



TÉCNICO
LISBOA

Answer Rephrasing and Emotion Tracking in a Multilingual Customer Service Chat

Luís Miguel Correia Alves

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisors: Prof. Dr. André Filipe Torres Martins
Dra. Ana Catarina Farinha

Examination Committee

Chairperson: Prof. Dr. João Manuel de Freitas Xavier
Supervisor: Prof. Dr. André Filipe Torres Martins
Member of the Committee: Prof. Dra. Maria Luísa Torres Ribeiro Marques da Silva
Coheur

November 2023

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

First and foremost, I would like to express my deepest appreciation to my supervisors, Prof. André Martins and Dra. Catarina Farinha, for all the support, guidance and inspiration, which were crucial for the development of this project.

Secondly, I would like to thank the Unbabel research team for all the shared knowledge essential to guide my thesis on the right track.

Thirdly, and most importantly, I would like to thank my family, especially my mother, my father and my brother for always giving me the best support they could offer during this particular period. They always pushed me to become a better version of myself and, without them, this work would not exist.

A special thanks to Catarina for always motivating me and being by my side this whole journey.

Finally, I would like to thank my friends. It would be impossible to finish this journey without all the support given by them.

Abstract

Providing exceptional customer service is crucial to promoting customer loyalty and ensuring a fulfilling client experience. Unfortunately, the client experience is often harmed when the agent's response is unhelpful or not empathetic. In this thesis, we aim to improve the quality of customer support by leveraging recent progress in language modelling. We focus on two critical tasks: emotion recognition in conversations and empathetic dialogue rephrasing.

In the context of emotion recognition, we adapt existing methods to the domain of customer service, an area that has received relatively little attention in the literature, and compare them with a prompting-based approach using ChatGPT. Our findings highlight the importance of fine-tuning models for specific tasks, as a smaller, task-specific model outperforms the more powerful ChatGPT in emotion recognition. Moreover, we observe the importance of providing the model with the conversational context to enhance performance.

In addition, we develop an empathetic rephrasing system that leverages the emotion recognition model to identify sentences in need of rephrasing to better address clients, building upon the emotion recognition model's strong capacity in empathy identification. These identified sentences are then fed to ChatGPT to enhance the quality and empathy of responses.

Our results demonstrate the significance of fine-tuning for emotion recognition tasks, as well as the efficacy of ChatGPT in producing high-quality, empathetic text for empathetic rephrasing. These findings underscore the potential for substantial improvements in customer service when implementing such systems compared to regular customer service agents.

Keywords

Emotion Recognition; Empathetic Rephrasing; Prompting.

Resumo

Fornecer um atendimento ao cliente de excelência é fundamental para promover a fidelização dos clientes e garantir-lhes uma experiência satisfatória. Infelizmente, a sua experiência é frequentemente prejudicada pela falta de empatia ou conteúdo das respostas do agente. Nesta tese, temos como objetivo melhorar a qualidade do suporte ao cliente, aproveitando os recentes avanços na modelação de linguagem. Focamo-nos em duas tarefas de grande importância: reconhecimento de emoções em conversas e reformulação empática de diálogos.

No contexto do reconhecimento de emoções, adaptamos métodos existentes ao atendimento ao cliente. Comparamos também esses métodos com uma abordagem baseada em prompting usando o ChatGPT. As nossas experiências destacam a importância de fazer fine tuning de modelos para esta tarefa, pois um modelo menor e específico para a tarefa supera o ChatGPT, mais poderoso, no reconhecimento de emoções. Adicionalmente, observamos a importância de fornecer ao modelo o contexto da conversa para aprimorar o desempenho.

Adicionalmente, desenvolvemos um sistema de reformulação empática baseado no modelo de reconhecimento de emoções para identificar frases que precisam de reformulação para um melhor atendimento ao cliente. Isto é possível devido à forte capacidade do modelo de reconhecimento de emoções na identificação de empatia. As frases identificadas como pouco empáticas são dadas ao ChatGPT para melhorar a qualidade e empatia das respostas.

Palavras Chave

Reconhecimento de Emoções; Reformulação Empática; Prompting;

Contents

1	Introduction	2
1.1	Motivation	3
1.2	Contributions	5
1.3	Outline	5
2	Background	7
2.1	Feedforward Neural Networks	8
2.2	Recurrent Neural Networks (RNNs)	8
2.3	Sequence-to-Sequence Models	11
2.4	Attention Mechanisms	11
2.5	Transformer	12
2.6	Transfer Learning	15
2.7	Word embeddings	15
2.8	Prompting	17
3	Emotion Recognition in NLP	18
3.1	Emotion Theories	19
3.2	Dialogue Emotion Labelled Datasets	19
3.3	Evaluation Metrics	22
3.4	Models for Emotion Recognition in Conversation	23
3.5	Text Generation Conditioned to Emotions	27
3.6	Empathetic Dialogue Generation	29
4	Emotion Recognition in Conversation	35
4.1	Datasets	36
4.2	Metrics	37
4.3	Models	38
4.3.1	RoBERTa-based model	39
4.3.2	ChatGPT API	45
4.4	Experiments using empathy label	47

4.4.1	RoBERTa-based model	48
4.4.2	ChatGPT API	48
5	Empathetic Dialogue Rephrasing	51
5.1	Dataset	52
5.2	Multilingual AI Agent Assistant (MAIA) empathy analysis	52
5.3	Methodology	53
5.4	Evaluation	55
5.4.1	Automatic Evaluation	55
5.4.2	Manual Evaluation	56
5.5	Consecutive response grouping approach	59
5.6	Models	60
5.7	Evaluation	61
5.7.1	Automatic Evaluation	61
5.7.2	Human Evaluation	62
6	Conclusion	69
6.1	Conclusions and Future Work	70
	Bibliography	71

List of Figures

1.1	The importance of emotion recognition in a conversation.	3
2.1	Feedforward neural network [Beqari, 2018]	9
2.2	Unfolded Recurrent Neural Network [Feng et al., 2017]	9
2.3	Long Short Term Memory (LSTM) architecture [dprogrammer, 2020]	10
2.4	GRU architecture [Pipis, 2020]	11
2.5	Attention mechanism function [Bahdanau et al., 2014]	12
2.6	Transformer architecture [Vaswani et al., 2017]	13
2.7	Multi-head self-attention layer [Vaswani et al., 2017]	14
3.1	ERC approaches	24
3.2	Emotional generation models diagram	27
3.3	Empathetic Generation Models	29
5.1	Emotion distribution of the MAIA dataset for the agent.	52
5.2	Predicted emotion distribution of the MAIA dataset for the agent after rephrasing.	55
5.3	Confusion Matrix before rephrasing	56
5.4	Confusion Matrix after rephrasing	56
5.5	Confusion Matrix before rephrasing	61
5.6	Confusion Matrix after rephrasing using ChatGPT	62
5.7	Confusion Matrix after rephrasing using Vicuna	62

List of Tables

3.1	Emotion distribution per language pair	22
3.2	Example of the different emotions of the MAIA dataset.	22
4.1	Correspondence between DailyDialog and MAIA emotions.	37
4.2	Distribution of utterances for the MAIA training, validation and test sets and DailyDialog utterances on train set.	38
4.3	Average results from three runs using a Robustly Optimized BERT Approach (RoBERTa)-based model for emotion recognition in conversations using different training strategies.	40
4.4	Average results from three runs using a RoBERTa-based model for emotion recognition in conversations for German-English language pair.	42
4.5	Average results from three runs using a RoBERTa-based model for emotion recognition in conversations for Portuguese-English language pair.	43
4.6	Study of the impact of context in Emotion Recognition in Conversations with RoBERTa-based model for the non-empathy labelled dataset.	44
4.7	Results of Emotion Recognition in Conversations with ChatGPT for the non-empathy labelled dataset.	47
4.8	Average results from three runs using a RoBERTa-based model for emotion recognition in conversations with the dataset labelled with empathy.	49
4.9	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.	50
5.1	Example of a non-empathetic dialogue. The client's source language is Brazilian Portuguese.	53
5.2	Example of an empathetic dialogue. The client's source language is Brazilian Portuguese.	54
5.3	Results of human evaluation for utterances' adequacy and empathy.	57
5.4	Results of human evaluation for meaning preservation.	57

5.5	Example demonstrating the adequacy and empathy achieved through rephrasing using ChatGPT. The client's source language is Portuguese.	57
5.6	Example of a polite rephrasing using ChatGPT. The client's source language is Brazilian Portuguese.	58
5.7	Example of a rephrasing that changes the meaning of the original utterance.	58
5.8	Example of a rephrasing that repeats the context. The client's source language is Brazilian Portuguese.	59
5.9	Emotion distribution per language pair.	60
5.10	Example of an utterance labelled with frustration where it is not visible in the utterance.	63
5.11	Results of human evaluation for utterances' relevance.	64
5.12	Results of human evaluation for utterances' fluency.	64
5.13	Results of human evaluation for utterances' empathy.	65
5.14	Example of a rephrasing in that Vicuna only corrects the original response.	65
5.15	Example of a rephrasing that translates the client utterance.	66
5.16	Example of a rephrasing that answers to the agent's original response.	66
5.17	Example of a rephrasing that repeats the context.	66
5.18	Example of a rephrasing in that ChatGPT repeats the structure for all answers.	67
5.19	Example of a rephrasing that enhances the carefulness used by ChatGPT.	67
5.20	Example of a rephrasing that clearly improves the structure and quality of the response for both Vicuna and ChatGPT's answers.	68

Acronyms

LLM	Large Language Model
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory
GRU	Gated Recurrent Unit
NLP	Natural Language Processing
ERC	Emotion Recognition in Conversation
seq2seq	Sequence-to-Sequence
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Approach
GPT	Generative Pre-training Transformer
CNN	Convolutional Neural Network
CRF	Conditional Random Field
T5	Text-to-Text Transfer Transformer
RL	Reinforcement Learning
MAIA	Multilingual AI Agent Assistant
LaBSE	Language-agnostic BERT Sentence Embedding

1

Introduction

Contents

1.1 Motivation	3
1.2 Contributions	5
1.3 Outline	5

1.1 Motivation

Providing good customer service is crucial for companies to retain clients since it leads to customer satisfaction [Hanif et al., 2010], a key factor for clients to remain loyal to a company [Jones et al., 1995]. Nonetheless, being able to provide appropriate customer service that satisfies the client's needs is not an easy task, presenting diverse challenges. Every client is unique and has their own specific needs and expectations. It is therefore critical for customer service providers to understand and address these differences in order to provide effective support and maintain customer satisfaction. This implies the use of qualified personnel who are able to handle a wide range of situations and provide personalized support to clients, turning it costly for a company [Bitner et al., 2002]. Moreover, in the globalized world we face today, many times the client and the service provider talk in different languages adding a translation barrier to the conversation, making it even harder to send the right message in the right tone.

Taking the above points into consideration, it is of extreme importance that we are able to evaluate the quality of the customer service provided in order to understand how well it is being performed and in order to improve how it is provided. One of the major indicators of how well-succeeded customer service is being carried out is to evaluate the fluctuation in the clients' emotions during the conversation [Rajaobelina et al., 2021]. So, it is important for the service distributor to understand the emotional state of the client in order to better fulfil the client's needs, as can be seen in Figure 1.1.

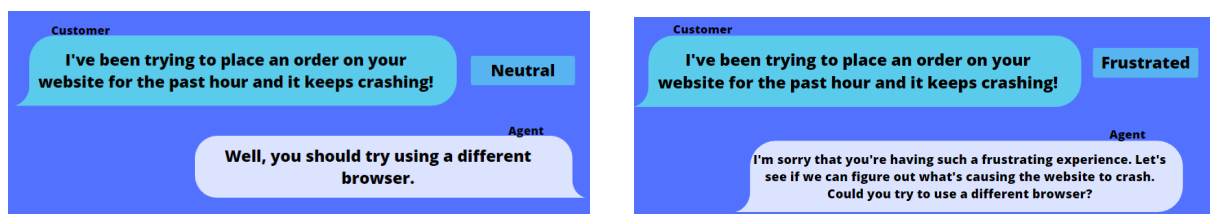


Figure 1.1: The importance of emotion recognition in a conversation. The response of an agent changes based on the emotion of the client's utterance to better address the emotional state of the client.

However, emotions are personal experiences that can be difficult to recognize in a conversation, and identifying the emotions being expressed can be time-consuming and costly [Xu et al., 2022]. It can be particularly challenging to distinguish between similar emotions and to recognize emotions in less structured and more casual conversations, such as those that occur in customer service contexts [Gonçalves et al., 2022]. To address these challenges, researchers have sought to automate the process of emotion recognition in order to avoid relying on manual identification. This field is known as Emotion Recognition in Conversation (ERC) and it has been seeing improvements throughout the last few years, with models largely improving F1-scores across different datasets [Pereira et al., 2022].

There have been different approaches in order to do ERC. The first ones were keyword-based approaches, such as [Ma et al., 2005], where the emotion is recognized by assigning it based on certain

keywords that appear in a given text and have a major influence on the emotion displayed. Moreover, machine learning approaches also came to try to solve the task of emotion recognition, starting with some simple approaches, such as [Alm et al., 2005], which uses a sparse network of linear functions. Additionally, rule-based approaches were also applied, such as [Lee et al., 2010], where the classification of the emotion in a text is given by a set of rules.

However, since deep learning has experienced major growth over the last decade, driven primarily by the great increase in computational power and available data, deep learning approaches appeared to address ERC. In a flash, deep learning managed to overcome the previous approaches achieving state-of-the-art results with some powerful models such as recurrent neural networks [Elman, 1990] and the transformer [Vaswani et al., 2017], suited for tasks that depend on previous inputs.

There are many challenges in order to apply deep learning models and achieve good performance when dealing with emotion detection in dialogues since the emotion does not depend only on the utterance of the speaker at that moment, but also on the other previous utterances from the same speaker and in the previous utterances from the other speaker. Nonetheless, deep learning is suited for this task due to its capacity to handle sequential data, making it effective in capturing context and dependencies in dialogues and also its scalability since as the volume of dialogues and text data grows, deep learning models can continue to improve their performance.

In recent years, there have been numerous research efforts that have utilized deep learning architectures for ERC, such as [Majumder et al., 2019] or [Wang et al., 2020], which will be described in Section 3. However, only the approach of [Wang et al., 2020] applies ERC in a customer service setting and none of the approaches covered tests the architecture in bilingual conversations.

Though modelling the emotions in a customer service conversation may improve its quality, since they relate directly to customer satisfaction [Oliver, 2014], it is not enough to solve all the challenges that customer service presents.

One of the challenges that still remains is the fact that it is a costly process. To address the cost, many companies turned to develop chatbots using deep learning, in an effort to replicate the tasks typically performed by humans. Nevertheless, creating an effective chatbot is a formidable task. Not only must it grapple with the inherent complexity of emulating human interaction, but it also needs to excel within the realm of customer service by accommodating diverse client needs, demonstrating adaptability in challenging scenarios, and possessing extensive domain-specific knowledge to provide valuable assistance to customers. Additionally, a critical aspect lies in how the agent interacts with customers and its ability to tailor support based on the client's circumstances, making customer experience a paramount concern.

