

Answer Rephrasing and Emotion Tracking in a Multilingual Customer Service Chat

Luis Alves
Instituto Superior Técnico
luis.c.alves@tecnico.ulisboa.pt

Abstract—As conversational agents are becoming widespread, tasks such as Emotion Recognition in Conversations (ERC) and Empathetic Dialogue Generation are gaining increasing importance. Our work addresses empathetic generation for customer support applications, a domain that is much less studied and where such task modules can empower human agents to provide a better interaction experience by promoting empathetic behaviour. First, we leverage the conversational context by feeding a RoBERTa-based model with the utterance to classify and the context of the conversation. Secondly, we investigate how the number of turns of context used influences the ERC module performance. We validate our results on the customer support MAIA dataset annotated with emotions and empathy information, for which the introduction of a certain number of conversational turns significantly improves model performance (a 3-point increase in Macro F1-score). Thirdly, we compare this approach with ChatGPT which is not fine-tuned for this task, highlighting the importance of fine-tuning as it considerably improved performance. Finally, we develop an empathetic rephrasing system by employing ChatGPT to rephrase text, aiming to enhance its empathy while preserving the original intended content. We perform automatic and human evaluation which shows that rephrasing improves empathy on the agents' side.

Index Terms—Emotion Recognition; Empathetic Rephrasing; Prompting.

I. INTRODUCTION

Dialogue systems have witnessed significant advancements in recent years in Natural Language Processing (NLP), mainly in the field of deep learning [1]. Not surprisingly, one of the branches that have been highly addressed is the field of Empathetic Dialogue Generation, since empathy is a key mindset and behaviour that supports an individual's active involvement in diverse dialogues [2].

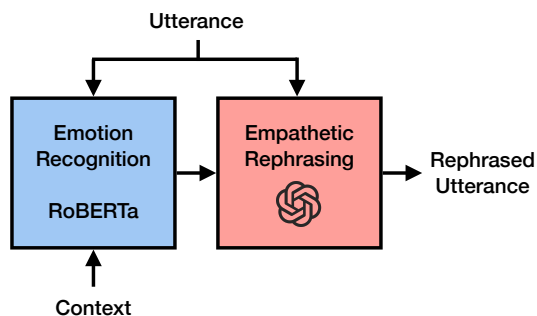


Fig. 1. Framework Architecture

Empathy is an emotion related to one's cognitive or emotional ability to share the emotions felt by another upon observance, increasing the likelihood of more successful communicative interactions and often leading to prosocial actions, such as offering help, support, or comfort to others [3]. The customer support agent should be empathetic, helping and solving user problems in a cooperating manner.

Moreover, given the strong connection between empathy and the capacity to understand and share the emotions of the other, it is crucial to initially identify the emotions expressed throughout a conversation. Subsequently, the development of an ERC module is important in order to accurately identify emotions within a conversation. ERC can be used in several applications such as automatic opinion mining and as an assisting module in therapeutic practices [4]. In customer support, it can measure customer and agent emotions to provide feedback for the agent in order to promote a better interaction experience and foster empathetic behaviour.

Aiming to empower customer support agents, we thus propose to tackle the tasks of Emotion Recognition in Conversations and Empathetic Dialogue Generation with the model from figure 1, on a customer support dataset, MAIA [5], equipped with emotion annotations and empathy information, enabling empathy's automatic evaluation. Our key contributions are:

- For the ERC task, the comparison of a relevant model for emotion prediction in predicting emotions in a customer service context.
- The study of the effect of the number of introduced context turns in the performance of our ERC module
- Comparison of a model fine-tuned in the MAIA dataset for the ERC task with prompting an LLM, ChatGPT
- Empathetic rephrasing using LLMs, thanks to its strong ability to generate text.

II. RELATED WORK

A. Emotion Recognition in Conversations

Knowledge and understanding of the conversational context, i.e., the previous conversational turns, are extremely valuable for identifying the emotions of the interlocutors [4]. Therefore, we describe approaches that leverage the conversational context in ERC.

The work from [6] leverages Long Short-Term Memory networks to extract contextual features from the utterances. DialogueRNN [7] uses three Gated Recurring Units (GRU)

in the classification module to model several aspects of the conversation.

One major problem in using RNNs is the long path of information flow, which difficult the capture of long-term dependencies. These dependencies can be better captured with the Transformer architecture [8] which has a shorter path of information flow. The Knowledge-Enriched Transformer (KET) [9] leverages such architecture.

Given a conversation’s sequential nature, work was also developed using explicit sequential-based methods. [10] proposes a model that uses a CRF [11] in order to learn the emotional coherence in the conversation.

Moreover, since recurrence-based methods had the limitation of updating the query utterance’s state with relatively limited information from nearby utterances, graph-based methods were proposed as an alternative to cope with this limitation [12].

