.g.*, joy and hope. Therefore, one of the post-processing steps is to normalize the values using a dictionary regarding the six considered emotions.[9] Using model-specific regular expressions[10] helps to remove unwanted characters and words, *e.g.*, line breaks or quotation marks, or the word JSON that appears explicitly with Gemma.

2) FINE-TUNING

As fine-tuning of Gemma, GPT-3.5 and LLaMA-3 is out of the scope of this work, this step was applied just to spaCy and BERT.

Being a transformer model, BERT was fine-tuned to recognize the six emotions. First, it was trained using the

[6] Available at https://flask.palletsprojects.com/en/3.0.x, September 2025
[7] Available at https://huggingface.co/google/gemma-1.1-7b-it, September 2025
[8] Available at https://huggingface.co/docs/text-generation-inference/index, September 2025
[9] Available at https://bit.ly/3W2TFpB, September 2025
[10] Available at https://bit.ly/3XYMNfD, September 2025

TABLE 2. BERT parameters.

Parameter	Value
num_classes	6
max_length	128
batch_size	16
num_epochs	4
learning_rate	2e-5

TABLE 3. RoBERTa parameters.

Parameter	Value
num_classes	6
max_length	128
batch_size	32
num_epochs	8
learning_rate	1e-5

aforementioned fine-tuning data partition with the parameters in Table 2. Table 3 details the parameters used in the RoBERTa model. This training produces a reduced model that can be executed on machines with limited resources.

spaCy is a traditional model that incorporates a series of functionalities, allowing for adaptation to any field or application. This is the case of pipes, which support model retraining for emotion recognition. Specifically, this work uses a test categorizer of six classes corresponding to the target emotions. Subsequently, pretraining was performed with the fine-tuning data partition described in Section IV-A to detect the six target classes solely.

Finally, the fine-tuned BERT, RoBERTa and spaCy models are ready for evaluation using the fine-tuning data partition also described in Section IV-A. Note that neither BERT, RoBERTa nor spaCy supports user prompting.

3) EVALUATION

The results are evaluated using classical Artificial Intelligence metrics, such as accuracy, recall, precision, and F-score.

To analyze the behavior of the selected models with different prompts and emotion groups, there are experiments with five types of prompts (see Section III-D) and three emotion groupings with six, three, and two classes (see Table 4). The six classes correspond to the sadness, joy, love, anger, fear, and surprise emotions; the three classes refer to positive (love), negative (fear), and neutral (surprise), whereas the two classes correspond to positive (joy/love) and negative (anger/sadness) feelings. The class grouping was based on the emotional charge.

D. SCENARIOS

The experiments consider three scenarios:

- S1 compares the performance of general-purpose LLMs versus pre-trained models using basic prompts (see Section III-C1).
- S2 analyses the performance of the LLMs with different prompts:
  - Basic prompt requests a single emotion from the available lists (see Section III-C1).
  - Mask prompt applies a binary mask to detect the emotions (see Table 5).
  - Percent prompt requests emotion percentages in JSON format. The model will be instructed that the desired output is a JSON list with the percentage of each analyzed emotion in the text. It will also be established that there must always be a dominant emotion.
  - Numerical prompt associates each emotion with a number.
  - Inverse prompt asks for the inverse emotion. The model is instructed to identify the inverse emotion to the one in the text. This prompt establishes whether the model can make complex associations between emotions.
- S3 assesses the impact of the distribution, choice, and grouping of emotions (six, three, and two classes) in the LLMs.

IV. EXPERIMENTAL RESULTS

Experiments were performed on a computer with the following hardware specifications:

- **Operating System**: Ubuntu 22.04.4 LTS 64 bits.
- **Processor**: IntelCore i7-13700K 3.40 GHz.
- **RAM**: 32 GB DDR4.
- **Disk**: 1000 GB NVME.
- **GPU**: Nvidia GTX-1050Ti 4 GB.

The LLM experiments were performed in a server with the following hardware specifications:

- **Operating System**: Debian 10 Buster 64 bits.
- **Processor**: IntelXeon Gold 5317 3.00 GHz.
- **RAM**: 128 GB DDR4.
- **Disk**: 100 GB SSD.
- **GPU**: Nvidia A10 20 GB.

A. EXPERIMENTAL DATA SETS

The experimental data is publicly available.[11] Table 6 shows the distribution by emotion category of two subsets (train and test). The first is used to fine-tune the traditional and BERT models. The second, with 16 000 samples, is intended for their evaluation.

B. THEORETICAL EVALUATION

In this section, a theoretical evaluation of the scenarios S1, S2, and S3 presented in Section III-D is performed.

1) S1: LLMS *VERSUS* PRE-TRAINED MODELS

Being m a model (either an LLM or a pre-trained model or PRE) and p_b a basic prompt, an evaluation metric (*e.g.*, accuracy, F-

[11] Available at https://www.kaggle.com/datasets/parulpandey/emotion-dataset, September 2025

TABLE 4. Emotion groupings.

(#)	Classes	Emotions
6	sadness, joy, love, anger, fear, surprise	sadness, joy, love, anger, fear, surprise
3	positive, negative, neutral	love, fear, surprise
2	positive, negative	joy/love, anger/sadness

TABLE 5. Relation between emotions and masks.

