Exploring Text-Generating Large Language Models (LLMs) for Emotion Recognition in Affective Intelligent Agents

Aaron Pico[1][a], Emilio Vivancos[1][b], Ana Garcia-Fornes[1][c] and Vicente Botti[1,2][d]

[1] *Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Valencia, Spain*

[2] *Valencian Graduate School and Research Network of Artificial Intelligence (valgrAI), Spain*

Keywords: Large Language Model, Emotion Recognition, Intelligent Agents.

Abstract: An intelligent agent interacting with a individual will be able to improve its communication with its interlocutor if the agent adapts its behavior according to the individual's emotional state. In order to do this, it is necessary for the agent to be able to detect the individual's emotional state through the content of the conversation the agent has with the individual. This paper investigates the application of text-generating Large Language Models (LLMs) for emotion recognition in dialogue settings with the aim of generating emotional knowledge, in the form of beliefs, that can be used by a BDI emotional agent. We compare the performance of several LLMs in recognizing the emotions that an affective BDI agent can employ in its reasoning. Results demonstrate the promising capabilities of diverse models in a Zero-shot prediction (without training and without examples), showcasing the potential for LLMs in emotion recognition tasks. The study advocates for further refinement of LLMs to balance accuracy and efficiency, paving the way for their integration into diverse intelligent agent applications.

1 INTRODUCTION

To enable intelligent agents to interact effectively with human beings, agents must be aware of the emotional state of their counterpart, and consider this information as part of the agent decision process (Fan et al., 2017; de Melo et al., 2014; Rincon et al., 2016; Irfan et al., 2020). In recent years, there have been important advances in the field of natural language processing (NLP) to create intelligent systems capable of understanding and generating natural language in a human-like way (Iqbal and Qureshi, 2022; Nagarhalli et al., 2021). The text that is produced or recognized by these systems will express not only the ideas that are to be communicated, but will also implicitly contain details that make it possible to deduce the emotional state of the person or agent who generated the text during the agent-human conversation. Consequently, Large Language Models (LLMs) open up new possibilities for addressing the complexities in the human-like text generation/recognition including the emotion recognition (Min et al., 2023).

Emotion recognition stands as a key part of the broader field of NLP, as it fosters the creation of systems that not only comprehend surface-level content, but also discern the intricate emotional hints of human expression. Conventional methods for emotion recognition have laid essential groundwork, they often struggle to capture the depth and complexity inherent in human communication. As we navigate this evolving landscape, the advent of LLMs has become a pivotal turning point.

This paper explores the potential of leveraging text-generating LLMs for the task of recognizing emotions in textual dialogues and generate beliefs for a BDI affective agent. The affective agent will use the counterpart recognized emotional state to adapt their behavior and/or interaction with the individual. The LLMs for this exploration have been chosen to provide diversity in terms of size, capabilities and training purposes. This selection contains GPT 3.5, Llama 2 chat 7B and 13B, Orca 2 7B and 13B, Mistral Instruct 7B 0.1 and 0.2, Zephyr 7B β, and StableLM Zephyr 3B. The task these models must perform is to classify the emotion of an interaction in a dialog according to a predefined list of emotional labels we provide depending on the specific assignment.

[a] https://orcid.org/0000-0002-5612-8033
[b] https://orcid.org/0000-0002-0213-0234
[c] https://orcid.org/0000-0003-4482-8793
[d] https://orcid.org/0000-0002-6507-2756

Pico, A., Vivancos, E., Garcia-Fornes, A. and Botti, V.
Exploring Text-Generating Large Language Models (LLMs) for Emotion Recognition in Affective Intelligent Agents.
DOI: 10.5220/0012596800003636
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 1*, pages 491-498
ISBN: 978-989-758-680-4; ISSN: 2184-433X

The rest of this paper is organized as follows. The following section describes the fundamental of intelligent BDI agents and Large Languages Models. Section 3 presents our comparative study of the ability of several LLMs to detect emotions in a text. In Section 4, the main results of the study conducted are discussed. The article ends with the main conclusions and some possible future extensions.