In order to overcome the problems that both recurrence-based methods and graph-based methods could not solve, a hybrid approach was proposed in [13] that tries to take advantage of the strong points of both architectures, leaving behind the weaknesses of those models. DAG-ERC [13] proposes a directed acyclic graph in order to better model the spread of information during a conversation.

Recently, [14] and [15] studied the emotional capacities of a powerful LLM such as ChatGPT, which came to revolutionize NLP due to its great performance in a wide range of tasks, coming as no surprise the use of this model in the Emotion Recognition task.

B. Empathetic Dialogue Generation

[16] proposed *EmpatheticDialogues*, a dialogue dataset specifically designed for the Empathetic Dialogue Generation task, that has served as a valuable resource for evaluating numerous approaches in this domain. Various techniques have been employed to enhance Dialogue Generation systems with empathetic capabilities, with a common focus on leveraging the emotions expressed within the conversations. MoEL [17] proposes a multi-decoder approach, where each decoder is useful to generate a response for a specific context emotion. MIME [18] relies on mimicking the context emotion to a varying degree, arguing that it is crucial for generating empathetic responses. Subsequently, various alternative methods emerged, namely, approaches using Reinforcement Learning, such as [19] and [20] or relying on the emerging power of LLMs, leveraging GPT-based architectures to generate empathetic text, such as [21] and [22]. In addition, strategies employing graph-based techniques to model emotional cause [23] and emotional context [24] were proposed. However, none of these approaches has yet integrated automated mechanisms for evaluating the empathetic value present in the generated dialogue.

C. Empathy Evaluation

Various approaches have been explored to evaluate empathy, although many of them are not specifically tailored for assessing empathy in dialogues but rather for other text types. One example is the WASSA shared task. In 2022, this shared

task [25] used a dataset labelled with empathy, focusing on classifying essays, and assessing their performance by evaluating the empathy using the Pearson correlation coefficient. Recently [26] introduced the first approach to automatically evaluate empathy in dialogues, a new metric called EMP-EVAL, employing a dataset centred around dialogues related to the mental health domain. To validate the quality of the metric it was compared to empathy’s human evaluation.

III. MAIA DATASET

The MAIA dataset [5] and its emotion annotations have a unique value in the field of customer support and conversational agents. It can be used as a training and benchmark dataset for emotion and empathy classification in these settings, as a base for the construction of dialogue systems that support customer-agent interaction processes. Classification models trained on this dataset could assist customer service agents by measuring customer and agent emotions to provide feedback for the agent.

The dataset consists of a total of 612 dialogues with a total of 24960 sentences. Figure 1 illustrates its emotions and distribution.

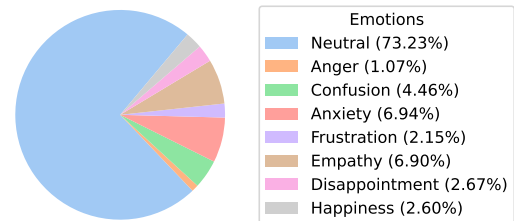


Fig. 2. Emotion distribution of the MAIA dataset

IV. EMOTION RECOGNITION IN CONVERSATIONS

Regarding the employed models, two distinct approaches were adopted. Firstly, we utilized the model outlined in [27], which leverages a RoBERTa model to perform ERC. This model incorporates the conversational context to enhance text comprehension, thereby improving classification outcomes as shown in [27]. However, it is important to note that this prior work did not assess the model’s performance in a customer support context, prompting our investigation in this specific environment.

Additionally, the ChatGPT API was employed, chosen for its exceptional performance in diverse NLP tasks [28]. This decision aimed to facilitate a comparison between a model pre-trained specifically for the task at hand, as demonstrated in [27], and one that exhibits strong overall performance across a range of NLP tasks but has not undergone specific training for this particular task.

A pivotal distinction between these two approaches lies in the methodology employed. For the model detailed in [27], our approach involves fine-tuning the model, whereas when employing ChatGPT, we adopted a different strategy. Specifically, we prompted the model with the sentence intended for

classification, accompanied by relevant context and examples similar to the sentence being classified.

A. RoBERTa-based model

Using the [27] model we investigate how the number of introduced context turns influences our model performance. The results of this investigation are presented in Table I. The displayed values represent the average of three runs, along with the corresponding standard deviation.

When investigating performance in relation to context, an evident trend emerges: augmenting the count of context utterances improves the model’s performance up to a certain point, where peak performance is achieved before a subsequent decline. From Table I it can be observed that the number of context turns that maximizes the Macro-F1 score and most of the individual emotion scores is 4.

The absence of a universally optimal count of context utterances is evident in Table I. Consequently, determining the optimal model depends on the specific emotion under scrutiny, due to the closely matched Macro-F1 scores across models. Additionally, it is worth noting that for the non-neutral emotions, the classifier yields the highest F1-score for the emotion empathy, suggesting that it is a suitable empathy automatic evaluation classifier.