Emotion	Mask
sadness	000001
joy	000010
love	000100
anger	001000
fear	010000
surprise	100000

TABLE 6. Experimental data set.

Partition	Class	Number of entries
Fine-tuning	sadness	581
	joy	695
	love	159
	anger	275
	fear	224
	surprise	66
	Total	2000
Evaluation	sadness	4666
	joy	5362
	love	1304
	anger	2159
	fear	1937
	surprise	572
	Total	16 000

score) is defined as:

$$\mathcal{M}(m, p_b)$$

The performance difference between LLMs and pre-trained models can be expressed as:

$$\Delta_{\mathcal{M}} = E_{m \in LLM}\left[\mathcal{M}(m, p_b)\right] - E_{m' \in PRE}\left[\mathcal{M}(m', p_b)\right]$$

where

$$E_{m \in \mathcal{A}}\left[\mathcal{M}(m, p_b)\right] = \frac{1}{|\mathcal{A}|}\sum_{m \in \mathcal{A}} \mathcal{M}(m, p_b)$$

2) S2: PROMPT COMPARISON

Given a model m, different prompt strategies are evaluated $p \in \mathcal{P} = \{p_b, p_m, p_{\%}, p_n, p_{\text{inv}}\}$, where:

- p_b is the basic prompt.
- p_m is the binary-masked prompt.
- $p_{\%}$ is the prompt based on percentages in JSON format.
- p_n is the numerical codified prompt.
- p_{inv} is the inverse emotion prompt.

The difference in performance between the two prompt strategies is defined as:

$$\Delta_{\mathcal{M}i,j} = \mathcal{M}(m, p_i) - \mathcal{M}(m, p_j) \quad \forall p_i, p_j \in \mathcal{P},\ i \neq j$$

3) S3: EMOTION GROUPING EVALUATION

Being $\mathcal{C}$ the set of emotional classess and Π_k a partition of $\mathcal{C}$ in k non-empty subsets (in our study $k = \{6,3,2\}$), Π_k fulfills that

$$\forall A \in \Pi_k,\ \exists B \in \Pi_{k'} / A \subseteq B, \textit{where}$$
$$k > k'$$

This relationship follows the partition refinement principle, where Π_k is a refinement of $\Pi_{k'}$, which implies a higher emotional granularity.

For a model m and a prompt p, we define the performance as $\mathcal{M}(m, p, \Pi_k)$. The difference in performance from a higher to a lower granularity, that is, grouping emotions, is defined as:

$$\Delta_{\mathcal{M}k,k'} = \mathcal{M}(m, p, \Pi_{k'}) - \mathcal{M}(m, p, \Pi_k)$$

This value represents the gain obtained by reducing the class space complexity. Under the specific emotion separability hypothesis, we may assume that if $H(\Pi_k)$ is the entropy associated with the portion Π_k.