2 INTELLIGENT AGENTS AND LLMs

Intelligent agents are designed to perceive their environment, make decisions based on acquired knowledge, and execute actions to achieve specific goals. Belief-Desire-Intention (BDI) (Rao et al., 1995) architecture serves as a foundational framework for modeling intelligent agents and is widely acknowledged in this discipline. The BDI framework divides their cognitive structure into three key components: beliefs, desires, and intentions. Desires represent the goals that the intelligent agent aims to achieve. These goals drive the agent's decision-making processes, motivating it to take specific actions in pursuit of desired outcomes. Intentions, represent the agent's concrete plans and decisions to perform certain actions based on its understanding of the environment and its goals. Finally, beliefs encapsulate the agent's knowledge about its environment. These encompass a range of information, including facts, perceptions, and interpretations of the surrounding context, including the emotional state of the individuals interacting with the BDI agent.

LLMs are computational models that learn the structure and patterns of language from vast amounts of text data. The evolution of LLMs is closely tied to the emergence of transformer architectures. Specifically, work developed on attention-based transformer models overcame the limitations associated with traditional recurrent and convolutional neural networks (Vaswani et al., 2017). That attention-based mechanism enabled transformers to capture long-range dependencies and parallelize computations effectively, making them highly efficient for processing sequences of information. The transformer architecture's success enabled the development of powerful LLMs such as OpenAI's GPT (Generative Pre-trained Transformer) series and BERT (Bidirectional Encoder Representations from Transformers) (Zhang et al., 2022; Alaparthi and Mishra, 2021). These models demonstrated remarkable language understanding capabilities, leading to breakthroughs in various NLP tasks. A novel and promising application of these models is the detection of emotions from the text of a human-human or human-machine conversation. All this makes these models suitable for evolving the way we interact with intelligent agents.

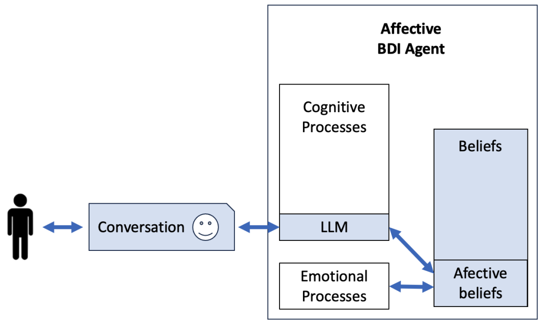


Figure 1: Incorporation of a LLM for emotion recognition in the affective multiagent architecture GenIA[3].

2.1 NLP Models for Emotion Recognition

The field of NLP has been progressing and improving since its beginnings addressing increasingly complex tasks. NLP models first focused on sentiment analysis (Kouloumpis et al., 2011; Nasukawa and Yi, 2003), which is concerned with identifying the overall sentiment (positive, negative or neutral) conveyed in a text (Ghosh et al., 2015). In recent years, machine and deep learning approaches have improved achieved satisfactory results in the emotion recognition task, in which specific emotions have to be classified.

The advent of transformer architectures has brought about a paradigm shift in the domain of emotion classification. Early transformer models, most notably BERT, were initially designed for diverse tasks among natural language processors, but can be specialized in specific tasks such as the sentimental analysis of text (Alaparthi and Mishra, 2021). These models have also allowed significant advances in the detection of possible emotions implicit in the text (Cortiz, 2022; Adoma et al., 2020). For instance, BERT, introduced in (Devlin et al., 2019), significantly improved performance in various NLP tasks, and can be used for sentiment analysis and emotion recognition. EmoBERTa (Kim and Vossen, 2021), a model based on RoBERTa (Liu et al., 2019), differs from its predecessor by its pre-training specialized in the detection of emotions, which was improved and was the cause of the model's better performance.

Currently, LLMs offer the advantage of being pre-trained on large linguistic datasets, allowing them to capture the nuances of human expression. Furthermore, the inherent flexibility of text-generating LLMs allows them to adapt to diverse emotional contexts without the need for explicit training on emotion-specific datasets. This adaptability coupled with the combined ability to generate contextually relevant, personalized, and emotionally resonant responses positions them as valuable tools for understanding and interpreting emotions in text-based conversations.