B. ERC task using ChatGPT

Besides using [27] to classify the emotions in the MAIA dataset, a comparison was also made with ChatGPT to understand the fluctuation of performance between a fine-tuned model and an LLM not tuned for the task.

To assess the effectiveness of ChatGPT in the ERC task, we conducted various tests in different scenarios. Specifically, we provided the model with the context of the conversation and included examples similar to the utterance being classified. We experimented with different configurations, including varying the number of utterances from the conversation context and adjusting the number of retrieved similar examples. The objective was to identify the optimal setup that yielded the highest performance results.

To retrieve the examples that exhibit a higher degree of similarity to each individual utterance to classify we use LabsE model [29], which is a model based on BERT’s architecture. This model, instead of being able to effectively learn monolingual sentence embeddings for semantic similarity, is able to learn multilingual embeddings, which is useful in this situation since the model is able to find similar examples from different languages.

The following is the prompt employed to obtain the results presented in Table II:

- *You are an emotionally intelligent assistant for customer support. Classify the emotion of the utterances with AT MOST ONE OF THE FOLLOWING EMOTIONS: [Emotions List].*

This is the format of the interaction:

"Context: [Previous Dialogue]

(Client/Agent):[Utterance to classify]

	Macro-F1	Emp	Hap	Disa	Conf	Frus	Ang	Anx	Neu	Inv-F1
c=0	42.50±1.06	67.44±0.66	43.22±0.35	23.6±4.26	38.17±3.41	16.72±8.67	7.77±4.52	53.24±2.71	89.83±0.53	38.45±2.39
c=1	43.46±0.65	65.93±0.47	45.02±2.18	31.18±1.36	36.22±1.31	17.27±1.94	11.24±6.06	50.58±2.26	89.92±0.33	40.48±1.41
c=2	41.32±3.05	68.31±1.28	43.57±2.99	22.32±18.65	31.02±6.58	14.75±4.24	6.35±5.51	54.12±2.22	90.16±0.26	36.73±3.86
c=3	42.28±4.62	69.32±3.13	40.77±3.52	23.86±16.89	35.26±8.94	15.13±5.75	4.89±6.34	58.47±1.16	90.52±0.42	37.68±5.40
c=4	45.50±1.9	70.55±1.79	44.86±1.92	36.06±5.03	42.95±1.30	16.80±1.62	8.18±3.69	53.76±5.02	90.86±0.14	41.77±1.74
c=5	42.44±2.50	69.88±2.31	42.63±1.87	24.37±20.13	40.60±1.10	12.45±6.28	1.67±2.89	57.40±0.57	90.52±0.59	37.88±2.72
c=6	44.56±1.10	70.74±0.1	46.15±2.32	32.59±5.94	42.14±2.25	11.43±3.03	7.16±2.35	55.62±4.25	90.68±0.40	40.32±1.04
c=7	44.02±1.58	69.28±2.89	45.94±0.2	33.41±4.90	41.16±1.02	8.19±3.87	7.38±6.69	56.15±3.81	90.60±0.57	40.14±2.67
c=8	43.60±3.67	68.93±3.99	46.58±4.21	23.48±19.14	42.03±1.44	16.48±8.21	3.78±4.52	56.81±2.86	90.64±0.76	39.33±3.73
c=9	42.60±4.24	66.81±7.26	46.43±6.05	24.19±17.68	39.01±2.77	13.21±8.87	5.32±4.72	55.46±3.76	90.37±0.36	38.29±4.80
c=10	42.98±2.98	67.44±4.2	45.39±5.45	26.56±13.94	40.81±2.24	13.60±7.53	3.19±2.77	56.46±1.38	90.40±0.60	39.12±2.96

TABLE I

RESULTS OF EMOTION RECOGNITION IN CONVERSATIONS USING A ROBERTA-BASED MODEL WITH THE DATASET LABELLED WITH EMPATHY. c IS THE NUMBER OF INTRODUCED CONTEXT TURNS.