Addressing the difficulty of building an effective conversational system, many studies have come out in the last few years in order to build empathetic generation systems for conversations, such as [Lin

et al., 2019] and [Majumder et al., 2020]. This is particularly crucial in the context of customer service, where the agent’s ability to be empathetic with the client is vital, given that empathy can be described as the emotional ability to share the emotions felt by another upon observance. However, no prior research has undertaken this task within a multilingual framework, particularly in the context of customer service, which presents unique challenges as conversations tend to be informal, frequently featuring ungrammatical linguistic structures, and exhibit blurry boundaries between written text and speech, often mimicking speech production [Farinha et al., 2022].

Moreover, evaluating a model’s performance regarding text quality and empathy is a lengthy process, with current models using separate measurements for each aspect. Various studies, including [Amjad et al., 2023], have attempted to construct metrics for assessing empathetic generation. However, the absence of a universally used method to measure empathy poses a significant challenge in the development of these models, since the majority of developed models rely on human evaluation for assessing empathy, such as [Lin et al., 2019] and [Majumder et al., 2020].

1.2 Contributions

In this thesis, there are two main problems that are addressed and closely related to each other.

- In our study, we conduct a comparative analysis by employing a relatively compact model for training an ERC model across three distinct training datasets within a customer service context. Moreover, we analyze the influence of the utterances of context that are used by the model to predict the emotion of the target utterance. Additionally, we assess the performance of a fine-tuned model in contrast to a Large Language Model (LLM) that has not undergone task-specific training but relies solely on prompting. Our investigation delves deep into strategies for enhancing the efficacy of prompting methods. Furthermore, we explore the potential benefits of detecting empathy in conversations, which can enable automated empathy evaluation.
- Furthermore, we propose an empathy rephrasing system that leverages an LLM exclusively through prompting, in conjunction with the previously fine-tuned model for the emotion recognition task, due to the high writing capabilities of LLMs. We compare two different-sized LLMs to understand the effect of model size on the capability of creating fluent and relevant text, that is also empathetic.

1.3 Outline

The remainder of the document is structured as follows:

- Chapter 2 presents a background to the theoretical deep learning concepts useful to understand the work developed in this thesis.
- Chapter 3 presents a deep dive into the state-of-the-art work in both ERC and empathetic dialogue generation tasks, the main tasks studied in the ambit of this thesis.
- Chapter 4 presents the different experimental setups used for the ERC task when fine-tuning a model and also when prompting an LLM together with the results obtained for each approach.
- Chapter 5 presents the experimental results derived from employing the empathetic rephrasing tool, using different LLMs as systems responsible for rephrasing.
- Chapter 6 sums up the main achievements of this thesis and leaves suggestions for future work.

2

Background

Contents

2.1 Feedforward Neural Networks	8
2.2 Recurrent Neural Networks (RNNs)	8
2.3 Sequence-to-Sequence Models	11
2.4 Attention Mechanisms	11
2.5 Transformer	12
2.6 Transfer Learning	15
2.7 Word embeddings	15
2.8 Prompting	17

In this chapter, we will cover several concepts of deep learning that will be essential for understanding the models described in Chapter 3.

First, concepts that came as the foundation of the deep learning field are presented, such as feedforward and recurrent neural networks. Then, we will dive into more advanced models, such as sequence-to-sequence models and transformer models, which are currently state-of-the-art in many applications in Natural Language Processing (NLP), such as machine translation and summarization. Moreover, we present different strategies to obtain word embeddings, which represent words in numerical vectors and are the input of textual deep learning models, since these models can only interpret numerical values. Finally, we do a small introduction to LLM prompting due to its high usability across NLP tasks nowadays.

2.1 Feedforward Neural Networks

Feedforward neural networks are the original type of artificial neural networks and the foundation of the deep learning field. Its purpose is to process input data, to identify patterns and relationships within the data, and to make decisions or predictions based on that analysis.

Feedforward neural networks consist of three main layers: an input layer, one or more hidden layers, and an output layer. The input layer receives the input data, and the hidden and output layers process the data in order to produce the output. The layers are connected by weights and biases, which are adjusted during training to reduce the error between the predicted output and the target output. An optimization algorithm, such as gradient descent, is usually employed to minimise the error. In order for these networks to be able to model any function, whether linear or not, the result produced by each layer is passed through a non-linear function called an activation function. This allows the network to model a broader range of functions and behaviours. Figure 2.1 illustrates how the network works.

2.2 Recurrent Neural Networks (RNNs)

Recurrent Neural Network (RNN)s [Elman, 1990] are an extension of feedforward neural networks that, as the name already suggests, have a recurrence mechanism. This is achieved by feeding the output of the previous step as the input of the current step. In this way, it is possible to say that this kind of network has a “memory” since the current steps’ output depends on all the previous steps. A graphical view of an RNN can be seen in Figure 2.2:

Given the architecture of a RNN it is suitable for dealing with sequential data since it holds information

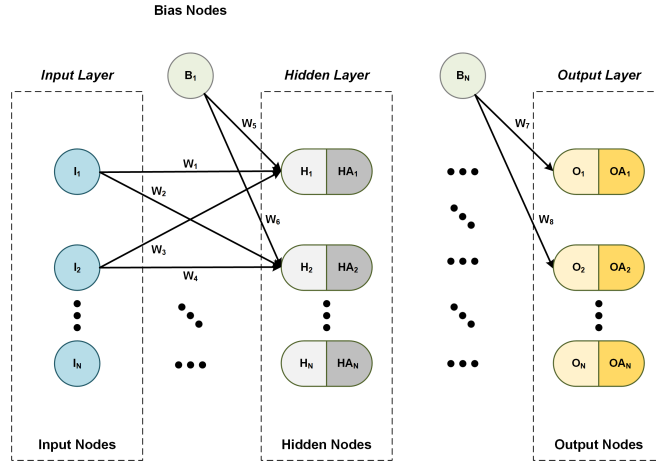


Figure 2.1: Feedforward neural network [Beqari, 2018]

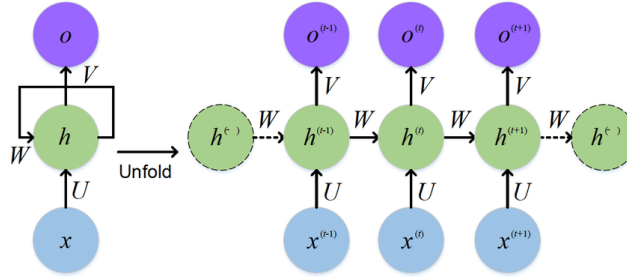


Figure 2.2: Unfolded Recurrent Neural Network [Feng et al., 2017]

about past inputs. The algorithm used in order to train RNNs is backpropagation through time. The output of an RNN, at time step t , is given by:

$$o_t = c + Wh_t, \quad (2.1)$$

where $h_t = f(Ux_t + Vh_{t-1})$ is the hidden state of the cell.

Nonetheless, RNNs have their drawbacks. One of the most significant ones is that this “memory” mechanism is limited due to the vanishing gradient problem [Hochreiter, 1998], which can be reported as the network’s difficulty in recalling a long-term dependency. When applying the backpropagation through time algorithm the gradients become smaller and smaller until they approach values near zero. This means that the network ends up giving more weight to the recent inputs and almost no weight to inputs that are far from the time step in question. Long Short Term Memory (LSTM), introduced in [Hochreiter and Schmidhuber, 1997], and Gated Recurrent Unit (GRU), introduced in [Cho et al., 2014a], were proposed in order to deal with this drawback.

Long Short Term Memory (LSTM)

LSTMs [Hochreiter and Schmidhuber, 1997] are mainly focused on solving the vanishing gradient problem that RNNs have. The key aspects of LSTM are that besides having a hidden state h_t it also has a cell state C_t and three gates that are introduced by this architecture. These gates have the ability to regulate which information should be kept and which one can be discarded from the previous hidden state h_{t-1} and the current input x_t . The three gates are called forget gate, input gate and output gate.

The forget gate is responsible for defining which information should remain and which should be discarded from the previous cell state and the input gate is responsible for deciding which information will be stored in the cell state. These gates are useful to calculate the cell state, which will then be used by the output gate, being this last one responsible for deciding the current hidden state.

A graphical representation of this architecture can be seen in Figure 2.3, where x_t is the current input, h_{t-1} is the previous hidden state, h_t is the current hidden state, C_{t-1} is the previous cell state, C_t is the current cell state.

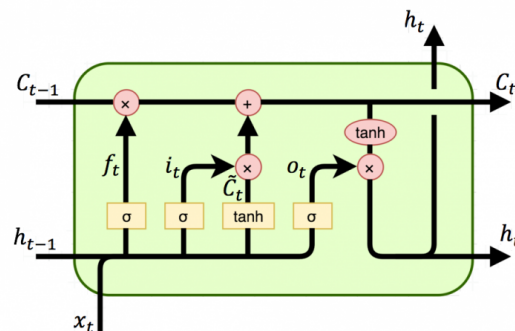


Figure 2.3: LSTM architecture [dprogrammer, 2020]

Gated Recurrent Unit (GRU)

GRU [Cho et al., 2014a] was also proposed in order to solve the problem of long-term dependencies, associated with the vanishing gradient problem. Unlike LSTM, GRU does not have a cell state and only has two gates: the update gate and the reset gate.

The update gate determines the amount of previous information to be retained and passed to the next state, whereas the model's reset gate selects how much past data should be ignored, deciding whether the previous states are important or not.

A graphical view of the model can be seen in Figure 2.4, where x_t is the current input, h_{t-1} is the previous hidden state and h_t is the current hidden state.

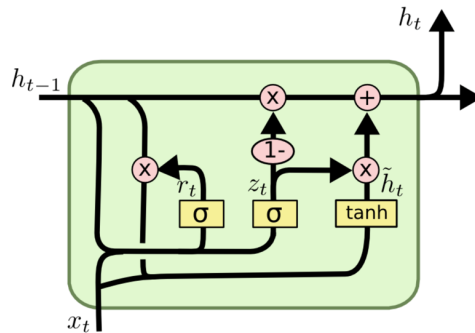


Figure 2.4: GRU architecture [Pipis, 2020]

Comparing GRU to LSTM it can be stated that the first one uses fewer training parameters, uses less memory and is faster to execute, while LSTM is better at handling larger datasets. However, none of these architectures can solve the fact that RNNs are not parallelizable due to their sequential nature. In Section 2.5 we will talk about transformers, which aim to solve this problem.

2.3 Sequence-to-Sequence Models

A Sequence-to-Sequence (seq2seq) model is a model that takes as input a sequence and also outputs a sequence and they were first proposed in [Cho et al., 2014b] and [Sutskever et al., 2014]. These models came to solve the problem that previous models like feed-forward neural networks and RNNs had, which is the fact that they could not map a sequence to a sequence. A seq2seq model is comprised of an encoder and a decoder. The encoder is responsible for trying to capture the information in the input vector in a smaller representation (a fixed-size vector). Then, the decoder is responsible for grabbing that information and creating an output sequence. Since this process is sequential most of the applications used RNNs for the encoder and for the decoder. Nonetheless, encoder-decoder models may encounter difficulties in generating accurate output for large input sequences due to the restricted amount of information contained in the fixed-size output vector produced by the encoder. To solve this problem, attention mechanisms were proposed and are going to be described in Section 2.4.

2.4 Attention Mechanisms

Attention mechanisms were proposed in [Bahdanau et al., 2014] in order to cope with the bottleneck of seq2seq models that, as mentioned in Section 2.3, is the fact that the encoder produces a fixed size

representation independently of the input given. Attention can be seen as an additional component of the model that aims to select which parts of the input are relevant in order to predict the output. So, every time the decoder tries to predict an output, it receives a context vector that helps the decoder to only focus on inputs that are relevant to the output that is being produced and in the previous outputs the decoder has already generated. A graphical view of the mechanism is represented in Figure 2.5.

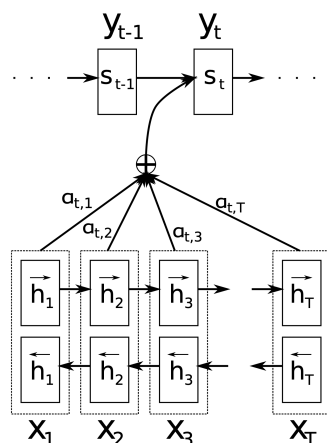


Figure 2.5: Attention mechanism function [Bahdanau et al., 2014]

When using attention in a seq2seq model, the decoder considers all encoder hidden states instead of just the last one. After having all the encoder hidden states, alignment scores are computed by taking the encoder hidden states and the current hidden state of the decoder and processing them through a feed-forward neural network. The network output is passed through a softmax function. Then, a context vector will be calculated by multiplying each encoder's hidden state by its alignment score, which is interpreted as the state's importance for the current output step.

2.5 Transformer

The transformer was proposed in [Vaswani et al., 2017]. Most of the seq2seq models could not be parallelized since both the encoder and the decoder were using RNNs. The transformer proposes the use of attention mechanisms to drop the RNNs from the decoder and the encoder and be able to parallelize their execution, enabling faster training. The transformer is comprised of three main components: the embedding layer, the encoder and the decoder, which are going to be covered below. A view of the entire architecture can be seen in Figure 2.6.

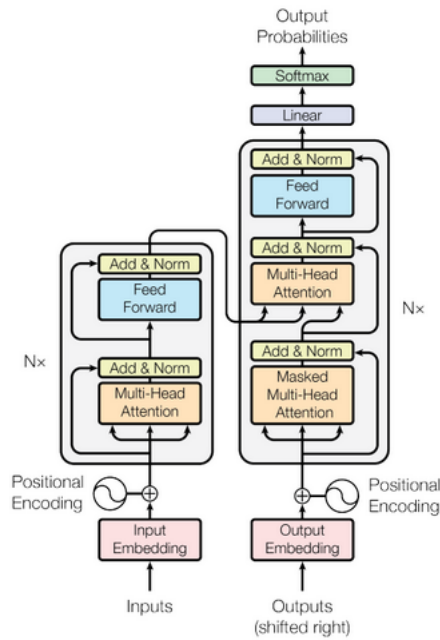


Figure 2.6: Transformer architecture [Vaswani et al., 2017]

Embedding layer

The first step in the process involves converting words into numerical vectors, as the architecture can only interpret numerical values. This process, known as word embedding, can be achieved through the use of a lookup table, which assigns a vector to each word. However, since the transformer does not use RNNs, the architecture lacks the means to capture the position of the words, unless there is also a positional encoding in the embedding layer, represented by a vector. This is done through the use of trigonometric functions, using the cosine function for every odd index of the input vector and the sine function for the even ones. This vector added to the word embedding vector creates a position-based word embedding.