Our study focuses on utilizing LLMs designed for text generation as a tool for emotion recognition in dialogues between a BDI affective multi-agent architecture, GenIA[3] (Alfonso et al., 2017; Taverner et al., 2019), and an individual. We show in Figure 1 the functional design of this BDI affective agent architecture. The conversation between the affective agent BDI and its human interlocutor takes place by means of a module consisting of an LLM. In order for the affective reasoning component of the agent to reason, it needs to dispose of the knowledge of the affective state of its interlocutor. The mode of representing knowledge in a BDI agent is by means of beliefs. For this reason, the function of the LLM module will be to detect the implicit emotions in the conversational text produced by the human interlocutor. These emotions, represented by a label, are translated into a belief and sent to the emotional belief base of the affective BDI agent.

In the subsequent sections, we outline our methodology for employing text-generating LLMs in the emotion recognition task, and a comparative study evaluating the effectiveness of various LLMs in order to select the LLM or LLMs best suited to the task.

3 LLMs FOR EMOTION RECOGNITION

As mentioned above, in this study we focus on the belief component as a first step towards enabling intelligent agents to act effectively in the emotional domain. This first stage consists of exploring and developing systems capable of understanding conversations with human beings, not only at a superficial or content level, but also by unraveling the emotional states present underneath them. Text-generating LLMs promise to be a potential tool to achieve this.

3.1 Methodology

3.1.1 Emotion Recognition Task

The Emotion Recognition task critically depends on several considerations to ensure accurate and meaningful results. One key factor is the careful selection of prompts during interactions with text-generating LLMs. Prompting plays a pivotal role in influencing the model's ability to classify responses effectively into predefined emotional categories. We emphasize the need for meticulous prompt design, as it is fundamental to guide the model to generate responses in a format favorable to an accurate classification of emotions.

Contextual information is also a critical component in deciphering the emotional content of a message. This is specially true in text-based interactions where non-verbal cues are absent. Leveraging the contextual understanding capabilities of LLMs, we hypothesize that providing historical conversation context enhances the model's ability to recognize and generate emotionally appropriate responses.

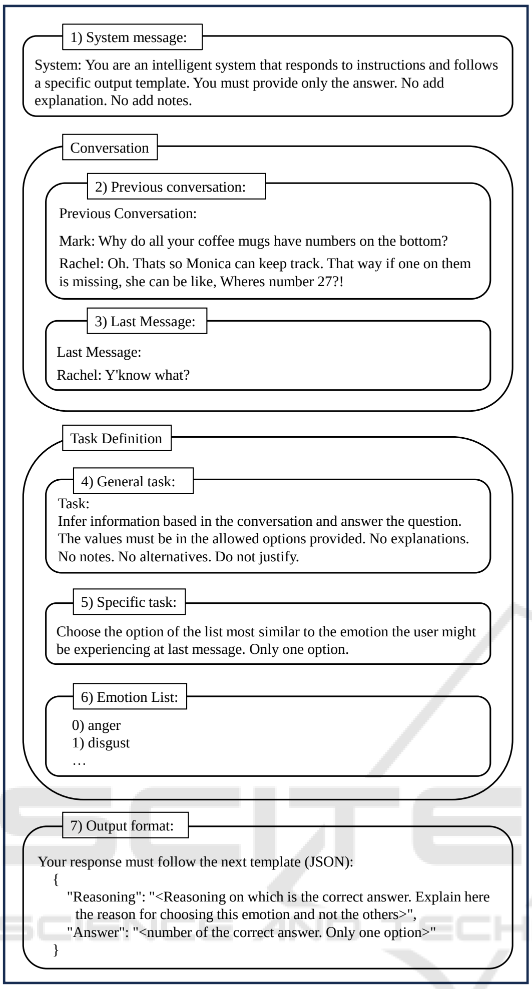
Furthermore, we delve into the concept of "reasoning" by LLMs in the context of emotion recognition. Unlike traditional approaches, LLMs exhibit a form of reasoning linked with text generation. To exploit this, we induce the model to generate a coherent line of reasoning prior to formulating its answer that serves as a mechanism to interpret and contextualize the emotional hints of the dialogue, enhancing the generation of more accurately predictions.