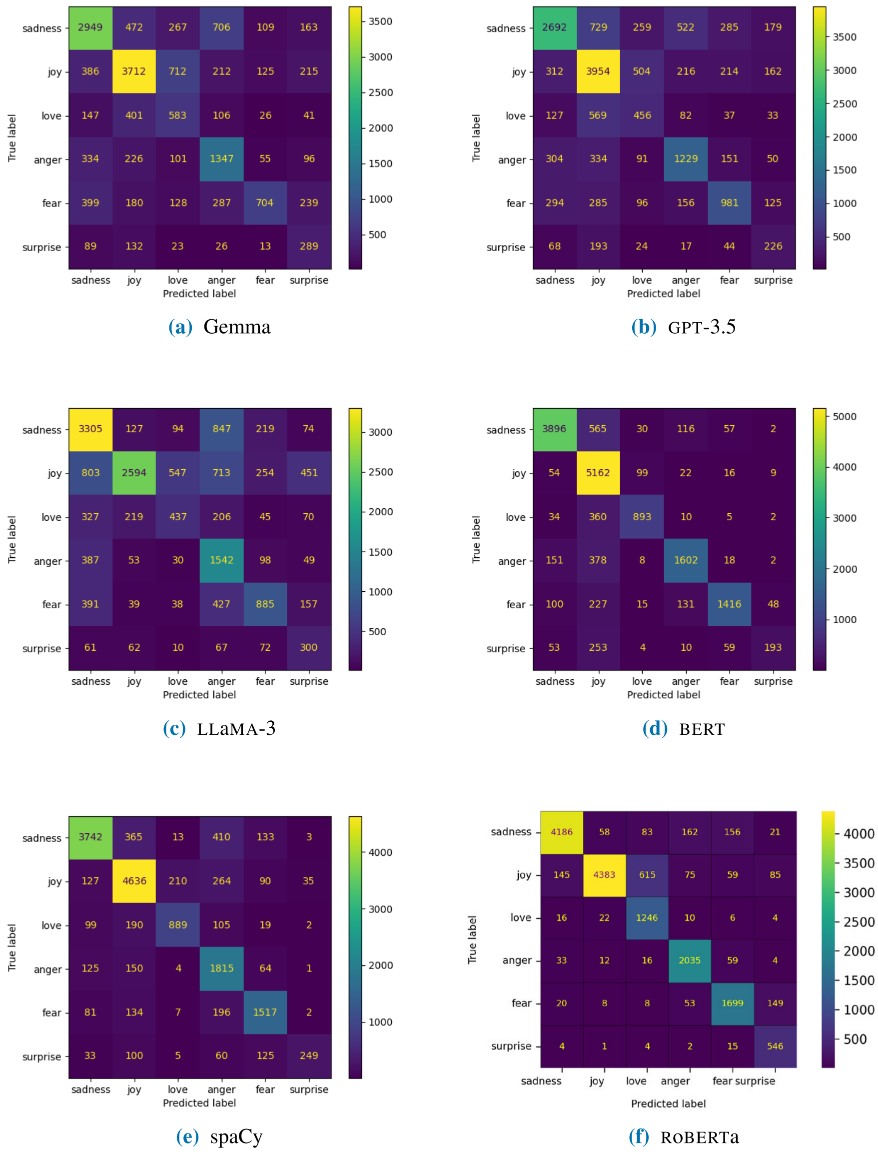
$$H(\Pi_k) > H(\Pi_{k'})$$

This formulation enables us to assess the impact on output space reduction as a supervised semantic compression phenomenon using set theory and information theory.

C. EXPERIMENTAL EVALUATION

Table 7 shows the results of the first scenario. The fine-tuning approaches (BERT, spaCy, and especially RoBERTa) yield the best results, with RoBERTa clearly outperforming all other models by reaching 90 % precision, 88 % accuracy, recall, and F-score. In contrast, general-purpose LLMs return results with accuracy close to 60 % and around 50 % for the remaining metrics. The confusion matrices (Figure 2) reveal that these models make many mistakes between close emotions (e.g., joy/love and sadness/anger) and even confuse opposite emotions (joy/sadness).

This highlights the advantage in this case of fine-tuned transformer models, which can distinguish fine-grained emotional categories thanks to their ability to capture deep contextual dependencies. Among them, RoBERTa achieves the best overall results due to several key improvements over BERT. It removes the next-sentence prediction objective, uses



(a) Gemma (b) GPT-3.5

(c) LLaMA-3 (d) BERT

(e) spaCy (f) RoBERTa

FIGURE 2. Comparison between fine-tuned and general-purpose models.

dynamic masking during pretraining, and is trained on a significantly larger corpus. These modifications allow RoBERTa to learn richer contextual representations and generalize better to unseen examples, which is crucial for separating semantically similar emotions. In contrast, general-purpose models without task-specific training struggle to separate the categories. These values may be the result of two possible factors. The issue may stem from the use of an inappropriate prompt or the existence of similar emotional categories that are difficult to classify. Both approaches are explored below.

Table 8 shows the variability in performance depending on the prompt engineering strategy used. Prompt 1 presents the best F-score for all models, while more complex formulations, such as inverse emotion (*i.e.*, strategy 5), lead to significantly lower performance, especially in Gemma and GPT-3.5.

TABLE 7. Emotion recognition results with six classes.

Model	Accuracy	Recall	Precision	F-score
Gemma	59.94	54.36	53.23	52.09
GPT-3.5	59.61	52.25	51.63	51.60
LLaMA-3	56.62	53.68	51.37	50.22
BERT	82.26	71.55	83.43	75.55
spaCy	80.30	73.46	79.62	75.07
RoBERTa	**88.00**	**88.00**	**90.00**	**88.00**

The lower results are obtained with prompts 2 (binary mask) and 5 (inverse emotion). The results with the inverse emotion prompt indicate that the models cannot establish complex relationships, such as detecting an opposite emotion. In this respect, LLaMA-3 is the most powerful model, with values of accuracy 10 % to 25 % above those of the others, as can be seen in Table 8 with the fifth prompt engineering strategy. The results obtained with the binary mask prompt are limited. In this case, the model trained solely on text was asked to establish a relationship between an emotion and a binary mask representation of that emotion.

Furthermore, the results of prompt 2 (binary mask) show the limited ability of the models to understand the translation to a binary space. However, the numerical interpretation (prompt 4), although it does not improve the results, does offer an acceptable translation compared to the basic prompt, except in the case of Gemma, which again drastically reduces its performance.

This suggests that LLMs are highly sensitive to prompt design and that complex reasoning, such as emotion inversion, is subject to improvement. The latter emphasizes the importance of clear and concise communication for achieving effective emotion detection in zero-shot scenarios. Furthermore, the limited ability of the models to address categorization tasks in highly granular contexts, where emotions with very similar semantic meanings exist, is identified.

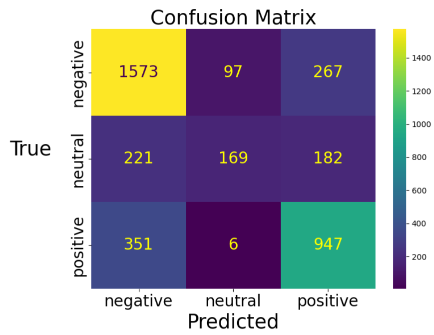
Given that LLMs detect emotions at the word level instead of the sentence level, the six emotion classes were grouped into three classes, applying the prompt 1 strategy: positive (love), negative (fear), and neutral (surprise). Table 9 and Figure 3 show the results of this approach. As can be seen in Table 9, the results show a notable improvement when reducing the number of emotion classes from six to three. F-score values increase 10 % points compared to the previous scenario, confirming that emotion grouping reduces ambiguity and confusion between semantically close classes (*e.g.*, joy and love). This result supports our claim that generalist LLMs have difficulty discriminating fine-grained emotion categories if they have not been previously fine-tuned.