Encoder

As in the other seq2seq models, the encoder is responsible for getting a representation of the input. The encoder proposed in [Vaswani et al., 2017] is composed of 6 stacked encoders. These 6 encoders have the same structure although not sharing the same weights. Each encoder is composed of a multi-head self-attention layer, that will be analyzed in detail, and a position-wise fully connected feed-forward network. After both, a residual connection is employed and is then followed by a layer normalization. As stated in [He et al., 2016], the residual connection enables an easier optimization and increasing accuracy to considerably deep networks. As far as layer normalization [Ba et al., 2016] is concerned, it

helps to stabilize the network enabling faster training.

Multi-head self attention

The transformer model uses a self-attention mechanism, rather than the attention mechanism described in Section 2.4. Whereas the attention mechanism focuses on identifying the words that are relevant for the output at a given time step, the self-attention mechanism in the transformer tries to understand the relationships between input words to fully comprehend the meaning of each word.

To compute self-attention, the first step is to obtain three vectors: query (Q), key (K) and value (V), which are obtained by multiplying the input vectors with matrices trained during the training process, which are represented as W_Q , W_K and W_V . With the three vectors computed, the next step is to multiply the query and key vectors, obtaining a score matrix, where each line of the matrix points out which input vectors are more important to the input vector in question. The values of this matrix are then divided by 8 in order to help stabilize the gradients and passed through a softmax layer to normalize the scores as a probability. Then, this result is multiplied by the value vector in order to obtain the output of the self-attention layer. This process is also entitled scaled dot-product attention and Equation 2.2 resumes the computations described.

$$\text{Self-attention} = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (2.2)$$

Multi-head attention can be described as computing this self-attention mechanism h times, with h being the number of heads. After calculating it h times the results are then concatenated into a single matrix and multiplied by a matrix W_O , which is computed during the training process, achieving the output of the multi-head self-attention layer. Both the computation of self-attention and multi-head attention can be visualized in Figure 2.7.

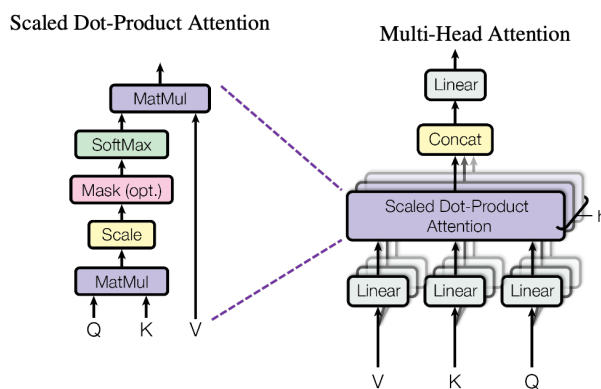


Figure 2.7: Multi-head self-attention layer [Vaswani et al., 2017]

Decoder

Similarly to the other seq2seq models, the purpose of the decoder is to grab the information given by the encoder and generate the output of the model. As well as the encoder, the decoder is composed of 6 stacked decoders and all of them have the same structure not sharing their weights. The structure of the decoder also has the multi-head self-attention layer and a position-wise fully connected feed-forward network, but an additional masked multi-head self-attention layer is needed because the decoder can only consider previously generated words and not future ones. This layer masks future outputs to allow the decoder to operate correctly. Similarly to the encoder, after every layer, a residual connection and a layer normalization are applied.

2.6 Transfer Learning

Deep learning is an area where the amount of computing power we have can significantly impact our ability to achieve groundbreaking results. There are many complex models that require a lot of expensive computing power and take a long time to train. To address this, transfer learning has become popular in the field of deep learning. Transfer learning [Pan and Yang, 2010] involves adapting a previously trained model to a new task in a similar field. This technique can save time and resources as it is more efficient to fine-tune a pre-trained model for a specific task than to train a model from scratch, allowing people with limited resources to compete with those who have greater resources. Some of the most used pre-trained models in NLP tasks are Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2018], a variant of BERT, Robustly Optimized BERT Approach (RoBERTa) [Liu et al., 2019] and Generative Pre-training Transformer (GPT) [Radford et al., 2018].

2.7 Word embeddings

Since the architectures described earlier do not have the ability to work with text, words must be transformed into numerical representations. This is done by converting it to a numeric vector. Word embeddings are learned representations of words where similar words have similar numeric representations.

One way to create word embeddings is to learn them jointly with a neural network model during an NLP task. This involves pre-processing the text and defining a vector space with randomly initialized vectors for each word. These vectors are then trained alongside the model as part of the overall process.

However, training word embeddings from scratch using this approach can be time-consuming and requires a large amount of data in order to be effective. To address these issues, pre-trained word embeddings have become a popular solution. Pre-trained word embeddings are word vectors that have already been trained on a large dataset and can be used as a starting point for a new NLP task. This allows for faster training and can improve performance, especially when working with smaller datasets.

Pre-trained embeddings

Pre-trained embeddings are a cost-effective solution for obtaining high-quality word embeddings. They are trained on a large corpus of text and can be used in place of training embeddings from scratch. This is an example of transfer learning since we use them in a different task than the one they were trained in. Two of the most popularly used pre-trained word embeddings are Word2Vec [[Mikolov et al., 2013](#)] and GloVe [[Pennington et al., 2014](#)].

Word2Vec uses two different learning models in its approach to learning word embeddings. One was the continuous bag-of-words (CBOW) model and the other one was the skip-gram model. While the CBOW model learns the embedding by predicting the current word based on its context, the skip-gram model uses the current word to predict the surrounding words to learn representations.

GloVe is an extension of Word2Vec. Besides using local context to learn a model for word embeddings it also leverages global statistics such as latent semantic analysis. By exploiting both global and local statistics, it is possible to create high-quality word embeddings that can improve performance on certain NLP tasks. Combining these two types of information can lead to embeddings that more accurately reflect the nuances and complexities of the language in analysis.

Contextual embeddings

However, both Word2Vec and GloVe are static word embeddings since they do not change depending on the corpora and the task they are used in. Nowadays, large language models have been utilised to create word embeddings. One of the most popular models used to create word embeddings is BERT. BERT can generate word embeddings using the output of its 12 transformer encoder layers for each token. Empirical evidence suggests that using the sum of the final 4 layers often produces the best results. By using BERT, when compared to the static approaches it is possible to achieve high-quality, context-dependent embeddings, outperforming previous methods.

2.8 Prompting

Prompting has emerged as a pivotal technique in the field of NLP, revolutionizing the way we harness the power of LLMs. It involves providing a specific instruction or query, known as a "prompt," to a LLM, to guide it in generating the required text. Given the instruction, the model processes it and delivers an output using only the previous knowledge acquired in previous training. This facilitates the utilization of LLMs by a broader user base, extending beyond those with higher computational resources, and enabling the customization of powerful models for specific tasks without the need for extensive computational capabilities. Given this, researchers do not rely on changing the training strategies to improve performance but rather on designing the prompt that is best suited for the specific task.

In summary, we conducted a comprehensive review of various models in the field of natural language processing, ranging from simpler models like feedforward neural networks and recurrent neural networks to more advanced models like transformers and LLMs. Additionally, we dived into topics such as word embeddings and prompting, which is an integral component of the emerging NLP paradigm. The insights gained from this review will be instrumental in comprehending the models discussed in the subsequent chapters. h

3

Emotion Recognition in NLP

Contents

3.1 Emotion Theories	19
3.2 Dialogue Emotion Labelled Datasets	19
3.3 Evaluation Metrics	22
3.4 Models for Emotion Recognition in Conversation	23
3.5 Text Generation Conditioned to Emotions	27
3.6 Empathetic Dialogue Generation	29

In this chapter, we will review several state-of-the-art models that are relevant to the work we will be developing.

We will begin by discussing various emotion theories to understand the challenges of detecting emotions in text. Next, we will present an analysis of datasets that provide dialogue conversations with emotional labelling for each utterance. We will then discuss the metrics used to evaluate emotion recognition models and, finally, we will examine various emotion recognition models and models for generating text conditioned to emotion.

3.1 Emotion Theories

The study of emotion involves examining a range of theories, including those that view emotions as fundamental forms of affect and those that consider emotions to be composed of numerical values in various dimensions. While the former approach adopts a categorical view of emotions, the latter approach proposes that emotions are dynamically defined through the interaction of these values. This section will explore these different theories and their implications in understanding emotion.

Categorical Theories

The categorization of emotions into a set of classes allows us to classify an emotion as a single label or multiple labels. Ekman's model [Ekman, 1999], which is considered the benchmark for this emotion description, had only 6 classes: anger, disgust, fear, happiness, sadness, and surprise. However, this approach started to be questioned in whether it could be fully representative of all the emotions that can be expressed [PS and Mahalakshmi, 2017].

Dynamic Theories

In order to deal with such limitations by the categorical theories, some other approaches were made in modelling emotions as a continuous space of values allowing it to have an infinite amount of emotion classifications, such as [Mehrabian and O'Reilly, 1980]. Nonetheless, since there is an infinite amount of emotions it is very hard to do annotation following a dynamic approach.

3.2 Dialogue Emotion Labelled Datasets

This section highlights the various dialogue emotion labelled datasets. There are various types of conversation datasets that can vary in terms of the type of interaction occurring in the conversation, the number of emotions considered, and the emotion theory used for labelling the dataset.

The four most used datasets in ERC are: DailyDialog [Li et al., 2017], EmoryNLP [Zahiri and Choi, 2018], IEMOCAP [Busso et al., 2008] and MELD [Poria et al., 2018], which will be analyzed in more detail. Three other datasets that will be analyzed in detail are AVEC [Schuller et al., 2012], which provides data labelled using a dynamical emotion theory, MSCTD [Liang et al., 2022a], which provides bilingual conversations and Multilingual AI Agent Assistant (MAIA) dataset [Farinha et al., 2022], a dataset with bilingual customer service conversations.

DailyDialog

DailyDialog [Li et al., 2017] is a multi-turn dialogue dataset constructed from raw data from websites where people learn how to talk in English by practising it in dialogues from daily life. This dataset covers categories that range from ordinary life to political topics, being a diversified dataset. It is labelled with the six Ekman's basic emotions and an additional option "other", being a highly unbalanced dataset, with 85572 counts of an utterance label being "other" when categories like fear and disgust only have 74 and 353 counts, respectively.

EmoryNLP

The EmoryNLP dataset [Zahiri and Choi, 2018] consists of 12,606 utterances from the TV show Friends, extracted from transcripts. The dataset is labelled with one of the six primary emotions in Willcox's feeling wheel [Willcox, 1982]: sad, mad, scared, powerful, peaceful, joyful, and neutral, with an unbalanced distribution that reaches a ratio of 1:4 between sad and neutral.

IEMOCAP

The IEMOCAP dataset [Busso et al., 2008] consists of dialogue from 151 videos featuring two speakers, totalling 7,433 utterances with an average length of 11.56 words. It includes labels for eight emotions: happy, sad, neutral, angry, excited, frustrated, fear, and disgust. Despite the availability of all eight emotions, many studies typically only evaluate their performance using the first six classes.

MELD

The MELD dataset [Poria et al., 2018] is a collection of 13,000 utterances extracted from 1,433 dialogues in the TV series Friends. It is annotated with the six basic emotions as identified by Ekman, in addition to a neutral label. The dataset exhibits imbalances, with the neutral category having the highest number of counts (6436) and the fear category having the lowest (358).

AVEC

A different dataset that has also been used is AVEC [Schuller et al., 2012]. Contrarily to the previously described datasets, AVEC is a dataset that is labelled according to a dynamical model of emotions containing interactions between humans and artificially intelligent agents. Each utterance in a conversation is labelled with four different values: valence ($[-1, 1]$), arousal ($[-1, 1]$), expectancy ($[-1, 1]$), and power ($[0, \infty)$).

MSCTD

Despite the fact that the MSCTD dataset [Liang et al., 2022a] labels utterances with sentiment rather than emotion, it is still relevant to our analysis because it is a multilingual dataset with a large number of Chinese-English and German-English bilingual conversation pairs. This is relevant to the problem we are addressing because it includes multilingual conversations.

MAIA

The MAIA dataset [Farinha et al., 2022] is labelled with 8 different emotions: happiness, empathy, disappointment, confusion, frustration, anger, anxiety and neutral, which differ from the six emotions in Ekman's model, but are crucial for the service provider to identify them in order to adjust their posture during the conversation. The MAIA dataset includes conversations in Portuguese-English, Portuguese(Brazil)-English, German-English, and Chinese-English bilingual pairs. In Table 3.1 the distribution of emotions for each language pair is shown.

In Table 3.2, we can observe examples of utterances from the dataset labelled with the different possible emotions of the dataset.

Table 3.1: Emotion distribution per language pair

Emotion	Portuguese-English	Portuguese (Brazil)-English	German-English	Chinese-English
Happiness	13	391	244	35
Empathy	4	206	1508	38
Neutral	674	6548	11057	6067
Disappointment	0	286	384	111
Confusion	24	422	666	456
Frustration	98	180	258	1577
Anger	1	67	198	68
Anxiety	1	53	1677	58
Total utterances	898	8153	15992	8410

Agent: I am happy I was able to help you today.	Emotion: Happiness
Agent: I am sorry that you have not received the order.	Emotion: Empathy
Agent: Hello, my name is #NAME#.	Emotion: Neutral
Client: Having issues on the computer, I'll try on the phone. Sorry the inconvenience.	Emotion: Disappointment
Client: If I switch screens, will the chat close?	Emotion: Confusion
Agent: Messages will not help me to request refund from the specialized department.	Emotion: Frustration
Agent: THE BOOK HAS ONLY 4 CHAPTERS. THEREFORE, IT IS INCOMPLETE.	Emotion: Anger
Agent: I'm at work, I can't stay here for too long.	Emotion: Anxiety

Table 3.2: Example of the different emotions of the MAIA dataset.

As seen in this section, the most used datasets in emotion recognition consider static emotion theories, since using dynamic emotions to label a dataset can be a very costly and time-consuming task.

3.3 Evaluation Metrics

Evaluation metrics that are commonly used in ERC tasks include accuracy, precision, recall, and F1-score. The F1-score is particularly useful in ERC tasks because it combines precision and recall into a single metric, giving equal weight to both. These metrics are commonly used in many classification tasks, including ERC, which is a multi-class classification task, as can be seen by the datasets presented in Section 3.2. In a multi-class classification task, it is important to use both weighted and unweighted metrics in order to get a complete picture of the performance of the classifier.

Weighted metrics

Weighted metrics use micro-averaging, which is to consider the relative frequency of each class, aggregating the contributions of all classes to estimate the metric. This can lead to minority classes having little or no impact on the metric since the major impact is coming from classes with a large number of elements. This indicates these metrics are not good enough to evaluate the performance of the model in minority classes.

Unweighted metrics

In contrast, unweighted metrics do not consider the relative frequency of each class and therefore treat each class as equally valuable. As a result, a model must perform well across all classes in order to achieve a good overall performance using these metrics.