It is crucial to note that in this Emotion Recognition task scenario we must specify a list of possible emotions. This list is flexible and can be adapted according to the needs of the context in which it is implemented. In this study, we have used two different sets of emotions, adjusting to the characteristics of each particular dataset we have used in the evaluations. The expected outcome of the model must be one of emotions specified in the list.

3.1.2 Prompt Building

In our exploration of text-generating LLMs for emotion recognition in dialogues, the methodology for prompt building played a crucial role. The design of effective prompts is fundamental to eliciting targeted emotional responses from LLMs.

The structured prompt scheme employed in the study consists of several key elements. It begins with a system message that sets the stage for the task, followed by the inclusion of the previous conversation to provide contextual information. The last message is explicitly specified, and the LLM is guided to classify the specific emotion by a general task definition, a specific task definition, and a list of emotion categories. Finally, the desired output format is determined, with 2 fields: reasoning and answer. Later, the answer is structured using a prolog-style approach to generate an affective agent belief as $emotion(emotion\_label)$, where emotion_label repre-



1) System message:

System: You are an intelligent system that responds to instructions and follows a specific output template. You must provide only the answer. No add explanation. No add notes.

Conversation

2) Previous conversation:

Previous Conversation:

Mark: Why do all your coffee mugs have numbers on the bottom?

Rachel: Oh. Thats so Monica can keep track. That way if one on them is missing, she can be like, Wheres number 27?!

3) Last Message:

Last Message:

Rachel: Y'know what?

Task Definition

4) General task:

Task:
Infer information based in the conversation and answer the question.
The values must be in the allowed options provided. No explanations.
No notes. No alternatives. Do not justify.

5) Specific task:

Choose the option of the list most similar to the emotion the user might be experiencing at last message. Only one option.

6) Emotion List:

0) anger
1) disgust
…

7) Output format:

```
Your response must follow the next template (JSON):
   {
      "Reasoning": "<Reasoning on which is the correct answer. Explain here
       the reason for choosing this emotion and not the others>",
      "Answer": "<number of the correct answer. Only one option>"
   }
```

Figure 2: Prompt example for emotion recognition.

sents one of the emotions specified in the prompt. Figure 2 shows an example of the structured prompt applied across all LLMs in our comparative study.

3.1.3 Text-Generating LLMs Selected

In this section, we introduce the text-generating LLMs selected for our comparative study on emotion recognition in dialogues. The criteria followed for the model selection has been the relevance of the models in the current moment of this study and their endorsement by reputable companies or research institutes in the field. An attempt has also been made to provide variety in the study, so models of different sizes, trained for different purposes, have been chosen. Thus, the final set of models includes models of 3B, 7B and 13B parameters (3, 7 and 13 billions of parameters, respectively), some being pre-trained for chat, others for following instructions and others for reasoning skills.

GPT 3.5, also known as ChatGPT, represents one of the prominent models in the GPT series developed by OpenAI. This model has been chosen for its exceptional performance, making it one of the leading LLMs available to the public.

Llama 2 (Touvron et al., 2023) is a family of LLMs developed and publicly released by Meta. Although we use the smaller versions (7B and 13B parameters) due to resource constraints and the need for speed in the task, this model series actually ranges from 7 billion to 70 billion parameters. Specifically, the fine-tuned LLMs within the Llama 2 family, known as Llama-2-Chat, are tailored for dialogue use cases. At the time of their release, these models exhibited superior performance over open-source chat models across multiple benchmarks. These models have been chosen because they have been an important step in the construction of open-source LLMs and numerous models have been derived from them.

Orca 2 (Mitra et al., 2023) is a model developed by Microsoft whose base model is Llama 2. It is a research-oriented model tailored for tasks such as reasoning, reading comprehension, math problem solving, and text summarization. This model is available in both a 7 billion parameter configuration and a 13 billion parameter version. While it is not explicitly optimized for chat, it has the capability to perform in that domain and showcases advanced reasoning abilities.