The confusions decrease, reaching 60 % in all metrics with Gemma and GTP-3.5. However, it is still far from the 90 % of the fine-tuned RoBERTa model with the six emotions.

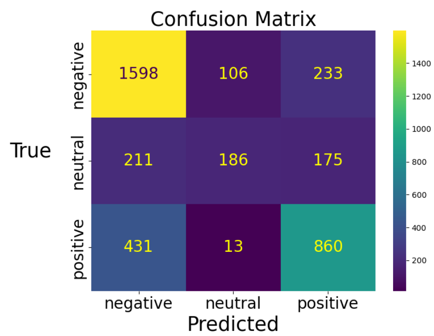
Additionally, the confusion matrices reflect a large number of errors committed by the neutral category toward positive and negative emotions. This effect translates into metrics that fail to exceed 70 %, hampered by the results obtained in this intermediate category.

Finally, Table 10 and Figure 4 elaborate on the binary scheme (positive *versus* negative emotions). As observed in Table 10, the models achieve accuracies and F-score values greater than 78 % in all cases. These results demonstrate that LLMs can be effective in simplified emotional analysis tasks, even without specialized training, validating their applicability in contexts where it is sufficient to identify the general emotional polarity of the text.

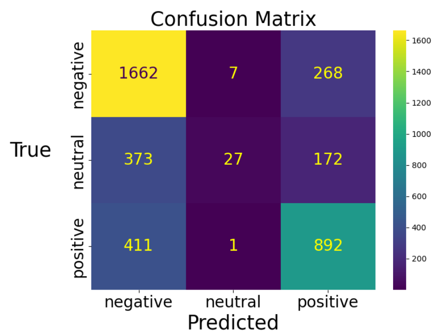
Furthermore, the experimental results support the conclusion that general-purpose LLM instances have difficulty detecting more than two classes and that optimal performance requires fine-tuning.



(a) Gemma



(b) GPT-3.5



(c) LLaMA-3

FIGURE 3. Confusion matrix with three classes.

D. DISCUSSION

Reference [50] examined the extent to which LLM understand and integrate human intentions and emotions in their open-ended answers. The approach consists of submitting human prompts from an online discussion forum and collecting and evaluating the LLM-generated responses. The evaluation relies on (*i*) humans to determine the reasoning quality of the LLM responses; (*ii*) statistical significance to establish the emotional dissimilarity between human and LLM responses; and (*iii*) metrics to quantify the semantic similarity and lexical overlap between human and LLM responses. Although with a different objective, this work on the human-like reasoning capabilities of LLMs addresses open-ended questions and identifies emotions and sentiments. Moreover, two of the

TABLE 8. Prompt engineering strategies for general-purpose LLM models.

Prompt strategy	Model	Accuracy	Recall	Precision	F-score
1	Gemma	**59.94**	**54.36**	**53.23**	**52.09**
	GPT-3.5	59.61	52.25	51.63	51.60
	LLaMA-3	56.62	53.68	51.37	50.22
2	Gemma	12.75	15.35	14.17	6.97
	GPT-3.5	12.12	13.95	15.35	10.84
	LLaMA-3	7.75	12.16	10.74	7.12
3	Gemma	50.62	42.45	49.28	41.77
	GPT-3.5	50.38	42.20	42.22	41.63
	LLaMA-3	58.0	49.57	49.62	49.08
4	Gemma	18.44	26.86	23.38	16.26
	GPT-3.5	52.12	50.78	46.93	46.89
	LLaMA-3	56.81	53.7	52.05	51.36
5	Gemma	6.69	6.11	6.82	5.81
	GPT-3.5	23.94	18.51	18.15	16.46
	LLaMA-3	32.88	35.11	42.18	27.47

TABLE 9. General-purpose LLM models with three classes.

Model	Accuracy	Recall	Precision	F-score
Gemma	66.75	**61.32**	60.47	60.81
GPT-3.5	**69.34**	60.32	66.72	**61.94**
LLaMA-3	67.69	52.98	**70.69**	50.80

TABLE 10. General-purpose LLM models with two classes.

Model	Accuracy	Recall	Precision	F-score
Gemma	**80.39**	**80.41**	**80.52**	**80.37**
GPT-3.5	79.83	79.83	79.84	79.82
LLaMA-3	78.58	78.61	79.17	78.49

general-purpose LLM used are related: LLaMA-2 and GPT-4 versus LLaMA-3 and GPT-3.5, respectively. Regrettably, the emotion recognition metrics are not comparable. Nonetheless, existing similarities allow the comparison of qualitative results regarding emotion recognition with the same-family models. Best results were obtained with GPT-4 followed by LLaMA-2 in the case of [50]. The identical behavior of the same-family models in the current proposal supports the findings.