	Macro-F1	Emp	Hap	Disa	Conf	Frus	Ang	Anx	Neu	Inv-F1
c = 0, e = 0	28.99	31.25	23.66	31.39	29.72	18.72	10.26	3.33	83.62	31.45
c = 0, e = 5	38.55	56.20	30.73	31.61	34.86	25.64	11.90	30.45	87.00	40.54
c = 0, e = 10	40.90	53.68	34.33	33.70	36.49	29.27	15.73	37.16	86.79	43.44
c = 0, e = 20	42.65	54.81	37.78	33.90	40.48	29.68	17.20	40.36	86.99	46.08
c = 5, e = 0	28.42	22.97	22.69	33.48	34.21	17.02	5.13	7.49	84.37	31.19
c = 10, e = 0	29.02	17.87	23.08	33.11	38.16	21.18	7.32	7.39	84.07	32.41
c = 5, e = 10	41.83	53.62	31.53	33.16	42.56	28.39	17.98	40.22	87.18	45.15
c = 5, e = 5	40.72	54.28	33.25	32.62	41.03	25.24	18.18	33.86	87.27	44.30
c = 10, e = 5	39.07	52.83	35.05	32.62	41.03	24.36	13.95	29.86	87.27	42.68
c = 10, e = 10	41.38	49.19	35.82	32.73	38.80	25.00	21.98	40.80	86.74	45.33
c = 5, e = 20	42.72	54.11	35.68	33.81	41.57	27.78	17.78	43.43	87.60	46.08

TABLE II

RESULTS OF EMOTION RECOGNITION IN CONVERSATIONS WITH CHATGPT. c IS THE NUMBER OF INTRODUCED CONTEXT TURNS AND e IS THE NUMBER OF SIMILAR EXAMPLES TO THE UTTERANCE TO CLASSIFY USED IN THE PROMPT.

Emotion: [Output]."

Here you have some examples similar to the utterance to classify: [Examples].

If you do not identify the emotion from the emotions list or the message is empty, please answer neutral.

Context: [Previous Dialogue]

(Client/Agent): [Utterance to classify]

Emotion:

where the blue text represents the input variables that vary depending on the specific utterance under analysis. [Previous Dialogue] refers to the preceding context of the utterance being classified, with each utterance identified by its speaker. [Utterance to classify] represents the specific utterance being analyzed, and [Emotions List] denotes the list of emotions from which the model must make a selection.

This prompt is similar to the one used in [30], in which they use ChatGPT to do ERC in a mental health dataset.

The results are presented in Table II, where is observed that adding both context and retrieved examples improves the performance of ChatGPT in emotion classification in the MAIA dataset, emphasizing that the importance of the retrieved examples far outweighs the significance of the added context. Although both context and retrieved examples improve results using ChatGPT, the performance still comes quite short when compared to the performance achieved when using the RoBERTa-based model.

The results are aligned with the ones presented in [31]. The RoBERTa-based model is trained and tested on the same dataset, allowing it to adapt to dataset-specific labelling guidelines. In contrast, ChatGPT predicts emotions based on its accumulated knowledge not being fine-tuned for this specific domain.

V. EMPATHETIC REPHRASING

As concluded in [32], appropriate empathetic responses can be relevant for the agent to efficiently solve the problem affecting the client.

In figure 3, it is evident that there are few situations where the agent demonstrates clear empathy and, in some minor cases, even exhibits negative emotions, potentially hindering the resolution of clients' problems. As can be seen in table III, in a conversation where the client is already frustrated, the agent is not comprehensive leading to the client continuing frustrated. An empathetic conversation can be seen in table IV.

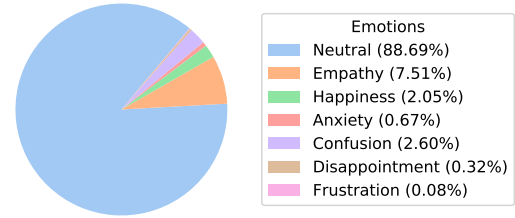


Fig. 3. Emotion distribution of the MAIA dataset for the agent

Client: I carry until supposedly say that it is turned off and nothing happens. I say again that it only enters sleep mode.

Emotion: Frustration

Agent: hold the power button

Emotion: Neutral

Client: How long?

Emotion: Neutral

Agent: until it let you know is powering off

Emotion: Neutral

Client: But I've told you that it doesn't happen. It is already the seventh time I do this and only goes into suspension.

Emotion: Frustration

Agent: perform a manual reset then

Emotion: Neutral

Client: How?

Emotion: Confusion

TABLE III

EXAMPLE OF A NON-EMPATHETIC DIALOGUE. THE CLIENT'S SOURCE LANGUAGE IS BRAZILIAN PORTUGUESE.

To solve this issue, we use ChatGPT and Vicuna [33] to rephrase agents' utterances that are not considered empathetic.

It is important to mention that we have made a subtle tweak to the MAIA dataset. This dataset contains dialogues with real-world interactions between agents and clients in continuous conversations. Occasionally, a participant provides multiple consecutive responses before receiving a reply. To maintain coherence, we group consecutive responses from the same

Client: I wanted to update this #PRS_ORG# that with bought the 3 years and has not been updated yet any is this #PRS_ORG#

Emotion: Frustration

Client: Nr series is this #ALPHANUMERIC_ID#

Emotion: Neutral

Agent: Hello, my name is #NAME#. Thank you for contacting #PRS_ORG# Customer Support.