Mistral 7B (Jiang et al., 2023) is a novel model developed by Mistral AI that includes new features that have made it achieve a good performance, matching or surpassing other models of even larger size. This features are the utilization of Grouped-Query Attention (GQA) for expedited inference and Sliding Window Attention (SWA) to effectively handle sequences of arbitrary length with reduced inference costs. We are using the instruction fine-tuned versions, Mistral 7B Instruct 0.1 and its new enhanced version 0.2.

Zephyr 7B β is part of a series of language models designed as helpful assistants. Notably, Zephyr-7B sets at its release a new benchmark in chat models for 7B parameter models, surpassing Llama2-Chat-70B, and excels in intent alignment.

StableLM Zephyr 3B is a lightweight LLM developed by Stability AI. The model is an extension of the pre-existing StableLM 3B-4e1t model and is inspired by the Zephyr 7B model. With 3 billion parameters, this model effectively satisfies a wide range of text generation needs, from simple queries to complex instructional contexts on edge devices.

3.1.4 LLMs Quantization

Quantization is a technique used in the field of deep learning to reduce the size of models, speed up inference and improve computational efficiency. In the case of large language models, quantization can be applied to the neural network weights. Instead of representing each weight with full precision, fewer bits can be used to represent them. This significantly reduces the size of the model and speeds up inference operations, although there may be some loss of accuracy.

In order to incorporate LLMs as an emotion recognition tool on our available hardware, and assuming that reducing the time and resources required is vital for its general use in intelligent agents, in the present study we quantify the models used (with the exception of GPT 3.5 because it is not possible for us). For a fair evaluation of the models, regardless of their size, they have all been quantized with the same characteristics using AutoGPTQ. These are a bit size of 4 bits, a group size of 32g and utilizing act order.

3.2 Experiment Design

3.2.1 Datasets

In this experiment, the primary objective is to assess the capability of text-generating LLMs in recognizing emotions during conversations. Unlike traditional methods, the models being compared have not been explicitly trained for this task or undergone pretraining on the specific datasets used in the tests. The approach involves directly evaluating the models' performance on selected datasets to take advantage of their flexibility. The datasets used for evaluating the models' performance include:

- **MELD:** Multimodal EmotionLines Dataset (Poria et al., 2019) is a dataset for emotion recognition that combines text, audio and video extracted from the Friends TV series. In this study, we exclusive focus in the analysis of the textual component. Each utterance is labeled with one of the following emotions: anger, disgust, sadness, joy, surprise, fear and neutral.
- **Topical Chat:** Topical Chat (Gopalakrishnan et al., 2023) is a dataset that consists of conversations between knowledgeable people on eight broad topics, with no explicitly defined roles for the participants. The emotion labels included are: angry, disgusted, sad, happy, surprised, fearful, curious, and neutral.

The selection of these datasets is based on their diversity and relevance to the task of emotion recognition in dialogues, aiming to evaluate the adaptability of LLMs to diverse emotional contexts present in everyday conversations.

3.2.2 Materials

The experiments in this study were conducted using a high-performance computing setup to effectively train and evaluate text-generating LLMs. The specified hardware resources for these experiments is composed of a NVIDIA A40 (48GB VRAM) GPU, an AMD EPYC 7453 28 cores processor with and 512 GB of RAM.

3.2.3 Metrics

To systematically evaluate the performance of the selected text-generating LLMs in emotion recognition, we employ the following metrics:

- **Accuracy:** Accuracy represents the ratio of correctly predicted emotions to the total number of instances in the dataset. It is estimated as:

$$\mathrm{A} = \frac{TP+TN}{N} \tag{1}$$

where TP represents the number of true positives, TN the number of true negatives, and N is the total number of instances.

- **Precision:** Precision represents the ratio of true positives to the total number of positive predictions made. It is calculated as:

$$\mathrm{P} = \frac{TP}{TP+FP} \tag{2}$$

where FP the number of false positives.