Reference [51] explored contextual information to improve emotion recognition with general-purpose LLM. The designed prompt always contains full or partial contextual data (the complete text or, alternatively, the emotions and polarities detected so far). The evaluation involved three public datasets and GPT-3.5. The Conversations and TED talks, manually labeled by the authors, achieved an F-score above 70 % for emotions and 78 % for polarities. The Short phrases, pre-labeled by default with positive and negative polarities, reached 62 % with emotions and 87 % with polarities. Since the data sets differ from those adopted in the current work, the results are not directly comparable. In the current work, with the emotion data set, GPT-3.5 achieved values of F1 around 52 % (Table 7), 62 % (Table 9) and 80 % (Table 10) in the detection of six classes, three classes, and two polarities, respectively. While the polarity results obtained with the emotion data set are aligned with those obtained by [51] with the Short phrases, the emotion recognition values are considerably lower. This difference may result from the impact of contextual data and the manual labeling of conversations and TED talks.

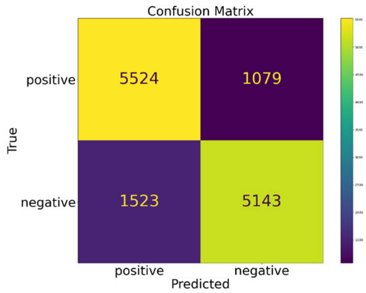
The basic prompt templates are tailored to each LLM, resulting in three prompt templates. These templates are more detailed than the one proposed by [41] and less detailed than the one adopted by [50]. The size of this more extended template is related to the specific context and the dimensions of the instructions. While all templates provide the task instructions, the list of emotions, and the sentence to be examined, the LLM prompt templates designed for this work and by [50] also provide some context. However, it is far from the total or partial context provided by [51]. While supplying context makes perfect sense, primarily when the data are organized by conversation, topic, or document, it has little impact when the data are made of unrelated short sentences.

The best overall results regarding the detection of six emotions were achieved with the fine-tuned pre-trained RoBERTa model, with all metrics above 88 %.

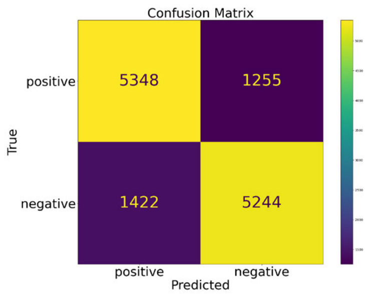
Table 11 presents a comparison with the most related competing work in the state of the art to validate our proposal further. As can be observed, our solution is the one that attains better performance in all metrics, with difference values of 35.63 % in accuracy compared to [30] and 33.55 % in F-measure compared to [43].

Regarding the most recent work, our proposal continues to be the one with better performance. Compared to the works by [44] and [45], the differences surpass the 20 % while being even higher between our proposal and that by [49], in which we attained an accuracy more than 35 % points superior.

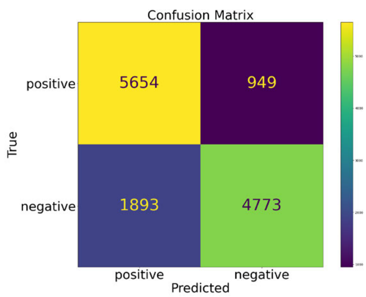
The results presented indicate that generalist LLMs show limitations in fine-grained and multi-class emotion recognition tasks. However, their performance improves significantly when both the prompt design and the emotional class



(a) Gemma



(b) GPT-3.5



(c) LLaMA-3

FIGURE 4. Confusion matrix with two classes.

TABLE 11. Benchmarking with existing research.

Authorship	Accuracy	Recall	Precision	F-score
[30]	46.63	-	53.80	47.90
[43]	-	-	-	42.00
[44]	61.49	-	-	60.52
[45]	57.87	65.39	-	58.43
[49]	45.81	41.56	-	-
Proposal	**82.26**	**71.55**	**83.43**	**75.55**

structure are simplified. While complex prompts reduce accuracy, simpler ones allow for better extraction of implicit knowledge from the model. Furthermore, by grouping emotions into more general classes, the models achieve competitive results without the need for fine-tuning. This empirical evidence supports the use of a hybrid strategy, which combines the specialization of trained models with the flexibility of generalist models guided by prompt engineering.

Despite the promising results, it is essential to elaborate on the limitations of the proposal. First, general-purpose LLMs may underperform in multi-class emotion recognition tasks without applying fine-tuning. Additionally, performance is highly dependent on prompt design, thereby increasing the complexity of implementation. Moreover, the experiments focused initially on six emotions, and further analyses will be required to reflect the emotional richness present in real-world applications. Context-aware prompt engineering strategies may also be appropriate.

V. CONCLUSION

With the advent of transformer models, the field of emotion recognition has experienced significant advancements. However, challenges remain regarding the exploration of open-ended queries and transformer models. Consequently, this study evaluates the performance of a large set of well-known LLMs with different prompt strategies and emotion groupings.

The evaluation stage comprises three scenarios to provide a comprehensive analysis of LLMs for emotion detection in open-ended queries: (*i*) performance of fine-tuned pre-trained models and general-purpose LLMs using simple prompts; (*ii*) effectiveness of different emotion prompt designs; and (*iii*) impact of emotion grouping techniques on the performance of LLMs.