Given this, it is important to compute unweighted metrics when analyzing an imbalanced dataset, where minority classes are much smaller in number compared to the majority classes. As we have observed in the datasets discussed in Section 3.2, most datasets used in emotion recognition classification tasks are highly imbalanced, making it necessary to use these metrics to evaluate the performance of a model accurately.

3.4 Models for Emotion Recognition in Conversation

In this section, the evolution of the deep learning approaches to ERC will be covered. Starting by first talking about simpler approaches and then deepening down into more sophisticated approaches that led to state-of-the-art results. Modelling emotion in a conversation is a difficult task, which can be explained by the fact that the emotion of a particular utterance in a conversation does not only depend on that utterance but also on the context of the speaker's previous utterances and the previous utterances of other speakers in the conversation. The reliance on previous utterances soon gave space for the development of approaches based on RNNs, which can provide information from the past to the current input.

The different approaches used for ERC can be seen in Figure's 3.1 diagram.

Recurrence-Based Methods

Recurrence-based methods are a class of deep learning techniques in which the output of the model depends not only on the current input but also on previous inputs. One example of a recurrence-based

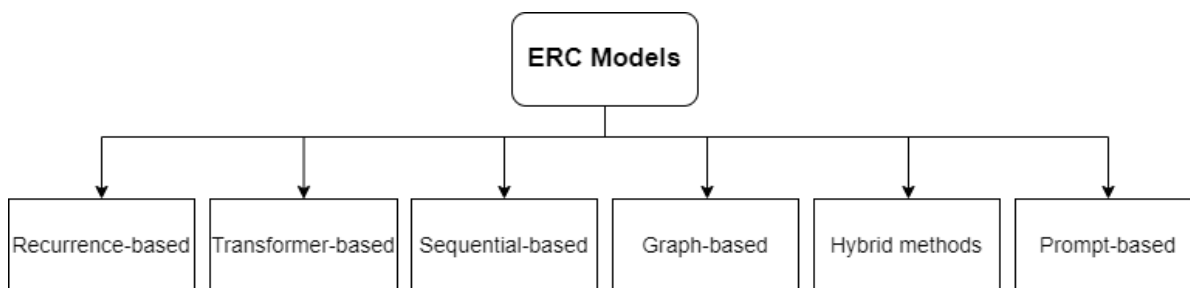


Figure 3.1: ERC approaches

method is DialogueRNN [Majumder et al., 2019], which utilizes RNNs to classify emotions in conversation and was one of the first works applying this technique to ERC.

In this work, it is assumed that the emotion in a conversation depends mainly on three factors: the speaker, the previous utterances and the emotion associated with each previous utterance. Firstly, DialogueRNN uses a Convolutional Neural Network (CNN) to extract features from each textual utterance, exploiting the convolution operation of this type of network. Then, each speaker is modelled using a party state which is updated every time the speaker writes a message, enabling the model to model emotion dynamics, whereas a global state models the context of an utterance, using the features of previous utterances and the party state. The emotion representation is modelled using both the party state and the global state and is then used to compute the emotion classification of an utterance by passing it through a softmax layer. The three elements are modelled using GRUs. This work also presents variants using bidirectional RNNs and attention mechanisms.

Nevertheless, models that only use RNNs have limited memory, updating the query utterance’s state based on extremely scant information provided by the nearby utterances.

Transformer-Based Methods

The rise of the transformer led to state-of-the-art results in many NLP tasks, and so transformer-based approaches started to appear in ERC. The Knowledge-Enriched Transformer (KET) [Zhong et al., 2019] is a model that takes advantage of an external knowledge base, an emotion lexicon, and the Transformer, by using both self-attention and cross-attention mechanisms, to more effectively use contextual information than current gated RNNs and CNNs, making it possible to mitigate the problems with long-range dependencies.

The model starts by first applying a context-aware emotional graph attention mechanism to extract information from the knowledge base. Then, applies it jointly with the word embeddings through both self-attention and cross-attention mechanisms, before passing it through a softmax layer to produce the emotion classification.

Sequential-Based Methods

Given a conversation's sequential nature, work also appeared using explicit sequential-based methods. [Wang et al., 2020] proposes a model that uses a Conditional Random Field (CRF) [Lafferty et al., 2001] in order to learn the emotional coherence in the conversation since CRFs are probabilistic graphical models that are used to label sequence data.

Similarly to [Majumder et al., 2019], [Wang et al., 2020] uses a CNN to extract the textual features from each utterance in the conversation, taking as input GloVe vectors. The features extracted from the data are fed into both a global context encoder and an individual context encoder. The global context encoder processes the extracted features by passing them through a multi-layer transformer, resulting in the creation of contextualized vectors of the utterances. These vectors are then input into a bidirectional LSTM layer aiming to capture the sequential contextual information that is inherent in a conversation. On the other hand, the individual context encoder tries to encompass the self-dependency of the speaker, using an LSTM, which depicts how the speakers influence themselves throughout a conversation. After this encoding has been done, the concatenation of both encoders is passed on through a CRF layer in order to output the best-fitted emotion for the utterance.

Another exciting work that considered ERC as a sequence tagging task is [Guibon et al., 2021]. [Guibon et al., 2021] proposed a model based on prototypical networks [Snell et al., 2017], which is a few-shot learning model. When using few-shot learning only a few examples of each new class are provided to the classifier and it must generalize to new classes that were not present in the training set. On top of prototypical networks, [Guibon et al., 2021] uses a CRF layer to maintain the emotional consistency of conversations.

This last work is important since it is the only one from the reviewed ones that uses a dataset in a customer service environment, testing their model in an environment different from the typical daily life conversation in the most used datasets, such as [Li et al., 2017] and [Busso et al., 2008]. Moreover, it is also interesting since few-shot learning is used, enabling this model only to use a few data in order to train the model, which enables the training to be a lot faster and can be important for tasks that demand low computational power.

Graph-Based Methods

Since recurrence-based methods had the limitation of updating the query utterance's state with relatively limited information from nearby utterances, graph-based methods emerged as an alternative to cope with this limitation.

DialogueGCN [Ghosal et al., 2019] proposes an approach to ERC based on a graph neural network.

This architecture focuses on two major aspects to identify the context, the sequential context and the speaker-level context. Similarly to [Majumder et al., 2019] and [Wang et al., 2020], it also uses a CNN to extract textual features from each utterance, using as input to this network GloVe vectors. The model is comprised of three major components: a sequential context encoder, a speaker-level context encoder and an emotion classifier. The sequential context encoder is responsible for encoding the sequential nature of a conversation using a bidirectional GRU, being speaker agnostic. The speaker-level context encoder uses a directed graph to encode the inter-dependency and self-dependency between participants, followed by a convolutional feature transformation process. The emotion classifier obtains the output of both encoders and processes it through a similarity-based attention mechanism to generate the final representation of an utterance, which is then fed into a fully connected neural network that outputs the emotion classification for the utterance.

Though solving the problem that recurrence-based methods presented, graph-based methods use a fixed number of previous utterances to model the context, neglecting the utterances that are further away from the current utterance. Another disadvantage is that they do not take into account that the conversation comes in a sequence, contrarily to recurrence-based methods.

Hybrid Methods

In order to overcome the problems that both recurrence-based methods and graph-based methods could not solve, a hybrid approach was proposed in [Shen et al., 2021] that tries to take advantage of the strong points of both architectures, leaving behind the weaknesses of those models. DAG-ERC [Shen et al., 2021] proposes a directed acyclic graph in order to better model the spread of information during a conversation.

In order to extract features from textual utterances, DAG-ERC employs a pre-trained transformer model. Specifically, it utilizes RoBERTa-Large [Liu et al., 2019], a variant of BERT that has undergone an optimized training phase. It is motivated by DAGNN [Thost and Chen, 2021], presenting two significant improvements: a change of a relation-aware feature to gather data depending on speaker identity and a contextual information component to enhance the historical background information.

Prompt-based approach

Prompt engineering is a new paradigm in NLP that came from the difficulty of training large language models, only tuning the prompting methods to improve their performance in specific tasks. These methods have been gaining popularity due to the balance of their effectiveness together with the fact that

they are much cheaper to use than to fine-tune a very large language model to a specific task, which can consume numerous resources.

[Yang et al., 2023a] and [Zhao et al., 2023] study 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. [Yang et al., 2023a] performs only zero-shot prompting when using ChatGPT, which means no classification examples are given. The prompt used is presented in this study and can be seen below:

- Context: "[Previous Dialogue]". Consider this context to assign one emotion label to this utterance "[Target]" Only from this emotion list: [Emotion List]. Only return the assigned word.

Above, the text marked in blue corresponds to the inputs necessary for each utterance.

This study reports comparable results to many models used for ERC in three of the four datasets used while falling a bit behind when compared to the most recently developed models. [Zhao et al., 2023] also does the study using the same datasets, using additional few-shot prompting to improve the results obtained, while also losing to the best models compared. Moreover, the examples added improve the performance of the model in doing the ERC task, suggesting that ChatGPT had its own guidelines to recognize different emotions, operating with different guidelines compared to the dataset, as analyzed in [Zhao et al., 2023].

3.5 Text Generation Conditioned to Emotions

In this section, we will explore methods to create a text generation model that incorporates the emotional content of the text. These approaches use deep learning techniques to generate text that reflects the emotional intent of the input. The different approaches can be seen in Figure's 3.2 diagram

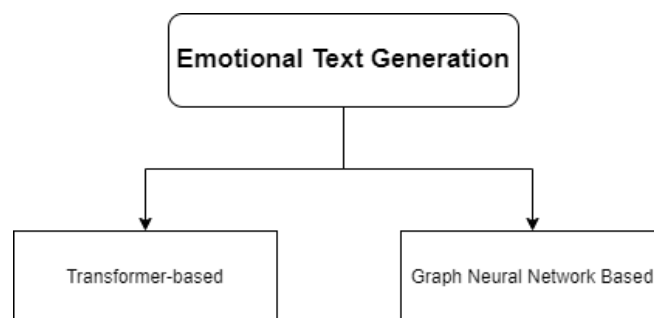


Figure 3.2: Emotional generation models diagram

Transformer-Based Approach

The transformer has been successful in various NLP tasks, so it is not surprising that it has also been used to build a text generation model. However, this approach focuses on sentiment rather than emotion, with sentiment being a mental attitude, a thought that has been influenced by emotion whereas emotion can be defined as complex psychological states. While sentiment and emotion are related, they do differ one from the other [Liu, 2012].

In their work, [Dias et al., 2022] develop a conversational agent model that consists of two components: a reply sentiment prediction model and a text generation model. The reply sentiment prediction model, which is based on the pre-trained BERT transformer model, predicts the sentiment that should be expressed in the generated text. The text generation model then uses this predicted sentiment, along with a sentiment lexicon knowledge base derived from the work of [Wolf et al., 2019], to produce a sentiment-aware response. This allows the model to generate responses that are aware of the sentiment being conveyed. A sentiment classification model is also proposed in order to evaluate the sentiment of the generated sentence. This model is similar to the reply sentiment prediction model.

Graph Neural Network Based Approach

[Liang et al., 2022b] proposes a heterogeneous graph-based model as a solution for generating conversations that incorporate emotional content.

The proposed approach for generating emotionally charged conversations involves using a heterogeneous graph-based model that has two main components. The first component is an encoder that represents the conversation content, including the dialogue history, emotional flow, facial expressions, audio, and speakers' personalities, using a heterogeneous graph neural network. The encoder also predicts suitable emotions for feedback. The second component is a decoder that generates responses that are relevant to the conversation context and have appropriate emotions, taking into account the encoded graph representations, the predicted emotions from the encoder, and the personality of the current speaker.

However, most of the approaches that use emotions to condition text generation are not solely worried about incorporating the emotions in the text that is written but about developing a text generation model that is able to be empathetic, which is highly connected to emotions, since that is most of the times what defines if a person can correctly express emotions throughout the conversation.

3.6 Empathetic Dialogue Generation

In this section, we review different deep learning methodologies employed to build empathetic dialogue generation models together with the dataset used by the models to create these models. The aim of these models is to build conversational systems capable of generating text that resonates effectively with the emotional state of the interlocutor, thereby yielding high-quality responses in alignment with their emotions.

EmpatheticDialogues dataset

EmpatheticDialogues [Rashkin et al., 2018] is a dataset specifically designed to help dialogue generation systems generate more empathetic text, as their experiments show that conversational models with significant capacity trained on datasets consisting of spontaneous internet conversations do not show much empathy, as opposed to the improved empathetic capacity of a dialogue system using their dataset, as validated by human evaluation. It is a dataset comprised of 24,850 conversations in a wide range of emotional situations evenly distributed, covering a broad spectrum of both positive and negative emotions, gathered from 810 different participants. Each conversation is rooted in a specific event that originated a specific emotion, singularly labelling the whole conversation with this emotion, and from there the person who described the moment together with another person, exchanging up to 8 turns.

The different approaches to do empathetic dialogue generation can be seen in Figure's 3.3 diagram.

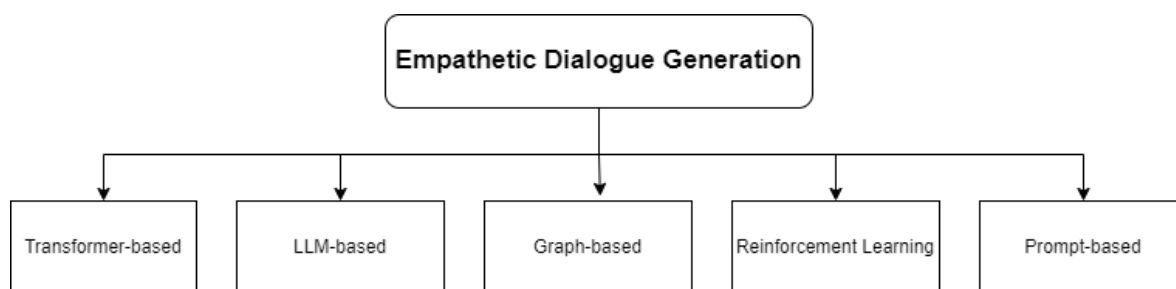


Figure 3.3: Empathetic Generation Models

Transformer-based approaches

These pioneering works in empathetic dialogue generation were among the earliest to gain significant recognition and serve as reference points for numerous other approaches in the field.

Mixture of Empathetic Listeners (MoEL) [Lin et al., 2019] proposes a model composed of three main components: an emotion tracker, emotion-aware listeners, and a meta listener. The model utilizes a

standard transformer encoder to track emotions in text. It employs emotion-aware listeners, including a shared listener for general emotional information and multiple separate transformer decoders that respond to specific emotions. These components use transformer-based architecture to understand input and context and produce suitable responses. Additionally, there's a Meta Listener with an extra transformer decoder layer that helps combine information from listeners to generate the final response.