Emotion: Neutral

Client: For the purpose of this

Emotion: Neutral

Agent: I am sorry for the issue you are facing while updating the device.

Emotion: Empathy

Client: and the Portuguese support

Emotion: Neutral

Agent: I will surely assist you to update the device. May I know the issue you are facing while updating the device?

Emotion: Neutral

TABLE IV

EXAMPLE OF AN EMPATHETIC DIALOGUE. THE CLIENT’S SOURCE LANGUAGE IS PORTUGUESE.

participant, helping to maintain natural conversation dynamics, which better fit the generation abilities of LLMs. Nonetheless, we had to select a single label that best characterizes the entire set of utterances. In this context, we opt to employ the emotion observed in the final non-neutral utterance of that source. This led to a reduction in the dataset to less than half its size.

We make use of our ERC model to determine which agents’ utterances lack empathy and need a rephrasing to effectively solve the client’s problem. We consider that only sentences that are labelled as empathetic or happy are empathetic enough so we rephrase all other agents’ utterances.

Presented below is the prompt used for obtaining the empathetically rephrased utterances:

- *Help a customer service agent to answer empathetically when they are unable to due to their emotions. This is the format of the interaction:*

"Context: [Previous Dialogue]

Agent: [Message to rephrase to be more empathetic]".

You must rewrite the response while keeping all the original content but making the tone more empathetic.

It is important to note that the definition of empathy is: Empathy is an emotion related to one’s cognitive or emotional ability to share the emotions felt by another upon observance. This sense of awareness of others’ emotional state increases the likelihood of more successful communicative interactions. Empathy often leads to prosocial actions, such as offering help, support, or comfort to others, WITHOUT UNDERESTIMATING THE OTHER. In a customer support interaction, empathy promotes a more seamless experience with friendlier customers and more sympathetic agents.

Here you have some examples of empathetic responses: [Empathetic Responses].

IMPORTANT: The answer must start with: "Agent: Answer that should be given".

Context: [Previous Dialogue]

Agent: [Message to empathetically rephrase]

where the blue text represents the input variables that vary depending on the utterance to rephrase. [Empathetic Responses] presents examples of agents’ empathetic responses, [Previous Dialogue] refers to the preceding context of the utterance being classified, with each utterance identified by its speaker. [Message to empathetically rephrase] represents the utterance to be rephrased.

VI. EVALUATION

The rephrased utterances are then evaluated, both automatically and manually. The automatic evaluation assesses whether more agents’ utterances are identified as empathetic by the ERC model, whereas the manual evaluation is important to understand if the automatic evaluation is validated.

A. Automatic Evaluation

Firstly, our ERC model automatically evaluates the rephrased dataset. In this manner, we employ our model to perform the ERC task and examine whether there is an increase in the number of empathetic utterances. Notably, as can be seen in figures 4, 5 and 6 the model assigns a higher number of utterances as empathetic in the original text compared to both the text rephrased by Vicuna and ChatGPT. However, this outcome may find its explanation in our prior practice of grouping utterances. Labelling utterances based on the original dataset may compromise the model’s ability to accurately identify long empathetic utterances due to the lack of sufficient examples of such utterances in the dataset.

Actual Class	Emp	127 77%	17 24%	110 7%	3 6%	0 0%	0 0%	0 0%	5 4%
	Hap	4 2%	21 30%	42 3%	0 0%	0 0%	0 0%	0 0%	2 1%
	Neu	22 13%	22 31%	1129 75%	9 19%	30 25%	4 16%	0 0%	13 10%
	Disa	1 1%	0 0%	39 3%	17 35%	6 5%	1 4%	0 0%	10 7%
	Conf	3 2%	4 6%	68 5%	0 0%	48 40%	2 8%	0 0%	17 13%
	Frus	1 1%	1 1%	23 2%	8 17%	9 7%	10 40%	0 0%	10 7%
	Ang	1 1%	0 0%	12 1%	1 2%	5 4%	6 24%	0 0%	6 4%
	Anx	5 3%	5 7%	76 5%	10 21%	22 18%	2 8%	1 100%	71 53%
			Emp	Hap	Neu	Disa	Conf	Frus	Ang
		Predicted Class							

Fig. 4. Confusion Matrix before rephrasing

B. Manual Evaluation

Regarding the manual evaluation, we evaluated the models using 3 different criteria: Empathy, Relevance, and Fluency. Empathy measures the understanding and expression degree of context emotion the response demonstrates, Relevance evaluates