The fine-tuned RoBERTa, one of the most widely used LLMs in the literature, was the best emotion detector with metrics above 88 %. Moreover, the spaCy model optimized for emotion recognition reports an average performance equivalent to that of the fine-tuned BERT. Conversely, the general-purpose LLMs using prompt engineering obtained results close to 50 % with six emotions. Their performance improved as the number of emotion groupings decreased, reaching values near 80 % for both positive and negative polarities. The prompts with the best results are the basic prompts.

Shortly, the plan is to do further research on (*i*) prompt design and refinement – define a unique user prompt template, automatically refine the submitted user prompts, and automatically translate them to the different requirements of distinct general-purpose LLM models; (*ii*) perform emotion recognition through general-purpose LLM models – experiment with additional general-purpose LLM models, namely ChatGLM and GPT-4o, with other publicly available benchmark data sets; and (*iii*) incorporating multimodal approaches that integrate text, audio, and images for emotion recognition in more complex contexts.

REFERENCES

[1] N. Alswaidan and M. E. B. Menai, “A survey of state-of-the-art approaches for emotion recognition in text,” *Knowl. Inf. Syst.*, vol. 62, no. 8, pp. 2937–2987, Aug. 2020.

[2] M. Khan, P.-N. Tran, N. T. Pham, A. El Saddik, and A. Othmani, “MemoCMT: Multimodal emotion recognition using cross-modal transformer-based feature fusion,” *Sci. Rep.*, vol. 15, no. 1, p. 5473, Feb. 2025.