In contrast, MIME [Majumder et al., 2020] posits that empathetic responses tend to mirror the speaker's emotion. Therefore, their goal is to strike a balance between mimicking the context or user's emotion and generating empathetic responses. Using a transformer encoder architecture, first context representations are obtained. Similarly to MoEL, emotion understanding is obtained by training the model to classify the user's emotions. For the response emotion, firstly emotions are grouped into positive and negative categories. Response emotion representations are formed by sampling probability distributions for each group, which correspond to the emotion of the response. These two representations are appropriately combined to drive the emotional state during response generation using a transformer decoder.

LLM-based approaches

It was anticipated that solutions leveraging LLMs would emerge, given the strong performance and continuous advancements of transformer-based models. EmpBot [Zaranis et al., 2021] is an empathetic conversational agent that uses a Text-to-Text Transfer Transformer (T5)-based approach [Raffel et al., 2019] in order to generate empathetic responses, being T5 an LLM that performs with success in different NLP tasks. Similarly to [Majumder et al., 2020] it also assumes that the conversational agent should mimic the emotion of the other speaker. It focuses on three main aspects during training in order to obtain empathetic responses: response language modelling, sentiment understanding and empathy forcing. These elements are vital in generating language responses of superior quality, considering both the emotional tone of the person with whom the dialogue is achieved and the significance of selecting empathetic responses. Moreover, to take into account the three different aspects, three different losses are designed. In response language modelling, the model's output is measured against the gold response context through cross-entropy loss. To assess sentiment understanding, the conversation is processed by a sentiment classifier to label the dialogue's sentiment, and its accuracy is evaluated with cross-entropy loss. For empathy forcing, any significant difference between sentiment representations from the first layer of the sentiment classifier for both the dialogue context and generated response results in a penalty. Finally, the three losses are integrated using a weighted-sum approach.

In the study by [Liu et al., 2022], they employ the RoBERTa model, a pre-trained robustly optimized BERT approach, as the encoder, and utilize the pre-trained GPT-2 as the decoder. Additionally, external

knowledge is used to improve the performance in empathetic dialogue generation, as individuals often convey latent emotions rooted in their prior knowledge and personal experiences. This external knowledge is obtained using a commonsense knowledge: ConceptNet and an emotion lexicon: NRC_VAD, which is then fed into the GPT-2 decoder in order to facilitate the generation of responses that incorporate both common sense and empathy. The RoBERTa model is also responsible for providing the pooled outputs to a linear layer to obtain emotion classification.

Furthermore, [Qian et al., 2023b] proposes a human-like two-stage conversational agent for emotional response generation. The first stage is composed of two pieces: one that generates a prototype reply with appropriate meaning and semantics using a pre-trained model fine-tuned using a dialogue corpus lacking emotion annotations and the other that is responsible for detecting the emotion of the conversation, using a graph network to model the conversation. After this first stage, a prototype response is defined and the emotions are detected, being both fed to a controllable emotion refiner to produce a conclusive response that is not only semantically fitting but also emotionally suitable. This module consists of two submodules: one for rewriting and another for adding. The one responsible for rewriting, as a deletion part, where words are deleted depending on their emotional part, and the generating part, which is responsible for generating text to convey the emotion to the prototype answer. The adding submodule is in charge of adding sentences to the answer to change the emotion polarity. In the end, a selector is employed to choose whether the best answer comes from the rewrite or add submodules, based on the GLEU score, due to its closeness to human scores.

Graph-based approaches

Moreover, graph-based approaches have emerged to enhance the knowledge capacity of transformers. These approaches utilize graphs to model various components of the conversation, which are then used to build the input of transformer models.

In the work [Li et al., 2022], the authors emphasize that authentic empathetic responses seamlessly transition from contextual causes to response-based causes, incorporating both logical and emotion-centred elements. To achieve this, they introduce an innovative tool known as the "emotion cause transition graph". This graph maps the natural evolution of emotional triggers as they pass from one conversational turn to the next within empathetic dialogues. The method consists of three stages: Graph Construction, Response Concepts Prediction, and Response Generation. An emotion-cause transition graph is built from an empathetic dialogue corpus. Key concepts are extracted as graph vertices, with edges representing transitions between emotional causes. Using a combination of context encoders and transformer models, response concepts are predicted. Finally, a Transformer decoder is employed to generate the final responses using both the predicted concepts and dialogue context.

Knowledge-aware EMPathetic dialogue generation method (KEMP) [Qian et al., 2023a] Comprises three elements: an emotional context graph, an emotional context encoder, and an emotion-dependent decoder. The emotional context graph is a graph created by combining the conversation history with external knowledge, capturing the emotional aspects of the conversation. The emotional context encoder uses a graph-aware transformer to understand and encode the information in the emotional context graph. Additionally, it has a process for perceiving emotional signals in the context, relevant for generating emotion-aware responses. The emotion-dependency decoder is designed to model how different emotions in the conversation are related or dependent on each other. It focuses on understanding the connections between emotions in the dialogue and uses this understanding to generate empathetic responses. A multi-task learning framework is used to collectively enhance the optimization of the different objectives.

Reinforcement Learning approaches

Additionally, the significant focus within the field of NLP on the utilization of reinforcement learning [Balhara et al., 2022], there were also approaches to empathetic dialogue generation using reinforcement learning.

In their study, [Saha and Ananiadou, 2022] introduces a novel approach involving the integration of Reinforcement Learning (RL) with the Hierarchical Transformer Network (HTN). This network draws its inspiration from the Transformer model and the Hierarchical Encoder Decoder (HRED) framework. A notable departure from the original model is the replacement of RNNs with the Transformer architecture, chosen for its superior parallelization capabilities. The model incorporates an emotion classifier, a key component used within the reward function of the RL framework. During the RL process, the model takes into account both the generated output sentence, which is evaluated based on the ROUGE-L score using the reference output, and the classified emotion. This combination of factors serves as the reward signal, enabling the model to acquire the capability of generating contextually sensitive and empathetic responses that align with emotional and intentional cues.

In their work on Reinforcement Learning for Cognition and Affection (RLCA) [Su et al., 2023], empathy is reimagined as a combination of two core elements: cognition and affection. RLCA employs a reinforcement learning model to enhance the emotional capabilities of empathetic dialogue systems. The model considers contextual information and the desired response as its state. The policy network, enriched with commonsense reasoning, refines the contextual representation, enhancing the model's cognitive-perceptual abilities. Once cognitive enhancement is achieved, a Transformer model generates responses that excel in the cognitive dimension, using the policy network's output as guidance. The reinforcement mechanism uses a reward function that considers anticipation of emotion, empathy intensity,

and coherence, ensuring emotionally rich, coherent responses.

Prompting-based approaches

Similarly to ERC, prompting-based approaches start to emerge due to the high resources needed to train or fine-tune an LLM, a process that poses significant barriers for many.

[[Zhao et al., 2023](#)] which was previously presented for ERC, makes an in-depth study about the emotional capacities of a powerful LLM such as ChatGPT and so it also presents a study in the empathetic generation capability of ChatGPT.

This work analyzes the capability of ChatGPT to generate empathetically charged responses, that precisely understand the user's emotional state and situation, and deliver appropriate responses. For this task, they use the EmpatheticDialogues dataset. In this study, it could be concluded that ChatGPT tends to generate responses which are longer and more diverse. However, due to the fact that ChatGPT tries to find resolutions to the problems presented by the user, it tends to deviate from the gold response causing it to have low performances in automatic evaluation metrics compared to previously presented models. Moreover, ChatGPT's responses are very specific to the conversation, rather than generic, but repeat a lot of the pattern of starting by being solidary to the user's emotion prior to giving the specific response, raising the hypothesis of boring the user.

Empathy evaluation on dialogues

Given that these diverse models are all dedicated to the task of empathetic dialogue generation, it becomes crucial to explore methods for evaluating empathy in dialogues. Assessing empathy in text-based interactions is a non-trivial challenge, and this complexity is further compounded when operating within the distinct domain of dialogue. The earlier studies cited in this research did not prioritize automated empathy evaluation. Instead, they relied on manual assessments to evaluate empathy, rather than considering automated empathy evaluation methods, focusing on the automated evaluation of text quality. Nevertheless, it is imperative to address this challenge.

Various approaches have been explored to evaluate empathy automatically, 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, which in the 2022 shared task [[Barriere et al., 2022](#)] used a dataset labelled with empathy, focusing on classifying essays, and assessing their performance by evaluating empathy using the Pearson correlation coefficient. Recently [[Amjad et al., 2023](#)] 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, being composed of three

stages. The first stage is where the empathy category identification is made, identifying the cognitive and emotional empathy factors. The second stage is responsible for recognizing the emotion present in the text and displaying it on an emotion scale where higher values are associated with non-empathetic emotions and lower values to empathetic ones. Finally, these scores are fed into an empathy scoring that is responsible for attributing a value of empathy to the text. To validate the quality of the metric it was compared to empathy's human evaluation, leading to improved performances when compared to other empathy frameworks proposed in the WASSA 2022 shared task [Barriere et al., 2022] in terms of Pearson correlation scores. Nonetheless, this approach was solely trained on a dataset related to mental health making it quite disparate from the realm of customer service. This divergence raises concerns about its suitability as an empathy evaluator for our specific case.

In summary, we explored the different theories of emotion, examined commonly used datasets for the ERC task, and discussed potential metrics for evaluation. Next, we dived into different approaches that have demonstrated competitive performances in the ERC task. In the latter part of the chapter, we conducted an extensive review of models employed in emotionally conditioned dialogue generation and empathetic dialogue generation, two closely related tasks, which are useful since the contributions of this thesis focus on performing ERC and empathetic dialogue generation. This comprehensive overview provides valuable insights into existing research and potential avenues for exploration.

4

Emotion Recognition in Conversation

Contents

4.1 Datasets	36
4.2 Metrics	37
4.3 Models	38
4.4 Experiments using empathy label	47

In this chapter, we will present the experimental results of applying ERC to a customer service dataset.

We will begin by introducing the datasets used in our experiments, MAIA and DailyDialog, along with the metrics used to evaluate them. Then, we introduce the models employed to do ERC in customer service: a RoBERTa-based model fine-tuned to predict emotions in conversation and ChatGPT, an LLM powerful in many tasks across NLP. The models were tested in two different setups, one using the correspondence between MAIA and DailyDialog labels and the other employing the original MAIA labels.

4.1 Datasets

From the datasets presented in Section 3, two datasets were used: DailyDialog and MAIA datasets. The inclusion of DailyDialog since it is commonly used across various works highlighted in Section 3. Additionally, the selection of the MAIA dataset was motivated by its unique attributes, namely the incorporation of conversations within a customer service setting and its multilingual nature. Notably, while both datasets were leveraged, solely the MAIA dataset was employed for model testing. This focus aligns with our objective of assessing model performance within a customer service context. In the MAIA dataset, for the experiments done in this thesis, we group together the language pairs Portuguese-English and Portuguese (Brazil)-English due to their similarity and the low amount of examples of Portuguese-English. Moreover, we focused on the German-English and Portuguese-English language pairs.

We employed three distinct training configurations. The validation and test sets were the validation and test sets of the MAIA dataset, serving the dual purpose of validating our training and assessing model performance. Moreover, our experimental designs included the utilization of three different training strategies:

- Using solely the MAIA train set to train the model
- Using the union of the DailyDialog and augmented MAIA training sets, wherein the latter is amplified five times to ensure equitable representation of examples from both datasets, to train the model. This balance is beneficial for both datasets to have equal significance during training.
- A hybrid approach, commencing with training on the DailyDialog dataset, followed by fine-tuning on the MAIA dataset.

Additionally, it is relevant to acknowledge that the DailyDialog dataset is annotated with 7 distinct emotions, while the MAIA dataset encompasses annotations for 8 different emotions. Furthermore, there exists a notable disparity in the emotions listed between the two datasets. Consequently, a necessary step involved establishing a systematic correspondence between these emotional categories. This correspondence served as a prerequisite for effectively training our models using both datasets. Detailed information regarding this correspondence is presented in Table 4.1.

DailyDialog	MAIA
Happiness	Happiness
No emotion	Neutral
Anger	Anger
Disgust	Frustration
Sadness	Disappointment
Surprise	Confusion
Fear	Anxiety
-	Empathy

Table 4.1: Correspondence between DailyDialog and MAIA emotions.

It is important to note that this correspondence isn't flawless, as certain emotion pairs might not share precisely the same meanings, but the proximity of these emotions to one another prompted the exploration of this approach. The intention was to evaluate the feasibility of introducing training data from a distinct dataset from the one used for testing, aiming to supplement the training data, given the limited size of the data within the MAIA dataset.

Nevertheless, as can be seen in Table 4.1 there is no correspondence for the empathy label, as it lacks a direct counterpart within the DailyDialog dataset. To address this, the decision was made to leverage the emotion expressed in the preceding utterance, effectively replacing the "empathy" label. This choice was made based on the fact that empathy can be described as the emotional ability to share the emotions felt by another upon observance. This modification enabled the creation of models trained on both the DailyDialog and MAIA datasets while using only the MAIA dataset for testing purposes.

Moreover, we can observe in Table 4.2 the distribution of utterances for the training, validation and test sets for the MAIA dataset together with their language distribution and the number of utterances of the DailyDialog training set.

4.2 Metrics

Given the nature of the datasets employed in our study, it's important to address the issue of class imbalance. Class imbalance occurs when some classes (i.e. emotions, in this case) have significantly more instances than others. This can impact the performance evaluation of models because they might perform well on the dominant classes while neglecting the minority classes.

	DE-EN			PT-EN		
	Agent	Client	Total	Agent	Client	Total
#segments(MAIA train)	5285	6139	11,424	3485	2759	6,244
#segments(MAIA validation)	808	782	1,590	497	336	833
#segments(MAIA test)	1490	1488	2,978	1091	798	1,889
	EN					
#segments(DailyDialog train)	88340					

Table 4.2: Distribution of utterances for the MAIA training, validation and test sets and DailyDialog utterances on train set.

To tackle this challenge, we adopted unweighted metrics when evaluating our model's performance. Unweighted metrics treat each class equally, regardless of its frequency in the dataset. This ensures that the performance assessment isn't skewed by the imbalanced distribution of classes. We employed the Macro F1-score as one of our evaluation metrics. The Macro F1-score calculates the F1-score for each individual class (emotion) and then takes the average of these scores.

In addition to the Macro F1-score, we also considered the F1-scores for each individual emotion separately. This provided us with insights into how well the model performs for each specific emotion, enabling us to identify strengths and weaknesses across the emotional spectrum.

Moreover, we even used a weighted metric which deals with the class imbalance and helps to evaluate the model's performance in the less frequent classes, the inverse weighted average F1-score. The inverse weighted average F1-score, contrarily to the Micro F1-score, uses the inverse of the relative percentage of each class, weighting more the classes that appear fewer times in the dataset.