- **F1 Score weighted :** The F1 Score is the harmonic mean of precision and recall. It provides a balanced measure that considers both false positives and false negatives, offering a consolidated view of the model's performance. It is measured as:

$$F1\,\mathrm{score} = \frac{2\cdot\mathrm{Precision}\cdot\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{3}$$

where Precision is the precision metric explained before and Recall is sensitivity of the model (number of true positives divided by the sum of true positives and false negatives). In the case of multiclass classification, it is calculated as the weighted average of the individual F1-scores, where the weight of each class is determined by the proportion of instances of that class to the total:

$$F1_{\mathrm{weighted}} = \frac{\sum_{i=1}^{C} w_i \cdot F1_i}{\sum_{i=1}^{C} w_i} \tag{4}$$

where C is the class number, $F1_i$ is the F1 score of the class i and w_i is the weight assigned to the class i.

3.3 Results

In this section, we present the results of our evaluation of the performance of the selected text-generating LLMs on emotion recognition. Table 1 and Table 2 show the values of the metrics that each model obtained with each data set. Table 3 shows the average time in seconds each model took to perform the task.

Table 1: Performance Metrics for MELD dataset.

MELD dataset			
LLM	Accuracy	Precision	F1 weighted
GPT 3.5	34.87	**56.78**	31.81
Llama 2 7B	28.66	48.13	25.78
Llama 2 13B	27.20	51.64	23.82
Orca 2 7B	36.36	49.42	36.93
Orca 2 13B	34.06	50.36	33.66
Mistral Instruct 0.1 7B	33.10	50.76	30.57
Mistral Instruct 0.2 7B	**46.63**	53.80	**47.90**
Zephyr 7B	22.61	44.48	17.41
StableLM Zephyr 3B	32.26	49.22	32.46

Table 2: Performance Metrics for Topical Chat dataset.

Topical Chat dataset.			
LLM	Accuracy	Precision	F1 weighted
GPT 3.5	31.30	40.86	**33.46**
Llama 2 7B	28.00	39.50	29.02
Llama 2 13B	25.44	37.53	27.29
Orca 2 7B	**33.02**	40.06	33.29
Orca 2 13B	30.16	**42.74**	27.86
Mistral Instruct 0.1 7B	31.16	40.12	32.07
Mistral Instruct 0.2 7B	32.98	41.79	32.65
Zephyr 7B	27.57	41.44	27.44
StableLM Zephyr 3B	28.36	38.26	27.21

Table 3: Average time in seconds for emotion recognition (using previous reasoning) for each dataset.

Average time for emotion recognition			
LLM	MELD	Topical Chat	Average
GPT 3.5	2.11	1.99	2.05
Llama 2 7B	2.12	1.85	1.99
Llama 2 13B	2.91	2.34	2.63
Orca 2 7B	2.47	2.19	2.33
Orca 2 13B	2.88	3.22	3.05
Mistral Instruct 0.1 7B	1.69	1.85	1.77
Mistral Instruct 0.2 7B	2.02	1.62	1.82
Zephyr 7B	2.56	2.32	2.44
StableLM Zephyr 3B	**1.55**	**1.43**	**1.49**

Regarding the MELD dataset we can see that GPT 3.5 obtains the best value in Precision, but is still outperformed by Mistral Instruct 0.2 in both accuracy and F1 weighted.

As for the Topical Chat dataset, with GPT 3.5 being the best performing model based on the F1 weighted metric. This indicates that there can be large differences in the performance of the models depending on the domain. For this case, the best of the open-source models is Orca 2 7B, achieving the second best F1 weighted and the best accuracy. However, the model with the best precision in this case is Orca 2 13B.

As for the average execution time, there is a direct correlation between the size of the model and the time required for the task. Thus, we find that the fastest model is the lightest one. For the rest, an exception can be noted for the Mistral models, which achieve a shorter execution time than the rest of the models with the same number of parameters.

4 DISCUSSION

Based on the results shown in the previous section, can we affirm that LLMs can potentially be a useful tool for emotion recognition? And if so, is it a valid tool to be integrated in intelligent agents?

Although all the metrics used in the experiment are of interest to evaluate the suitability of an LLM for emotion recognition, we consider the F1-score weighted metric is the most important for the selection of the LLM model for our BDI agent. The F1-score is the metric that best reflects the performance of a LLM for this multiclass classification since the quantity of utterances of each of the emotions in the datasets used in the experiment is not balanced and this metric is the only one that is immune in these cases.