[3] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia Tools Appl.*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023.
[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Annu. Conf. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
[5] Y. Zheng and X. Zhou, "Modeling multi-factor user preferences based on transformer for next point of interest recommendation," *Expert Syst. Appl.*, vol. 255, Dec. 2024, Art. no. 124894.
[6] N. Obergassel, S. Heitmann, A. Grund, S. Fries, K. Berthold, and J. Roelle, "Adaptation of quizzing in learning psychology concepts," *Learn. Instruct.*, vol. 95, Feb. 2025, Art. no. 102028.
[7] K. Kosimov, "Statistical analysis of empirical data in the process of sociological research," in *Proc. Int. Sci. Res. Conf.*, 2025, vol. 3, no. 31, pp. 77–81.
[8] S. V. Helm, V. J. Little, and C. Frethey-Bentham, "'No marketing on a dead planet': Rethinking marketing education to support a restoration economy," *J. Macromarketing*, vol. 44, no. 2, pp. 307–323, Jun. 2024.
[9] D. M. Olvet, T. B. Fulton, M. Kruidering, J. M. Brenner, J. B. Bird, and J. M. Willey, "Are open-ended question assessments an emerging trend in U.S. medical education?" *Teaching Learn. Med.*, vol. 2025, pp. 1–10, Aug. 2025.
[10] K. Hansen and A. Świderska, "Integrating open- and closed-ended questions on attitudes towards outgroups with different methods of text analysis," *Behav. Res. Methods*, vol. 56, no. 5, pp. 4802–4822, Oct. 2023.
[11] H. Gweon and M. Schonlau, "Automated classification for open-ended questions with BERT," *J. Surv. Statist. Methodology*, vol. 12, no. 2, pp. 493–504, Apr. 2024.
[12] A. M. Vyas, "Exploring the impact of 'emotion-recognition-AI' on consumer trust and satisfaction," in *Proc. IEEE Int. Students' Conf. Electr., Electron. Comput. Sci. (SCEECS)*, Feb. 2024, pp. 1–6.
[13] R. Ajayi and B. S. Adedeji, "Neural network-based face detection for emotion recognition in mental health monitoring," *Int. J. Res. Publication Rev.*, vol. 5, no. 12, pp. 4945–4963, Dec. 2024.
[14] D. Thiripurasundari, K. Bhangale, V. Aashritha, S. Mondreti, and M. Kothandaraman, "Speech emotion recognition for human–computer interaction," *Int. J. Speech Technol.*, vol. 27, no. 3, pp. 817–830, 2024.
[15] W. H. Park, D. R. Shin, and H. Mutahira, "An integrated approach to Bayesian weight regulations and multitasking learning methods for generating emotion-based content in the metaverse," *Expert Syst. Appl.*, vol. 259, Jan. 2025, Art. no. 125197.
[16] R. Ahamad and K. N. Mishra, "Exploring sentiment analysis in handwritten and E-text documents using advanced machine learning techniques: A novel approach," *J. Big Data*, vol. 12, no. 1, p. 11, Jan. 2025.
[17] A. R. Mishra, A. Rai, D. Nandan, U. Kshirsagar, and M. K. Singh, "Unveiling emotions: NLP-based mood classification and well-being tracking for enhanced mental health awareness," *Math. Model. Eng. Problems*, vol. 12, no. 2, pp. 647–656, Feb. 2025.
[18] E. H. Houssein, S. Mohsen, M. M. Emam, N. Abdel Samee, R. I. Alkanhel, and E. M. G. Younis, "Leveraging explainable artificial intelligence for emotional label prediction through health sensor monitoring," *Cluster Comput.*, vol. 28, no. 2, p. 86, Apr. 2025.
[19] J. S. Lerner, C. A. Dorison, and J. Klusowski, "How do emotions affect decision making?" in *Emotion Theory: The Routledge Comprehensive Guide*. U.K.: Routledge, 2024, pp. 447–468.
[20] A. Alslaity and R. Orji, "Machine learning techniques for emotion detection and sentiment analysis: Current state, challenges, and future directions," *Behaviour Inf. Technol.*, vol. 43, no. 1, pp. 139–164, Jan. 2024.
[21] M. Kastrati, Z. Kastrati, A. Shariq Imran, and M. Biba, "Leveraging distant supervision and deep learning for Twitter sentiment and emotion classification," *J. Intell. Inf. Syst.*, vol. 62, no. 4, pp. 1045–1070, Aug. 2024.
[22] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Netw. Anal. Mining*, vol. 11, no. 1, pp. 1–19, Dec. 2021.
[23] N. Zainuddin and A. Selamat, "Sentiment analysis using support vector machine," in *Proc. Int. Conf. Comput., Commun., Control Technol. (I4CT)*, Sep. 2014, pp. 333–337.
[24] J. Wehrmann, W. Becker, H. E. L. Cagnini, and R. C. Barros, "A character-based convolutional neural network for language-agnostic Twitter sentiment analysis," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2384–2391.
[25] M. Wongkar and A. Angdresey, "Sentiment analysis using naive Bayes algorithm of the data crawler: Twitter," in *Proc. 4th Int. Conf. Informat. Comput. (ICIC)*, Oct. 2019, pp. 1–5.
[26] T. Parvin and M. M. Hoque, "An ensemble technique to classify multi-class textual emotion," *Proc. Comput. Sci.*, vol. 193, pp. 72–81, Jan. 2021.
[27] X. Liu, T. Shi, G. Zhou, M. Liu, Z. Yin, L. Yin, and W. Zheng, "Emotion classification for short texts: An improved multi-label method," *Humanities Social Sci. Commun.*, vol. 10, no. 1, pp. 1–9, Jun. 2023.
[28] R. Sujatha, J. M. Chatterjee, B. Pathy, and Y.-C. Hu, "Automatic emotion recognition using deep neural network," *Multimedia Tools Appl.*, vol. 84, no. 28, pp. 33633–33662, Jan. 2025.
[29] X. Tang, J. Huang, Y. Lin, T. Dang, and J. Cheng, "Speech emotion recognition via CNN-transformer and multidimensional attention mechanism," *Speech Commun.*, vol. 171, Jun. 2025, Art. no. 103242.
[30] A. Pico, E. Vivancos, A. Garcia-Fornes, and V. Botti, "Exploring text-generating large language models (LLMs) for emotion recognition in affective intelligent agents," in *Proc. 16th Int. Conf. Agents Artif. Intell.*, 2024, pp. 491–498.
[31] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comput. Linguistics*, vol. 37, no. 2, pp. 267–307, Jun. 2011.
[32] M. Ren, "Advancements and applications of large language models in natural language processing: A comprehensive review," *Appl. Comput. Eng.*, vol. 97, no. 1, pp. 55–63, Nov. 2024.
[33] Z. Li and X. Lu, "Research on compressed input sequences based on compiler tokenization," *Information*, vol. 16, no. 2, p. 73, Jan. 2025.
[34] J.-H. Kim, C.-H. Kim, S.-M. Rho, and K.-S. Chung, "A low power attention and softmax accelerator for large language models inference," in *Proc. IEEE Int. Conf. Consum. Electron.-Asia (ICCE-Asia)*, Nov. 2024, pp. 1–4.
[35] H. Al-Omari, M. A. Abdullah, and S. Shaikh, "EmoDet2: Emotion detection in English textual dialogue using BERT and BiLSTM models," in *Proc. 11th Int. Conf. Inf. Commun. Syst. (ICICS)*, Apr. 2020, pp. 226–232.
[36] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 1743–1753, Jul. 2023.
[37] S. Peng, L. Cao, Y. Zhou, Z. Ouyang, A. Yang, X. Li, W. Jia, and S. Yu, "A survey on deep learning for textual emotion analysis in social networks," *Digit. Commun. Netw.*, vol. 8, no. 5, pp. 745–762, Oct. 2022.
[38] N. Wake, A. Kanehira, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, "Bias in emotion recognition with ChatGPT," 2023, *arXiv:2310.11753*.
[39] Z. Cheng, Z.-Q. Cheng, J.-Y. He, J. Sun, K. Wang, Y. Lin, Z. Lian, X. Peng, and A. Hauptmann, "Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 1–37.
[40] Z. Ma, W. Wu, Z. Zheng, Y. Guo, Q. Chen, S. Zhang, and X. Chen, "Leveraging speech PTM, text LLM, and emotional TTS for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 11146–11150.
[41] L. Peng, Z. Zhang, T. Pang, J. Han, H. Zhao, H. Chen, and B. W. Schuller, "Customising general large language models for specialised emotion recognition tasks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 11326–11330.
[42] S. Lei, G. Dong, X. Wang, K. Wang, and S. Wang, "InstructERC: Reforming emotion recognition in conversation with a retrieval multi-task LLMs framework," in *Proc. CoRR*, 2024, pp. 1–16.
[43] D. Venkatesh, P. Prasanjith, and Y. Sharma, "BITS pilani at SemEval-2024 task 10: Fine-tuning BERT and llama 2 for emotion recognition in conversation," in *Proc. 18th Int. Workshop Semantic Eval. (SemEval)*, 2024, pp. 811–815.
[44] Y. Zhang, M. Wang, Y. Wu, P. Tiwari, Q. Li, B. Wang, and J. Qin, "DialogueLLM: Context and emotion knowledge-tuned large language models for emotion recognition in conversations," *Neural Netw.*, vol. 192, Dec. 2025, Art. no. 107901.
[45] X. Hong, Y. Gong, V. Sethu, and T. Dang, "AER-LLM: Ambiguity-aware emotion recognition leveraging large language models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2025, pp. 1–5.
[46] M. Hussain, C. Chen, S. S. Albouq, K. Shinan, F. Alanazi, M. W. Iqbal, and M. U. Ashraf, "Low-resource MobileBERT for emotion recognition in imbalanced text datasets mitigating challenges with limited resources," *PLoS ONE*, vol. 20, no. 1, Jan. 2025, Art. no. e0312867.