Actual Class	Emp	49 43%	12 22%	163 10%	8 17%	12 11%	5 15%	0 0%	13 12%
	Hap	1 1%	12 22%	53 3%	0 0%	2 2%	1 3%	0 0%	0 0%
	Neu	49 43%	26 48%	929 58%	30 65%	78 71%	23 68%	2 100%	92 84%
	Disa	2 2%	0 0%	69 4%	3 7%	1 1%	0 0%	0 0%	0 0%
	Conf	5 4%	1 2%	116 7%	3 7%	11 10%	4 12%	0 0%	2 2%
	Frus	3 3%	2 4%	51 3%	0 0%	4 4%	1 3%	0 0%	0 0%
	Ang	1 1%	0 0%	30 2%	0 0%	0 0%	0 0%	0 0%	0 0%
	Anx	4 4%	1 2%	181 11%	2 4%	2 2%	0 0%	0 0%	2 2%
		Emp	Hap	Neu	Disa	Conf	Frus	Ang	Anx
		Predicted Class							

Fig. 5. Confusion Matrix after rephrasing using ChatGPT

Actual Class	Emp	62 40%	7 13%	156 10%	9 22%	12 11%	2 7%	1 33%	13 13%
	Hap	1 1%	12 22%	54 3%	0 0%	1 1%	1 3%	0 0%	0 0%
	Neu	62 40%	33 61%	927 59%	24 60%	78 72%	22 73%	2 67%	81 83%
	Disa	4 3%	0 0%	68 4%	1 2%	1 1%	1 3%	0 0%	0 0%
	Conf	9 6%	0 0%	114 7%	3 7%	11 10%	3 10%	0 0%	2 2%
	Frus	1 1%	0 0%	56 4%	2 5%	2 2%	0 0%	0 0%	0 0%
	Ang	3 2%	0 0%	27 2%	0 0%	0 0%	1 3%	0 0%	0 0%
	Anx	13 8%	2 4%	171 11%	1 2%	3 3%	0 0%	0 0%	2 2%
		Emp	Hap	Neu	Disa	Conf	Frus	Ang	Anx
		Predicted Class							

Fig. 6. Confusion Matrix after rephrasing using Vicuna

whether the responses are on-topic with the dialogue history and Fluency measures the grammatical correctness and readability of the response.

We used 4 non-expert human annotators to evaluate our framework. To measure inter-annotator agreement, we used Fleiss' Kappa inter-annotator agreement and obtained a value of 0.43, which can be classified as a moderate agreement.

The results of the manual evaluation are presented in tables V, VI and VII.

As can be seen in table V, both ChatGPT and Vicuna improve in relevance compared to the original text, being ChatGPT a stronger improvement than Vicuna. Nonetheless, the

Relevance	Win	Loss	Tie
ChatGPT vs Original	47%	26%	27%
Vicuna vs Original	31%	45%	24%
ChatGPT vs Vicuna	49%	18%	33%

TABLE V

RESULTS OF HUMAN EVALUATION FOR UTTERANCES' ADEQUACY AND EMPATHY.

Fluency	Win	Loss	Tie
ChatGPT vs Original	83%	8%	9%
Vicuna vs Original	62%	24%	14%
ChatGPT vs Vicuna	71%	10%	19%

TABLE VI

RESULTS OF HUMAN EVALUATION FOR UTTERANCES' ADEQUACY AND EMPATHY.

Empathy	Win	Loss	Tie
ChatGPT vs Original	86%	7%	7%
Vicuna vs Original	50%	38%	12%
ChatGPT vs Vicuna	75%	11%	14%

TABLE VII

RESULTS OF HUMAN EVALUATION FOR UTTERANCES' ADEQUACY AND EMPATHY.

improvement is not substantial and certain answers with lower relevance may sacrifice valuable information from the original text, making it challenging to definitively consider the models as better relevance-wise. On the other side, tables VI and VII clearly show a significant improvement from both Vicuna and ChatGPT in terms of fluency and empathy when compared to the original text, with ChatGPT particularly standing out.

Nevertheless, it is important to do an extensive analysis of examples to understand the critical failures of both ChatGPT and Vicuna.

Starting with the Vicuna model, the most prominent aspect is its tendency to maintain the original content on numerous occasions, except by correcting punctuation and misspellings. An illustrative instance can be seen in Table VIII:

Context:

Client :#PRS_ORG#

Agent: Thank you so much.We are going to work on your device, can you help me, please?

Client: Sim se poder

Original

Agent: Thank yu very much.Let's open the #PRS_ORG# app, and tap on the "Menu" icon at the top of the screen, please. Then tap "Settings", and scroll to the bottom and tap Repair your account.Finally tap "Repair", please.

Vicuna Rephrasing

Agent: Thank you very much.Let's open the #PRS_ORG# app, and tap on the "Menu" icon at the top of the screen, please. Then tap "Settings", and scroll to the bottom and tap Repair your account.Finally tap "Repair", please.

TABLE VIII

EXAMPLE OF A REPHRASING IN THAT VICUNA ONLY CORRECTS THE ORIGINAL RESPONSE. THE CLIENT'S SOURCE LANGUAGE IS PORTUGUESE.