In view of the results obtained, two models stood out above the rest in the emotion recognition task, being capable of generating emotion beliefs with a good ratio correct predictions. These are Mistral Instruct 0.2 7B, which showed the best performance for the MELD dataset, and Orca 2 7B, which obtained the best score for the Topical Chat dataset. It is important to note that both models are the second best performers for the other dataset.

Through the results, we can observe that lighter models have shown good performance. On the one hand the open-source models shown of 3, 7 and 13 billion parameters can be compared with GPT 3.5 having a performance equal or superior to this one, despite GPT 3.5 being a very large model (although of unspecified size for the best of out knowledge). On the other hand, because for both two pairs of models we have with different size versions (7 and 13 billion), Llama 2 models and Orca 2 models, the 7 billion parameter versions have demonstrated better performance in both datasets. This adds to the good performance shown by the lighter model, StableLM Zephyr 3B, which has achieved metrics surpassing several models for the MELD dataset. This may indicate that for specific tasks, such as emotion recognition, the number of parameters is not a determining factor and may even be a counterpart.

Despite this, these results are not superior to to the state of the art, but this is not the focus of the current study. It should be remembered that this is a preliminary study in which we have evaluated the understanding of this type of models for the emotional context by the nature of their massive pre-training.

We have to emphasize that our study employs LLMs without specific training for emotion recognition. To the best of our knowledge, these models haven't been trained specifically for the task of recognizing emotions, and we haven't retrained them with the specific datasets used in our comparative study. Therefore the results of our experiment are Zero-shot predictions, as the models are not even provided with classification examples in the prompt. Considering these conditions and that the MELD and Topical Chat datasets have 6 and 7 categories of emotions to classify respectively, we can conclude that they have transversely acquired a certain level of emotion recognition skills in their pre-training. In general, considering the results of the experiment, we can conclude that, the use of text-generating LLMs for emotion recognition is valid and it is an excellent starting point for further improvements by means of retraining or fine-tuning these models specifically for this task.

Once it has been proven that LLMs are a potentially suitable tool for emotion recognition in conversational contexts, now we wonder if the high computational costs of these models are suitable for intelligent agents in their interaction. The use of LLMs for emotion recognition in affective intelligent agents will be valid in those cases where the two main disadvantages of this approach are not present. The first is that these models are computationally expensive, so hardware with sufficient performance is required. The second disadvantage is the time needed to obtain a response. Although the average time required by these models to recognize emotions in conversations is relatively short, around 2 seconds, it may not be short enough for systems acting in real time, because this time will be added to that of the rest of the agent's processes, delaying its response. Unfortunately, a user waiting for a response may find this time unacceptable

However, both of these drawbacks are likely to be mitigated as the field progresses, given the ongoing development and optimization of smaller language models. As an illustration of this trend, it is noteworthy that not only do we have these 7 billion parameter models, but there is also the availability of StableLM with 3 billion parameters, which demonstrates competitive performance.

In addition to the emergence of smaller language models, various techniques are being developed to execute such models on lighter hardware and/or with reduced time requirements. An example of this techniques is quantization, a method that demands less VRAM to load the model and requires less execution-time, although at a slight cost to accuracy. For intelligent agent systems, the best balance between model accuracy and time and space requirements should be pursued.

5 CONCLUSIONS AND FUTURE WORK

In this work we have proposed the use of LLMs for the recognition of emotions in a conversation between a individual and an intelligent affective BDI agent. We have used the promting technique for the LLM to generate beliefs with the detected emotion to be inserted into the belief base of a BDI agent in the GenIA^3 architecture.

The success of Zero-shot predictions suggests that these models can serve as a foundation for future endeavors in retraining or fine-tuning, specifically targeting emotion recognition tasks. LLMs show promising capabilities in both recognition and belief generation. Mistral Instruct 0.2 7B and Orca 2 7B are the best candidates to be trained in emotion recognition and employed by our affective BDI architecture. We recommend Mistral Instruct 0.2 7B because of its good task performance and lower time cost. For contexts where a shorter response time is needed, the lightest and therefore fastest model is StableLM Zephyr 3B.