[47] M.-H. Yi, K.-C. Kwak, and J.-H. Shin, "HyFusER: Hybrid multimodal transformer for emotion recognition using dual cross modal attention," *Appl. Sci.*, vol. 15, no. 3, p. 1053, Jan. 2025.

[48] Y. Fu, J. Wu, Z. Wang, M. Zhang, L. Shan, Y. Wu, and B. Liu, "LaERC-S: Improving LLM-based emotion recognition in conversation with speaker characteristics," in *Proc. Int. Conf. Comput. Linguistics*, 2025, pp. 6748–6761.

[49] S. Bo-Hao, S. G. Upadhyay, and L. Chi-Chun, "Toward zero-shot speech emotion recognition using LLMs in the absence of target data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2025, pp. 1–5.

[50] M. Amirizaniani, E. Martin, M. Sivachenko, A. Mashhadi, and C. Shah, "Do LLMs exhibit human-like reasoning? Evaluating theory of mind in LLMs for open-ended responses," 2024, *arXiv:2406.05659*.

[51] A. Pajón-Sanmartín, F. D. Arriba-Pérez, S. García-Méndez, J. C. Burguillo, F. Leal, and B. Malheiro, "Emotional evaluation of open-ended responses with transformer models," in *Proc. World Conf. Inf. Syst. Technol.*, 2024, pp. 23–32.

[52] F. Alqarni, A. Sagheer, A. Alabbad, and H. Hamdoun, "Emotion-aware RoBERTa enhanced with emotion-specific attention and TF-IDF gating for fine-grained emotion recognition," *Sci. Rep.*, vol. 15, no. 1, p. 17617, May 2025.

ALEJANDRO PAJÓN-SANMARTÍN received the B.S. degree in telecommunication technologies engineering from the University of Vigo, Spain, in 2023. He is currently a Researcher with the Information Technologies Group, University of Vigo. His research interests include natural language processing techniques and large language models.

FRANCISCO DE ARRIBA-PÉREZ received the B.S. degree in telecommunication technologies engineering, and the M.S. and Ph.D. degrees in telecommunication engineering from the University of Vigo, Spain, in 2013, 2014, and 2019, respectively. He is currently a Researcher with the Information Technologies Group, University of Vigo. His research encompasses the development of machine learning solutions for various domains, including finance and healthcare.

SILVIA GARCÍA-MÉNDEZ received the Ph.D. degree in information and communication technologies from the University of Vigo, in 2021. Since 2015, she has been a Researcher with the Information Technologies Group, University of Vigo. She is collaborating with foreign research centers as part of her postdoctoral stage. Her research interests include natural language processing techniques and machine learning algorithms.

FÁTIMA LEAL received the Ph.D. degree in information and communication technologies from the University of Vigo, Spain. She is an Auxiliary Professor with Universidade Portucalense, Porto, Portugal, and a Researcher with REMIT (Research on Economics, Management, and Information Technologies). Recently, she has been exploring blockchain technologies for responsible data processing. Her research is based on crowdsourced information, including trust and reputation, big data, data streams, and recommendation systems.

BENEDITA MALHEIRO received the M.Sc. and Ph.D. degrees in electrical engineering and computers from the University of Porto. She is currently a Coordinator Professor with the Instituto Superior de Engenharia do Porto, School of Engineering, Polytechnic of Porto, and a Senior Researcher with INESC TEC Porto, Portugal. Her research interests include artificial intelligence, computer science, and engineering education. She is a member of the Association for the Advancement of Artificial Intelligence (AAAI), the Portuguese Association for Artificial Intelligence (APPIA), the Association for Computing Machinery (ACM), and the Professional Association of Portuguese Engineers (OE).

JUAN CARLOS BURGUILLO-RIAL received the M.Sc. degree in telecommunication engineering and the Ph.D. degree in telematics from the University of Vigo, Spain. He is currently a Full Professor with the Department of Telematic Engineering and a Researcher with the atlanTTic Research Center in Telecom Technologies, University of Vigo. His research interests include intelligent systems, evolutionary game theory, self-organization, and complex adaptive systems. He is the Area Editor of the *Simulation Modelling Practice and Theory* (SIMPAT).