Additionally, at times, Vicuna may exhibit a lack of comprehension regarding the task at hand and for example, translate the utterance of the client or answers to the utterance to rephrase.

Regarding ChatGPT, it exhibits a distinctive response style that may lead to client frustration, characterized by overly repetitive interactions non-natural in a conversation, which is

reflected also in the repetition of previous context leading to long sentences. One example can be seen in table IX:

Client: No momento não sei quando os comprei; mais ou menos de 3 a 5 anos

Agent: I understand that you're experiencing an issue with the #PRS_ORG# desktop app, where it is showing that your library is empty even though you have purchased ebooks. I apologize for any inconvenience this may have caused. Could you please provide me with your email address so that I can assist you in recovering your ebooks?

Client: Sim

Agent: Thank you for reaching out. I understand that you're unsure about the purchase date of your books, approximately 3 to 5 years ago. I'm here to assist you with any concerns you may have. So, if I understand correctly, you're experiencing missing books on the #PRS_ORG# App, is that correct?

Client: Aguardo.Obrigado.Boa tarde.Não mensagem diz que está vazia.

Agent: I understand. I apologize for any inconvenience. May I kindly ask if you have downloaded the App from our official website?

Client: Baixei novamente, mas continua com a mesma mensagem.

Agent: I understand. Let's try to troubleshoot this issue together, shall we?

TABLE IX

EXAMPLE OF A REPHRASING IN THAT CHATGPT REPEATS THE STRUCTURE FOR ALL ANSWERS. THE CLIENT'S SOURCE LANGUAGE IS PORTUGUESE.

On the other side, ChatGPT approaches the clients carefully, effectively preventing those with negative emotions from further escalating their negativity. A notable example is in Table X

Context:

Client : Vejo Tem ima conta?Entrar e acima tem o botão de ajuda

Agent: Please click on Login, and select #PRS_ORG# as login option. How is it going?

Client: Não encontro o link

Original

Agent: Would be possible to see a screenshot ?

ChatGPT Rephrasing

Agent: Could you please provide a screenshot?

TABLE X

EXAMPLE OF A REPHRASING THAT ENHANCES THE CAREFULNESS USED BY CHATGPT. THE CLIENT'S SOURCE LANGUAGE IS PORTUGUESE.

A clear illustration of the enhancements achieved by both Vicuna and ChatGPT can be found in Table XI. In comparison to the original response, both models significantly improve the structure and quality of the generated response.

In summary, we can see the potential of LLMs to produce high-quality text but we have to take into account that it has its failures and cannot be trusted without reviewing the text

Context:

Client : continua sem abrir mas vou mandar vou mandar a informação que me aparece

Agent: Ok, please send it

Client: Visitor uploaded: informacao-#PRS_ORG#.jpg
URL: #URL#Type: image/jpegSize: #PHONENUMBER#

Original

Agent: Ok, please do a sign out, with this steps. This will help you to update your account Go to your Home screen.Tap the Menu icon at the top of the screen.Tap Settings.Tap Accounts.Under #PRS_ORG#, tap Sign out.A confirmation screen will appear.Tap Sign out.

Vicuna Rephrasing

Agent: Once you have completed these steps, please let me know and we will proceed further to resolve the issue. B. Please perform a sign-out process to update your account. To do this, follow these steps:

1. Go to your Home screen.
2. Tap the Menu icon at the top of the screen.
3. Tap Settings.
4. Tap Accounts.
5. Under #PRS_ORG#, tap Sign out.
6. A confirmation screen will appear.
7. Tap Sign out.

ChatGPT Rephrasing

Agent: Ok, I understand that you are still unable to open the information. Thank you for sharing the details with me. In order to assist you better, could you please follow these steps to sign out of your account? This will help us update your account and resolve the issue.

1. Go to your Home screen.
 2. Tap the Menu icon at the top of the screen.
 3. Tap Settings.
 4. Tap Accounts.
 5. Under #PRS_ORG#, tap Sign out.
 6. A confirmation screen will appear.
 7. Tap Sign out.
-

TABLE XI

EXAMPLE OF A REPHRASING THAT CLEARLY IMPROVES THE STRUCTURE AND QUALITY OF THE RESPONSE FOR BOTH VICUNA AND CHATGPT'S ANSWERS. THE CLIENT'S SOURCE LANGUAGE IS PORTUGUESE.

produced. This leads us to believe that this can be an important tool for empathetic dialogue rephrasing, not to substitute the customer service agent but rather to help him/her be more successful with clients.

VII. CONCLUSIONS AND FUTURE WORK

In this work, we proposed a framework to empower customer support agents, consisting of an ERC module and an empathetic rephrasing system.

Regarding the ERC module, we applied [27] to customer service, and demonstrated that the introduction of an adequate number of context turns, significantly improves performance. Furthermore, its strong performance on empathy identification suggests its suitability as an automatic empathy classifier. The

comparison with ChatGPT proved the importance of fine-tuning for the ERC task.