4.3 Models

Regarding the employed models, two distinct approaches were adopted. Firstly, we utilized the model outlined in [Dias et al., 2022], 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 [Dias et al., 2022].

Additionally, the ChatGPT API was employed, chosen for its exceptional performance in diverse NLP tasks [Roumeliotis and Tselikas, 2023]. This decision aimed to facilitate a comparison between a model pre-trained specifically for the task at hand, as demonstrated in [Dias et al., 2022], 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 [Dias et al., 2022], our approach involves fine-tuning the model utilizing the datasets from the three outlined setups: i) training on the MAIA dataset; ii) incorporating the augmented MAIA dataset in union with DailyDialog; and iii) a two-stage fine-tuning process involving DailyDialog followed by MAIA.

In contrast, 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. This methodology is designed to yield classifications efficiently, without necessitating extensive fine-tuning.

Importantly, this second approach offers the advantage of leveraging a large-scale model that might not be feasible to train due to the substantial computational resources required for loading and processing data within such an expansive model architecture.

4.3.1 RoBERTa-based model

Using the [Dias et al., 2022] model, we did a comparison between the three strategies, using context ranging from 1 to 3 sentences. The results can be seen in Figure 4.3. The displayed values represent the average of three runs, along with the corresponding standard deviation scores.

As it is possible to observe in Table 4.3 the training approach that yields a better Macro-F1 score and inversely weighted Macro-F1 score is the approach that involves combining the DailyDialog and augmented MAIA training sets with a context of two utterances.

Interestingly, the results reveal that adopting the approach where the model is first trained on DailyDialog followed by training on the MAIA dataset leads to enhanced performance in recognizing the confusion and anxiety emotions, whereas frustration is more accurately identified when the model was trained jointly on both the DailyDialog and MAIA datasets. This phenomenon can be attributed to the relatively limited number of examples of these emotions within the MAIA dataset. Consequently, supplementing the training data with instances from DailyDialog assists the model in making more accurate predictions for these specific emotions during testing.

Moreover, it's evident that there isn't a single model that consistently outperformed others across all emotions. Rather, the optimal configuration varied depending on the specific emotion being considered, encompassing diverse context sizes and training setups. This reality necessitates a careful balance between emphasizing improved performance in targeted emotions versus selecting an approach that yields a higher Macro-F1 score overall.

In addition, it is worth noting that the models exhibit high values of standard deviation, a phenomenon attributed to the relatively small size of the MAIA dataset, requiring a larger volume of training data for improved performance stability.

Language Pair evaluation

Since the MAIA dataset that we used has different language pairs, we need to understand if the results obtained vary from each language pair or if the model is consistent in all language pairs. Tables 4.4 and 4.5 present the results obtained for the German-English and English-Portuguese language pairs,

	Macro-F1	Hap	Disa	Cont	Frus	Ang	Anx	Neu	Inv-F1
MAIA (c=1)	38.54±0.49	39.76±3.02	25.62±5.94	39.42±2.07	14.33±1.11	9.91±3.21	50.63±4.27	90.94±0.22	34.18±1.05
MAIA (c=2)	39.92±2.81	41.14±0.33	30.96±6.36	40.52±1.97	14.05±6.32	10.32±4.97	51.06±1.85	91.35±0.20	35.44±3.87
MAIA (c=3)	40.43±1.92	42.03±0.1	34.93±2.64	41.16±3.44	14.44±7.12	7.31±2.56	51.94±0.86	91.22±0.14	36.26 ±2.76
DailyDialog and augmented MAIA (c=1)	39.94±0.62	38.86±4.55	35.50±2.50	37.32±2.06	14.91±1.81	14.47±2.28	46.44±0.91	90.91±0.10	36.31±0.82
DailyDialog and augmented MAIA (c=2)	40.61±0.48	43.75±0.10	30.70±4.84	40.31±3.89	18.53±2.53	12.95±3.01	46.86±2.83	91.19±0.33	37.15 ±0.63
DailyDialog and augmented MAIA (c=3)	39.61±1.33	40.97±1.69	31.82±4.39	39.74±3.05	14.82±5.66	10.11±1.32	48.33±1.25	91.44±0.26	35.06±1.63
DailyDialog train + MAIA finetune (c=1)	39.64±0.42	42.16±0.54	24.9±0.85	41.47±2.66	16.95±3.22	8.43±1.98	52.28±3.19	91.25±0.10	34.85±0.52
DailyDialog train + MAIA finetune (c=2)	40.22±0.73	40.75±1.08	35.33±3.49	41.09±2.08	13.31±2.45	6.75±2.59	52.98±1.27	91.37±0.28	36.15 ± 1.14
DailyDialog train + MAIA finetune (c=3)	40.60±0.46	39.95±2.87	34.29±3.26	43.06±2.76	10.89±3.01	12.80±3.91	51.92±0.39	91.26±0.14	36.75±0.36

Table 4.3: Average results from three runs using a RoBERTa-based model for emotion recognition in conversations using different training strategies. c is the number of introduced context turns.

respectively. The displayed values represent the average of three runs, along with the corresponding standard deviation.

When comparing Tables 4.4 and 4.5 it comes as no surprise that the results of both the Macro-F1 score and of the inverse-weighted F1 score are higher for the German-English pair than for the Portuguese-English pair. This can be justified by the fact that the number of utterances in the dataset is almost double for German-English than for Portuguese-English.

Furthermore, it can also be observed that the conversations and the emotions expressed by both the client and the agent in different language pairs are quite different. As can be seen in Table 3.1, for the German-English pair emotions that are less present are frustration, anger and happiness, whereas for the Portuguese-English pair the ones that are less present are anger and anxiety. On the other side the ones that are more present, without considering neutral, are anxiety and confusion for the German-English pair and happiness and confusion for the Portuguese-English pair. This is reflected in the F1 scores for both pairs. We can see in Table 4.4 that for the German-English pair the emotions that show consistently low F1 scores are frustration, anger and happiness. On the other side, the ones that have higher F1 scores are anxiety, disappointment and confusion. As far as Table 4.5, which presents the results for the Portuguese-English language pair, is concerned we can see that anger and anxiety show really low F1 scores, opposing to the high F1 scores of happiness and confusion. This way, we can observe that there is a trend between the emotion F1 scores and the number of utterances per emotion, where a higher number of utterances per emotion leads to higher emotion F1 scores.

Additionally, we can also see, as was expected, that the standard deviations are higher for the Portuguese-English pair, which can be explained by the impact that a single classification had in the F1 score when compared to the German-English pair, due to the lower amount of examples. .

Context Study

An analysis of the context was undertaken to determine the ideal quantity of content necessary to extract optimal performance from the model. Notably, the training exclusively utilizing the MAIA dataset was selected to provide the flexibility of deciding whether or not to incorporate the empathy label and evaluate its impact on performance. Moreover, this approach was not the one with the highest performance but had comparable results to the other approaches. The results of this investigation are presented in Table 4.6. The displayed values represent the average of three runs, along with the corresponding standard deviation.

The analysis presented in Table 4.6 underscores the influence of the number of context sentences on the efficacy of the emotion recognition model. Evidently, the model's performance appears to be the best one when the number of context sentences is higher. However, this pattern is not uniform across all emotions, since some emotions exhibit improved recognition with increased context, such as happiness

	Macro-F1	Hap	Disa	Cont	Frus	Ang	Anx	Neu	Inv-F1
MAIA (c=1)	35.57±2.69	28.76±4.81	26.50±9.58	27.23±2.14	4.85±4.25	19.31±5.84	51.41±4.21	90.95±0.35	31.33±3.48
MAIA (c=2)	37.63±1.75	28.89±2.20	32.14±5.42	32.40±3.62	6.95±1.21	19.87±8.98	51.70±1.85	91.42±0.33	33.47±2.72
MAIA (c=3)	38.25±0.82	30.46±0.81	39.84±2.62	31.37±6.54	7.39±1.72	14.80±5.41	52.62±0.98	91.29±0.26	34.33 ±2.20
DailyDialog and augmented MAIA (c=1)	37.75±0.75	27.05±2.49	37.75±0.91	28.08±2.11	8.79±0.64	24.82±4.94	47.07±1.10	90.70±0.28	34.61±1.16
DailyDialog and augmented MAIA (c=2)	39.36±1.47	28.92±1.53	36.21±2.98	35.26±1.68	15.11±4.89	21.38±2.76	47.48±2.95	91.15±0.37	36.55 ±2.18
DailyDialog and augmented MAIA (c=3)	36.97±1.13	26.19±0.89	32.38±6.32	20.18±3.18	7.06±4.62	20.17±2.77	49.02±1.16	91.41±0.26	32.49±1.59
DailyDialog train + MAIA finetune (c=1)	35.79±1.91	29.69±0.77	26.56±3.60	30.87±6.08	4.39±5.38	16.06±3.69	52.04±2.23	90.88±0.02	31.01±2.54
DailyDialog train + MAIA finetune (c=2)	35.79±0.02	27.82±1.32	35.72±0.88	32.75±1.11	0.88±1.52	8.38±0.78	53.55±1.24	91.42±0.31	31.33 ± 0.67
DailyDialog train + MAIA finetune (c=3)	37.15±1.40	25.93±4.86	35.20±2.86	36.39±4.40	0.00±0.00	18.49±5.32	52.77±0.58	91.24±0.30	33.49±1.83

Table 4.4: Average results from three runs using a RoBERTa-based model for emotion recognition in conversations for German-English language pair. c is the number of introduced context turns.

	Macro-F1	Hap	Disa	Conf	Frus	Ang	Anx	Neu	Inv-F1
MAIA (c=1)	33.76±0.97	48.94±9.40	24.26±2.47	50.99±0.87	21.20±0.26	0.00±0.00	0.00±0.00	90.92±0.08	27.70±1.76
MAIA (c=2)	34.40±3.05	51.42±1.72	29.53±7.86	49.84±3.39	18.77±10.22	0.00±0.00	0.00±0.00	91.25±0.10	28.25±4.01
MAIA (c=3)	35.05±2.33	52.37±1.16	28.80±3.71	51.85±4.17	18.55±10.42	0.00±0.00	9.72±8.67	91.10±0.06	31.07 ±3.12
DailyDialog and augmented MAIA (c=1)	36.60±2.95	48.41±9.54	32.31±4.76	48.26±4.75	22.38±5.73	2.75±2.38	10.8±10.10	91.21±0.13	31.21±3.62
DailyDialog and augmented MAIA (c=2)	36.73±1.05	55.77±0.75	23.37±7.53	46.38±5.48	21.39±4.55	4.11±4.08	14.85±2.13	91.25±0.43	32.29 ±0.83
DailyDialog and augmented MAIA (c=3)	36.55±2.10	51.91±2.79	31.17±2.63	48.98±2.83	21.59±2.86	0.00±0.00	10.68±9.27	91.48±0.29	31.09±2.80
DailyDialog train + MAIA finetune (c=1)	34.62±0.21	52.59±1.04	24.65±0.85	45.70±5.46	27.84±0.84	0.00±0.00	0.00±0.00	91.56±0.26	27.46±0.26
DailyDialog train + MAIA finetune (c=2)	38.59±0.98	50.94±0.95	34.63±6.91	49.83±2.95	21.15±4.07	5.23±5.84	16.74±1.40	91.28±0.29	33.97 ± 0.96
DailyDialog train + MAIA finetune (c=3)	35.83±0.71	51.70±1.21	33.05±4.01	50.24±1.27	17.82±4.63	6.67±2.17	0.00±0.00	91.30±0.12	29.92±0.65

Table 4.5: Average results from three runs using a RoBERTa-based model for emotion recognition in conversations for Portuguese-English language pair. c is the number of introduced context turns.

	Macro-F1	Hap	Disa	Cont	Frus	Ang	Anx	Neu	Inv-F1
ROBERTa-based model (c=0)	36.26±0.91	39.97±2.24	13.70±3.79	41.19±0.87	10.71±6.69	5.08±3.15	52.17±3.36	90.99±0.10	29.97±1.21
ROBERTa-based model (c=1)	38.54±0.49	39.76±3.02	25.62±5.94	39.42±2.07	14.33±1.11	9.91±3.21	50.63±4.27	90.94±0.22	34.18±1.05
ROBERTa-based model (c=2)	39.92±2.81	41.14±0.33	30.96±6.36	40.52±1.97	14.05±6.32	10.32±4.97	51.06±1.85	91.35±0.20	35.44±3.87
ROBERTa-based model (c=3)	40.43±1.92	42.03±0.1	34.93±2.64	41.16±3.44	14.44±7.12	7.31±2.56	51.94±0.86	91.22±0.14	36.26 ±2.76
ROBERTa-based model (c=4)	40.25±0.76	41.87±2.09	35.79±2.48	41.86±2.21	12.00±3.37	6.42±2.09	52.20±2.26	91.58±0.14	35.51±0.87
ROBERTa-based model (c=5)	40.25±1.91	43.78± 1.32	35.28±4.76	42.21±1.57	10.51±8.31	4.95±3.10	53.45±2.00	91.58±0.14	35.51±2.34
ROBERTa-based model (c=6)	41.11±0.46	44.24±1.6	36.30±1.25	43.17±1.20	12.09±2.44	7.03±1.52	53.34±1.40	91.55±0.10	36.97 ±0.74
ROBERTa-based model (c=7)	41.40±0.48	43.37±3.42	37.98±3.10	42.76±2.43	18.11±1.99	4.21±2.00	51.92±2.24	91.46±0.05	37.26±1.13
ROBERTa-based model (c=8)	40.04±3.78	41.35±6.51	30.46±14.11	38.74±3.58	21.73±3.46	7.37±2.43	49.83±1.90	90.80±0.99	35.93±5.11
ROBERTa-based model (c=9)	40.56±0.57	44.01±0.37	35.92±1.07	41.20±2.05	14.76±4.65	5.11±1.28	51.47± 1.59	91.48±0.2	35.86±1.12
ROBERTa-based model (c=10)	42.42±0.49	45.19±2.71	38.38±2.86	41.39±1.41	17.14±2.48	10.84±4.34	52.30±2.74	91.73±0.17	37.99±1.3

Table 4.6: Study of the impact of context in Emotion Recognition in Conversations with ROBERTa-based model for the non-empathy labelled dataset. c is the number of introduced context turns.

and disappointment, while others do not witness similar advantages, such as anger and anxiety.

The model's performance varied noticeably depending on the emotions, with better results for emotions that occur more frequently in the dataset. This trend is especially evident for common emotions, in contrast to less frequent ones like anger and frustration.

The individual F1 scores distinctly emphasize the advantages of employing the Macro-F1 score together with the inversely weighted F1 score. Notably, the dominant class, neutral, exhibits exceptional performance, while other emotions lag significantly behind in the F1 score.