Future work could involve training the selected models using emotion-labeled datasets to enhance their performance and adapt them to the specific requirements of the GenIA^3 architecture. Further research could focus on identifying the optimal trade-off between model size, response time, and accuracy. Finally, we should delve into the ethical implications of deploying emotion recognition systems, ensuring fairness, transparency, and mitigating potential biases.

ACKNOWLEDGEMENTS

Work partially supported by Generalitat Valenciana CIPROM/2021/077, Spanish Government projects PID2020-113416RB-I00 and TED2021-131295B-C32; TAILOR project funded by EU Horizon 2020 under GA No 952215; and TED2021-131295B-C32.

REFERENCES

Adoma, A. F., Henry, N.-M., and Chen, W. (2020). Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *17th ICCWAMTIP*, pages 117–121. IEEE.

Alaparthi, S. and Mishra, M. (2021). Bert: A sentiment analysis odyssey. *Journal of Marketing Analytics*, 9(2):118–126.

Alfonso, B., Vivancos, E., and Botti, V. (2017). Toward formal modeling of affective agents in a BDI architecture. *ACM Trans. on Internet Technology*, 17(1):5.

Cortiz, D. (2022). Exploring transformers models for emotion recognition: a comparision of bert, distilbert, roberta, xlnet and electra. In *3rd Int. Conf. on Control, Robotics and Intelligent System*, pages 230–234.

de Melo, C., Gratch, J., and Carnevale, P. (2014). The importance of cognition and affect for artificially intelligent decision makers. In *Proc. of the AAAI'14*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc.NAACL-HLT 2019*, page 4171–4186.

Fan, L., Scheutz, M., Lohani, M., McCoy, M., and Stokes, C. (2017). Do we need emotionally intelligent artificial agents? first results of human perceptions of emotional intelligence in humans compared to robots. In *17th Int. Conf. on Intelligent Virtual Agents*, pages 129–141. Springer.

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., and Reyes, A. (2015). Sentiment analysis of figurative language in twitter. In *Proc. 9th Int. Workshop on Semantic Evaluation*, pages 470–478.

Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R., and Hakkani-Tur, D. (2023). Topical-chat: Towards knowledge-grounded open-domain conversations. *arXiv preprint arXiv:2308.11995*.

Iqbal, T. and Qureshi, S. (2022). The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences*, 34(6):2515–2528.

Irfan, B., Narayanan, A., and Kennedy, J. (2020). Dynamic emotional language adaptation in multiparty interactions with agents. In *Proc. 20th ACM Int. Conf. on Intelligent Virtual Agents*, pages 1–8.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kim, T. and Vossen, P. (2021). Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*.

Kouloumpis, E., Wilson, T., and Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! In *Proc. of the Int. AAAI Conf. on Web and Social Media*, volume 5, pages 538–541.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., and Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Mitra, A., Del Corro, L., Mahajan, S., Codas, A., Simoes, C., Agarwal, S., Chen, X., Razdaibiedina, A., Jones, E., Aggarwal, K., et al. (2023). Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

Nagarhalli, T. P., Vaze, V., and Rana, N. (2021). Impact of machine learning in natural language processing: A review. In *2021 3rd Int. Conf. on Intelligent Communication Technologies and Virtual Mobile Networks*, pages 1529–1534. IEEE.

Nasukawa, T. and Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proc. 2nd Int. Conf. on Knowledge capture*, pages 70–77.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proc.57th Annual Meeting of Association for Computational Linguistics*, page 527–536.

Rao, A. S., Georgeff, M. P., et al. (1995). BDI agents: from theory to practice. In *Icmas*, volume 95, pages 312–319.

Rincon, J., Bajo, J., Fernandez, A., Julian, V., and Carrascosa, C. (2016). Using emotions for the development of human-agent societies. *Frontiers Information Technology & Electronic Engineering*, 17(4):325–337.

Taverner, J., Vivancos, E., and Botti, V. (2019). Towards a computational approach to emotion elicitation in affective agents. In *Proc. ICAART'19*, pages 275–280.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. (2022). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.