As far as the Empathetic Rephrasing system is concerned, we successfully built a system capable of enhancing the empathy of the agents' responses in a customer service environment. Results were validated with manual evaluation

For future validation, it would be beneficial to expand our methods to additional language pairs. Moreover, it would be worth exploring fine-tuning a pre-trained model specifically for the empathetic rephrasing task, as opposed to relying on prompting an LLM. Regarding the ERC module, it could be valuable to use larger models with an increased number of parameters and experiment using datasets with more language pairs.

REFERENCES

- [1] Jinjie Ni, Tom Young, Vlad Pandealea, Fuzhao Xue, and Erik Cambria. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, 56(4):3055–3155, 2023.
- [2] Fabrizio Macagno, Chrysi Rapanta, Elisabeth Mayweg-Paus, and Merce Garcia-Mila. Coding empathy in dialogue. *Journal of Pragmatics*, 192:116–132, 2022.
- [3] Mohammadreza Hojat, Joseph S Gonnella, Thomas J Nasca, Salvatore Mangione, Michael Vergare, and Michael Magee. Physician empathy: definition, components, measurement, and relationship to gender and specialty. *American Journal of Psychiatry*, 159(9):1563–1569, 2002.
- [4] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953, 2019.
- [5] Ana C. Farinha, M. Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José G. C. de Souza, Helena Moniz, and André F. T. Martins. Findings of the WMT 2022 Shared Task on Chat Translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT22)*, Abu Dhabi, December 2022. Association for Computational Linguistics.
- [6] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883, 2017.
- [7] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguerrn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825, 2019.
- [8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [9] Peixiang Zhong, Di Wang, and Chunyan Miao. Knowledge-enriched transformer for emotion detection in textual conversations. *arXiv preprint arXiv:1909.10681*, 2019.
- [10] Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue*, pages 186–195, 2020.
- [11] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.
- [12] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguecgn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*, 2019.
- [13] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. Directed acyclic graph network for conversational emotion recognition. *arXiv preprint arXiv:2105.12907*, 2021.
- [14] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *arXiv preprint arXiv:2304.03347*, 2023.
- [15] Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*, 2023.
- [16] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*, 2018.
- [17] Zhaoliang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*, 2019.
- [18] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. Mime: Mimicking emotions for empathetic response generation. *arXiv preprint arXiv:2010.01454*, 2020.
- [19] Tulika Saha and Sophia Ananiadou. Emotion-aware and intent-controlled empathetic response generation using hierarchical transformer network. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.
- [20] Yun Su, Haoran Bian, Bozhen Fan, Bingxu Lian, Chengrong Zhang, Bingtao Zhang, and Runhe Huang. Rlca: Reinforcement learning model integrating cognition and affection for empathetic response generation. *IEEE Transactions on Computational Social Systems*, 2023.
- [21] Ye Liu, Wolfgang Maier, Wolfgang Minker, and Stefan Ultes. Empathetic dialogue generation with pre-trained roberta-gpt2 and external knowledge. In *Conversational AI for Natural Human-Centric Interaction: 12th International Workshop on Spoken Dialogue System Technology, IWSDS 2021, Singapore*, pages 67–81. Springer, 2022.
- [22] Yushan Qian, Bo Wang, Shangzhao Ma, Wu Bin, Shuo Zhang, Dongming Zhao, Kun Huang, and Yuexian Hou. Think twice: A human-like two-stage conversational agent for emotional response generation. *arXiv preprint arXiv:2301.04907*, 2023.
- [23] Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10993–11001, 2022.
- [24] Yushan Qian, Bo Wang, Ting-En Lin, Yinhe Zheng, Ying Zhu, Dongming Zhao, Yuexian Hou, Yuchuan Wu, and Yongbin Li. Empathetic response generation via emotion cause transition graph. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [25] Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. Wassa 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 214–227, 2022.
- [26] Bushra Amjad, Muhammad Zeeshan, and Mirza Omer Beg. Emp-eval: A framework for measuring empathy in open domain dialogues, 2023.
- [27] Isabel Dias, Ricardo Rei, Patrícia Pereira, and Luisa Coheur. Towards a sentiment-aware conversational agent. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–3, 2022.
- [28] Konstantinos I. Roumeliotis and Nikolaos D. Tselikas. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192, May 2023.
- [29] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852, 2020.
- [30] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *arXiv preprint arXiv:2304.03347*, 2023.
- [31] Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*, 2023.
- [32] Colin Mackinnon Clark, Ulrike Marianne Murfett, Priscilla S Rogers, and Soon Ang. Is empathy effective for customer service? evidence from call center interactions. *Journal of Business and Technical Communication*, 27(2):123–153, 2013.
- [33] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.