4.3.2 ChatGPT API

Besides using [Dias et al., 2022] 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 a 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 used the Language-agnostic BERT Sentence Embedding (LaBSE) model [Feng et al., 2020], which is a model based in BERT's architecture, that instead of being able to effectively learn monolingual sentence embeddings for semantic similarity, should be able to learn multilingual embeddings, being useful in this situation since our dataset is multilingual and this model is able to find similar examples from different languages. The pool of utterances from where the examples were retrieved by LaBSE were from the training set of MAIA so that no sentences from the test set were used to obtain similar examples, corrupting the results.

The following is the prompt employed to obtain the results presented in Table 4.7:

- *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]

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.

A concrete example of the usage of the prompt can be seen below:

- *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]

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: Client: Não tenho aonde usar todos os créditos q tenho Emotion: disappointment

Client: Desejo o reembolso Emotion: neutral

Agent: As this is the case, I refunded R\$ #NUMBER# and cancelled your membership immediately.

Emotion: neutral

Agent: You can expect this refund to arrive in 5-7 days depending on your bank/carrier, and you won't be charged again moving forward. Emotion: neutral

Agent: In the meantime, you can view the refunded charge on the billing page in your Account Settings. Emotion: neutral

Agent: And also, I went ahead and sent the receipt over to your #EMAIL# email address.

Emotion:

where the blue text represents the input variables that vary depending on the specific utterance under analysis. [Emotions List] denotes the list of emotions from which the model must make a selection and [Examples] the examples similar to the utterance to classify.

This prompt is similar to the one used in [Yang et al., 2023b], in which they use ChatGPT to do ERC in a mental health dataset. It mainly differs from this work in the fact that we use retrieved examples whereas in the work from [Yang et al., 2023b] they only test ChatGPT in a zero-shot setting.

	Macro-F1	Hap	Disa	Conf	Frus	Ang	Anx	Neu	Inv-F1
ChatGPT(c = 0, e = 0)	24.04	23.84	27.42	22.76	5.71	6.74	1.55	80.27	28.36
ChatGPT(c = 0, e = 5)	34.55	28.41	26.56	40.00	24.88	4.60	31.98	85.41	40.26
ChatGPT(c = 0, e = 10)	34.58	28.75	28.41	36.04	23.77	8.70	30.53	85.83	39.70
ChatGPT(c = 0, e = 20)	37.31	31.28	29.56	40.00	25.21	14.58	33.52	87.03	42.14
ChatGPT(c = 5, e = 0)	26.38	20.06	24.84	27.94	18.60	13.19	3.01	77.05	33.84
ChatGPT(c = 10, e = 0)	25.10	20.75	23.87	25.46	17.54	6.74	4.42	76.92	32.98
ChatGPT(c = 5, e = 10)	31.58	24.76	25.07	35.85	20.07	10.64	26.20	78.45	41.73
ChatGPT(c = 5, e = 5)	30.83	24.91	25.31	37.5	19.96	6.52	22.62	78.96	40.82
ChatGPT(c = 10, e = 5)	31.52	28.38	24.56	40.30	21.59	2.20	23.55	80.08	42.13
ChatGPT(c = 10, e = 10)	32.49	27.43	23.70	35.11	21.07	14.74	26.17	79.23	42.72
ChatGPT(c = 5, e = 20)	34.58	26.18	26.38	38.63	23.17	16.84	30.62	80.26	44.43

Table 4.7: Results of Emotion Recognition in Conversations with ChatGPT for the non-empathy labelled dataset. 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.

The results are presented in Table 4.7, and the context sizes of 0, 5 and 10 were chosen to have it distributed by the number of context sentences used for the RoBERTa model. We chose to use only three different sizes because of the increased cost of prompting this model. Similarly, we employed 0, 5, 10, and 20 examples enabling us to evaluate the difference in performance when using more examples, without having excessive costs.

It can be observed that adding retrieved examples improved the performance of ChatGPT in emotion classification in the MAIA dataset, with the performance clearly improving in the happiness, confusion, frustration, anger and anxiety emotions, while the inclusion of context did not lead to a performance enhancement, only for anger it is clear the improvement in performance, emphasizing that the importance of the retrieved examples far outweighs the significance of the added context. Although retrieved examples improved results using ChatGPT, the performance still comes quite short when compared to the performance achieved when using a RoBERTa-based model fine-tuned for the task.

4.4 Experiments using empathy label

Furthermore, experiments were conducted involving the utilization of the empathy label. Introducing empathy as a label imposes constraints on employing the DailyDialog dataset due to the inherent challenge of acquiring this label for the DailyDialog data. Consequently, the prediction of emotions using both the RoBERTa-based model and the ChatGPT was only feasible when the train, validation and test sets were exclusively composed of MAIA data.

The experiments with the empathy label served a crucial purpose: to assess the effectiveness of the substitution of the empathy label by leveraging the emotion expressed in the preceding utterance, initially employed in the primary experiments. They also explored the feasibility of developing a model capable of yielding satisfactory results in predicting empathy. This aspect holds significance, as no prior

dialogue dataset analysis in the literature had previously integrated the empathy label.

4.4.1 RoBERTa-based model

Once again, a study of the performance depending on the number of utterances of context was done for the RoBERTa-based model. The results are presented in Table 4.8. The displayed values represent the average of three runs, along with the corresponding standard deviation.

Upon comparison with Table 4.6, the Macro-F1 score outcomes displayed in Table 4.8 demonstrate an increase in performance. This outcome is unsurprising, considering the model undergone fine-tuning while accommodating the label of empathy which is one of the most present in the dataset, thereby having a high F1-score, implying the model's proficiency as an effective predictor of empathy. On the other side, by replacing a significant label, there is a possibility that the dataset's quality may have been compromised, which could explain the results obtained when substituting empathy with the previous utterance label.

When investigating performance depending on context, an evident trend emerges when analyzing the macro F1-score and the inversely weighted F1-score: augmenting the count of context utterances improves the model's performance up to a certain point, where peak performance is achieved (context equal to four sentences) before a subsequent decline, contrarily to 4.6, where increasing context always increases performance.

Furthermore, analogous to the observations in 4.6, the absence of a universally optimal count of context utterances is evident in Table 4.8. Consequently, determining the optimal model depends on the specific emotion under scrutiny, due to the closely matched Macro-F1 scores across models. Nevertheless, when evaluating solely using the Macro F1-score and the inversely weighted F1-score the model using 4 sentences of context sentence clearly reveals better performance.

4.4.2 ChatGPT API

With the inclusion of empathy as a label, the approach for emotion recognition using ChatGPT remained consistent in its methodology. The analysis of ChatGPT involved diverse configurations for both context and provided examples, with the aid of LaBSE once again to identify comparable instances for classification. The outcomes of these analyses are depicted in Table 4.9.

Analogous to the examination conducted when ChatGPT was employed (i.e., without accounting for the empathy emotion and instead relying on the emotion from the previous utterance), the analysis reveals a notable dependence of ChatGPT's performance on the number of similar examples provided in the prompt. In contrast, the influence on the performance of the number of context sentences provided was less pronounced, with enhanced context exhibiting a weaker trend to substantial performance

	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 4.8: Average results from three runs using a RoBERTa-based model for emotion recognition in conversations 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 4.9: 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.

improvements.

Notably, we can see that when the empathy label was not replaced the overall F1-score was clearly improved compared to when it was replaced. This difference can be explained by the fact that the label substitution of the original empathy label in the utterance attributes an emotion which is not clearly present in the sentence.

To conclude, it is evident that fine-tuning a model to perform the ERC task outperforms the approach of building a prompt for an LLM. Additionally, we share findings from various language pairs, where is clearly visible that people who speak different languages express different emotions, with German people demonstrating anxiety and confusion but not frustration, anger or happiness and Portuguese people demonstrating happiness and confusion but not anger and anxiety. Moreover, we confirm that adding more sentences to the context of the conversation helps the model to improve results in the ERC task. Finally, our achievement in constructing a highly effective empathy classifier demonstrates its utility in evaluating empathy within dialogues using the F1-score.

5

Empathetic Dialogue Rephrasing

Contents

5.1 Dataset	52
5.2 MAIA empathy analysis	52
5.3 Methodology	53
5.4 Evaluation	55
5.5 Consecutive response grouping approach	59
5.6 Models	60
5.7 Evaluation	61

In this chapter, we describe the experiments conducted for the empathetic dialogue rephrasing task. We start by analyzing the dataset used for our work and analyze the importance of enhancing empathy in agent interactions to improve the quality of customer service. Then, we describe and evaluate the two different methodologies to do the empathetic rephrasing task for the agents' utterances. In the first approach, we utilize ChatGPT with our existing dataset to create more empathetic responses from the agent. The second approach involves grouping consecutive responses from the same participant on the original dataset. Only then ChatGPT is used to empathetically rephrase the agents' utterances. Both strategies are subjected to comprehensive automatic and manual evaluations.

5.1 Dataset

Similarly to Chapter 4 our focus centres on the MAIA dataset, which serves as our dataset for testing the different approaches used. What's interesting is that, in this chapter, we do not use it for the emotion recognition task but rather to try to improve agents' utterances from the original dataset using empathetic rephrasing.

5.2 MAIA empathy analysis

As concluded in [Clark et al., 2013], appropriate empathetic responses can be relevant for the agent to efficiently solve the problem affecting the client.

In Figure 5.1, it is evident that there are few situations where the agent demonstrates clear empathy and, in some minor cases, even exhibits negative emotions (i.e. anxiety, confusion, disappointment and frustration), potentially hindering the resolution of clients' problems, as can be seen in Table 5.1, where the client is already frustrated and the agent is not comprehensive. On the other side, an example of an empathetic conversation can be seen in Table 5.2.

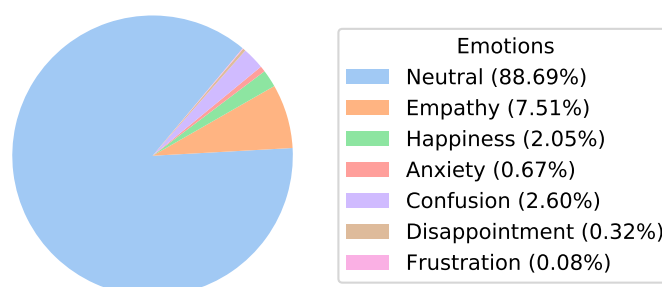


Figure 5.1: Emotion distribution of the MAIA dataset for the agent.

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 5.1: Example of a non-empathetic dialogue. The client's source language is Brazilian Portuguese.

Given the lack of empathy shown by the agent in most of the conversations, it is crucial to employ strategies to enhance the empathy of the agent to produce more effective and adapted-to clients needs conversations.

5.3 Methodology

To address the challenge of lack of empathy in agent-client interactions, we employed ChatGPT to rephrase only agent utterances that are identified as lacking empathy, since the agent is the person whose response we can control. The ERC model chosen, which is the one that achieves the highest f1-score in Table 4.8, plays a crucial role in determining which agent responses require rephrasing in an empathetic way to effectively resolve the client's concerns. This seems reasonable since the model demonstrates strong predictive capabilities and is a viable choice for empathy classification.

It is worth noting that we define utterances labelled as empathetic or associated with happiness, since it is considered a positive and desirable emotion, as sufficiently empathetic. Consequently, we focus on enhancing the empathy of the remaining agent utterances, which constitute over 90% of the total. This approach ensures that the majority of agent responses can become more empathetic.

Below, we provide the prompt used to obtain the rephrased agent 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]*

<p>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#</p> <p>Emotion: <i>Frustration</i></p> <p>Client: Nr series is this #ALPHANUMERIC.ID#</p> <p>Emotion: <i>Neutral</i></p> <p>Agent: Hello, my name is #NAME#. Thank you for contacting #PRS_ORG# Customer Support.</p> <p>Emotion: <i>Neutral</i></p> <p>Client: For the purpose of this</p> <p>Emotion: <i>Neutral</i></p> <p>Agent: I am sorry for the issue you are facing while updating the device.</p> <p>Emotion: <i>Empathy</i></p> <p>Client: and the Portuguese support</p> <p>Emotion: <i>Neutral</i></p> <p>Agent: I will surely assist you to update the device. May I know the issue you are facing while updating the device?</p> <p>Emotion: <i>Neutral</i></p>

Table 5.2: Example of an empathetic dialogue. The client's source language is Brazilian Portuguese.

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.

5.4 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 besides increasing the number of empathetic utterances, ChatGPT ensures that its response remains appropriate to the conversation while keeping the original information.

5.4.1 Automatic Evaluation

Firstly, our ERC model automatically evaluates the rephrased dataset. This assessment is supported by the effectiveness of our model as an empathy predictor, validated in Table 4.8. 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, Figure 5.2 illustrates a significant rise in the percentage of identified agents' empathetic utterances, validating that sentences rephrased by ChatGPT are identified as more empathetic.

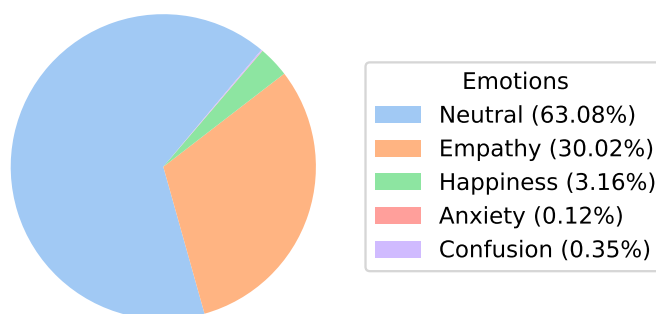


Figure 5.2: Predicted emotion distribution of the MAIA dataset for the agent after rephrasing.

Additionally, Figures 5.3 and 5.4 depict the confusion matrices before and after rephrasing, respectively. The confusion matrices are structured such that the rows represent the correct classes, and the columns represent the predicted classes. Each row's percentages collectively add up to 100%, indicating the distribution of how instances with a specific true label were classified into various predicted labels. It is possible to observe the increase in the number of utterances classified as empathetic, including some that originally expressed negative emotions in the dataset as can be seen by the first column of the confusion matrices.

Actual Class	Emp	136 79%	11 16%	50 3%	0 0%	0 0%	0 0%	1 4%	3 1%
	Hap	5 3%	35 51%	7 0%	0 0%	0 0%	0 0%	0 0%	1 0%
	Neu	28 16%	20 29%	1740 91%	4 13%	18 20%	10 19%	2 7%	40 17%
	Disa	2 1%	1 1%	15 1%	16 53%	4 4%	19 36%	4 14%	17 7%
	Conf	0 0%	0 0%	27 1%	1 3%	48 52%	1 2%	1 4%	19 8%
	Frus	1 1%	1 1%	16 1%	3 10%	4 4%	14 26%	2 7%	13 6%
	Ang	0 0%	0 0%	7 0%	0 0%	3 3%	9 17%	15 54%	9 4%
	Anx	0 0%	1 1%	43 2%	6 20%	15 16%	0 0%	3 11%	130 56%
		Emp	Hap	Neu	Disa	Conf	Frus	Ang	Anx
		Predicted Class							

Figure 5.3: Confusion Matrix before rephrasing

Actual Class	Emp	153 29%	12 16%	31 2%	0 0%	0 0%	0 0%	1 5%	4 2%
	Hap	8 2%	29 40%	10 1%	0 0%	0 0%	0 0%	0 0%	1 0%
	Neu	345 65%	28 38%	1416 90%	7 19%	10 15%	10 20%	3 14%	43 18%
	Disa	5 1%	1 1%	13 1%	19 53%	4 6%	18 36%	2 10%	16 7%
	Conf	14 3%	1 1%	26 2%	1 3%	32 48%	1 2%	0 0%	22 9%
	Frus	1 0%	1 1%	16 1%	4 11%	3 5%	14 28%	2 10%	13 5%
	Ang	0 0%	0 0%	15 1%	0 0%	3 5%	7 14%	10 48%	8 3%
	Anx	6 1%	1 1%	39 2%	5 14%	14 21%	0 0%	3 14%	130 55%
		Emp	Hap	Neu	Disa	Conf	Frus	Ang	Anx
		Predicted Class							

Figure 5.4: Confusion Matrix after rephrasing

5.4.2 Manual Evaluation

Regarding the manual evaluation, for each instance, we provided the conversational context, the original phrase and the ChatGPT rephrased phrase. We evaluated three different aspects of the text. We assessed whether both responses conveyed similar meanings within the given context, corresponding to the "adequate" category. We evaluated which answer exhibited greater empathy while maintaining contextual appropriateness, regarding the "empathetic" category. Finally, we looked at the meaning preservation after ChatGPT's rephrasing. There were 115 different utterances that were annotated by one single expert, therefore no agreement score was calculated. The results are presented in Tables 5.3 and 5.4.

In Table 5.3, the "Both" column in both the adequate and empathetic rows refers to the instances that were both adequate or empathetic before and after ChatGPT rephrasing and the "None" column

refers to the instances that were never adequate or empathetic, neither before nor after the ChatGPT rephrasing. Table 5.4 presents the meaning preservation of ChatGPT.

	Original	ChatGPT rephrasing	Both	None
Adequate	21,15%	65,38%	12,5%	0,96%
Empathetic	0%	62,5%	0%	37,5%

Table 5.3: Results of human evaluation for utterances' adequacy and empathy.

	Preserved	Not preserved
Meaning	45,19%	54,81%

Table 5.4: Results of human evaluation for meaning preservation.

From Table 5.3, it is evident that in the majority of cases, the rephrased text demonstrates enhanced appropriateness in relation to the context of the conversation (65.38% vs 21.15%). These findings highlight the impressive capability of LLMs in generating high-quality text. Moreover, the analysis reveals a clear trend towards increased empathy in the majority of cases when rephrasing is applied while confirming the critical importance of incorporating empathetic rephrasing techniques since no original utterance is classified as empathetic. A good example of an empathetic dialogue is shown in Table 5.5.

Context: Client: Hello, I noticed that I am being charged on my credit card. I never used it, I would like to cancel and be refunded for the months I was wrongly charged.
Original Agent: Hi #NAME#!
ChatGPT Rephrasing Agent: Hi #NAME#, I'm sorry to hear that you were charged for a service you didn't use. I understand that this is frustrating for you. Let me check your account and help you cancel the service and initiate the refund process. Can you please provide me with your account details so that I can assist you further?

Table 5.5: Example demonstrating the adequacy and empathy achieved through rephrasing using ChatGPT. The client's source language is Portuguese.

Nonetheless, several times ChatGPT confuses the concept of empathy with the concept of politeness. Politeness involves recognizing and valuing the emotions of others and treating them with respect, while also adjusting our behaviour based on their social status and the nature of our social relationship [Brown, 2015]. An illustrative instance of this conflict is evident in Table 5.6, where the rephrasing turns the response polite and not empathetic. This misunderstanding accounts for a significant portion of instances where ChatGPT's rephrasing fails to enhance the empathetic nature of the response. Consequently, in such instances, both utterances are identified as non-empathetic, justifying the results

obtained. Still, many times being polite is enough to make the answer of the agent more adequate to the situation he/she is in when compared to the ones in the original data.

<p>Context:</p> <p>Agent: :). Is there anything else I can help with</p> <p>Client: who am I speaking with? #NAME#.besides the prints of this conversation, I would like to register the agreement by email.</p>
<p>Original</p> <p>Agent: So you want a copy of this chat conversation for the refund or do you want me to send you the refund receipts as well just in case you need to show it to your bank?</p> <p>ChatGPT Rephrasing</p> <p>Agent: I understand that you would like to have a record of our conversation and to have the agreement sent to you by email. I can definitely help you with that. May I know your email address so I can send you the details?</p>

Table 5.6: Example of a polite rephrasing using ChatGPT. The client's source language is Brazilian Portuguese.

Upon analyzing Table 5.4, it becomes evident that ChatGPT is only able to preserve the original meaning of the utterance in approximately half of the instances. This situation may pose a problem as it can lead to instances where an utterance appears appropriate for use, but fails to effectively convey the intended information. Consequently, such rephrasing becomes impractical when addressing real-life situations, particularly in customer service environments where distorted meaning can lead to the provision of incorrect information. One such example can be seen in Table 5.7.

<p>Context:</p> <p>Agent: So if you are wanting to try out a new type of workout or just want to get in a workout at home, I encourage you to check those out. Here are the available plans in your area, R\$35 5 credits, R\$99 15 credits, R\$189 30 credits'</p>
<p>Original</p> <p>Agent: R\$299 #NUMBER# credits</p> <p>ChatGPT Rephrasing</p> <p>Agent: I understand that the prices may seem high, but unfortunately, those are the standard rates for our available plans in your area. Let me know if you have any other questions or concerns and I'll do my best to assist you.</p>

Table 5.7: Example of a rephrasing that changes the meaning of the original utterance.

Furthermore, one additional problem of ChatGPT is that it tends to repeat a lot of the context, which

in a real-life customer service situation is not adequate to solve the problem instead being a reason for the client to get frustrated due to the unnatural conversational flow. This context repetition is clearly visible in Table 5.8.

<p>Context:</p> <p>Client: still not opening</p> <p>Agent: Please try this steps. Go to your Home screen. Tap the More icon at the bottom of the screen. Tap Settings.</p>
<p>Original</p> <p>Agent: Tap Device information.</p> <p>ChatGPT Rephrasing</p> <p>Agent: I understand that you're having trouble opening the app. Let's walk through some steps together. Can you please go to your Home screen and tap the More icon at the bottom of the screen? Then, tap Settings and after that, please tap on Device information.</p>

Table 5.8: Example of a rephrasing that repeats the context. The client's source language is Brazilian Portuguese.

Despite the downsides of using ChatGPT to empathetically rephrase utterances, the manual evaluation shows that it can be useful in most situations, enlightening more adequate and empathetic responses. This finding holds valuable implications as it can serve as a meaningful complement to an agent's capabilities. This enables him/her to effectively filter the improved responses and discard the irrelevant ones, enhancing its overall performance in addressing clients needs.

5.5 Consecutive response grouping approach

Given the lack of meaning preservation of ChatGPT in the previous experiences together with the long repetitions of the context, we have made a subtle tweak to the MAIA dataset. This dataset contains dialogues from real-world interactions between agents and clients in continuous conversations. Occasionally, a participant may provide multiple consecutive responses before receiving a reply. To maintain coherence, we group such consecutive responses from the same participant, aiding in the handling of natural conversation dynamics. This grouping is better fitted to the generation abilities of LLMs and has the potential to considerably improve the weak points of ChatGPT's previous performance. Nonetheless, an important consideration arises when consolidating consecutive utterances from the same source: the need 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 adjustment results in a modified emotion distribution, as illustrated in Table 5.9.

Similarly to the previous approach, we still use our ERC model to understand the utterances that

Table 5.9: Emotion distribution per language pair.

Emotion	Portuguese-English	German-English
Empathy	33	229
Happiness	55	14
Neutral	571	659
Disappointment	41	34
Confusion	70	72
Frustration	40	22
Anger	15	16
Anxiety	7	185
Total utterances	791	1159

need rephrasing and the rephrasing is done for the negative emotions and the neutral emotion.

5.6 Models

Contrary to the previous approach where we solely use ChatGPT to do rephrasing, here we use 2 different models, ChatGPT [Brown et al., 2020] and Vicuna [Chiang et al., 2023]. Whereas ChatGPT is an LLM which is thought to have 175 billion parameters, Vicuna is an LLM with a considerably lower number of parameters, in our case, we used the one with 33 billion parameters. Vicuna is an open-source chatbot trained by fine-tuning LLaMA [Touvron et al., 2023], a popular open-source model, on user-shared conversations. The objective of using 2 different sized LLMs is to understand the influence of the size of the model on the quality of the rephrased text, therefore analyzing if a smaller model could have a comparable performance with ChatGPT.

Given that we use two different models it is important to distinguish the prompts used for each one of them. In the case of ChatGPT, we were able to adapt the prompt from the previous approach, enhancing it with elements of politeness, which proved beneficial for improving the original text. However, when working with the Vicuna model, we needed to construct a suitable prompt from scratch. Considering the experiences done using different prompts for Vicuna and the size of the model, it became evident that the prompt should be simpler in order to produce good results, since as the prompt size increased, Vicuna encountered greater challenges in comprehending the task. The prompt can be seen below:

- *Context:* [\[Previous Dialogue\]](#)
Sentence: *Agent:* [\[Message to empathetically rephrase\]](#).
Rephrase this last utterance in English to be more empathetic and polite, using only and exclusively the content of the utterance.

where the blue text represents the input variables that vary depending on the utterance to be rephrased. [\[Previous Dialogue\]](#) 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.

5.7 Evaluation

With the introduction of the two models, it is crucial to evaluate and compare the performance of the different-sized models. Therefore, similar to our previous analysis, we conduct both automatic and manual evaluations.

5.7.1 Automatic Evaluation

In order to do the automatic evaluation, we use the ERC model to examine if there is an increase in the number of empathetic responses of the agent. We do this comparison between the original data, the ChatGPT rephrased data and the Vicuna rephrased. The confusion matrices of the original text, the ChatGPT rephrased text and the Vicuna rephrased text are presented in Tables 5.5, 5.6 and 5.7, respectively.

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	Anx
		Predicted Class							

Figure 5.5: Confusion Matrix before rephrasing

In Figures 5.5, 5.6 and 5.7, we observe that similar to our previous approach, the original text contains a higher proportion of correctly labelled utterances, which comes as no surprise since the utterances are labelled according to the original text. Surprisingly, 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 since the majority of the empathetic examples consist of brief utterances. Furthermore, the precision of the labels could have been compromised during our process of identifying the underlying emotions in the utterances, resulting in an inadequately labelled dataset that could potentially lead to a less effective model. One example of possible incorrect labels can be seen in Table 5.10, where no visible signs of frustration can be seen

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								

Figure 5.6: 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								

Figure 5.7: Confusion Matrix after rephrasing using Vicuna

despite being the utterance label.

5.7.2 Human Evaluation

Recognizing that relying solely on automatic evaluations would make it even harder to assess the level of empathy and text quality, given the novel dataset modification approach, it is essential to perform human evaluations.

Contrary to the previous human evaluation, we changed the approach used to evaluate the quality of the text generated by the LLMs. We evaluated the models using 3 different criteria: Empathy, Relevance, and Fluency, each one of them individually evaluated. Empathy measures the understanding and expression degree of context emotion the response demonstrates, Relevance evaluates whether the responses are on-topic with the dialogue history and Fluency measures the grammatical correct-

Context:

Agent: well lets try a manual reset! going to share the steps

Client: ok!Yes if possible

Agent: Connect your eReader to a power source by doing one of the following:Turn on your computer and connect the included USB power cord to your computer first, and then your eReader. Plug the wall charger (not included) to a power outlet, and then connect your eReader to the wall charger.Press and hold the power button until the power light on the top-right corner of your eReader turns off.You'll see a 'Powered off' screen when your eReader is turned off.Release the power button. Press and hold the power button on your eReader for 30 seconds.Wait for the Restore screen to appear. Release the power button.Your eReader screen will turn black and begin the restoring process.After the restoring process finishes, follow the onscreen instructions to set up your eReader. You'll need access to an active Wi Fi or internet connection to complete setup.You can also use the #PRS.ORG# app to set up your eReader using your computer.Could you get the manual factory reset?

Emotion: Frustration

Table 5.10: Example of an utterance labelled with frustration where it is not visible in the utterance. The client's source language is Brazilian Portuguese.

ness and readability of the response. This change was made to be more aligned with practices done in the literature and to have a clearer view of the results obtained since relevance takes into account the "adequate" and meaning preservation categories and we even had a fluency category.

Moreover, we increased the number of human annotators to four instead of only one as used previously. It is worth highlighting that these annotators are not experts compared to the one who annotated the previous approach. To measure inter-annotator agreement, we used Fleiss' Kappa inter-annotator agreement [Fleiss, 1971], which is an extension of Cohen's Kappa agreement [Cohen, 1960] for more than 2 annotators. For the inter-annotator agreement, we obtained a value of 0.43, which can be classified as a moderate agreement according to [Altman, 1990]. For this evaluation, we used 100 different examples randomly pooled from the test set that were rephrased by our models.

The results of the human evaluation can be seen in Tables 5.11, 5.12 and 5.13 and they are a pairwise comparison between every 2 models in terms of Empathy, Relevance, and Fluency. The responses were randomly ordered from the 2 models and the original answer to remove possible bias from the annotators. Annotators could either choose one of the responses or select a tie when the provided options are either both good or both poor in the criteria being evaluated.

As can be seen in Table 5.11, both ChatGPT and Vicuna show an improvement in relevance when compared to the original text, with ChatGPT providing a stronger improvement than Vicuna. Nonethe-

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

Table 5.11: Results of human evaluation for utterances' relevance.

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

Table 5.12: Results of human evaluation for utterances' fluency.

less, the improvement is not significant and some of the answers that show less relevance can lose valuable information from the original text, so it is difficult to clearly consider the models as better relevance-wise.

Moreover, Tables 5.12 and 5.13 clearly show a high improvement from both Vicuna and ChatGPT in terms of fluency and empathy when compared to the original text. ChatGPT clearly outstands with a performance much superior to the one of Vicuna, both on fluency (71% vs 10%) and empathy (75% vs 11%).

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

Starting first with an examination of 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 of this behaviour is presented in Table 5.14, where the Vicuna model only corrects a misspelt word.

Additionally, according to some examples, Vicuna may exhibit a lack of comprehension regarding the task at hand and for example, translate the utterance of the client. One such example can be seen in Table 5.15.

Similarly, on certain occasions, rather than rephrasing the initial text, it answers to the utterance to be rephrased, as can be seen in Table 5.16.

As far as ChatGPT, the characteristic that stands out the most is the repetition of previous context leading to long sentences which are not practical nor natural in a customer service conversation. One such example is presented in Table 5.17, where ChatGPT repeats all the context of the conversation.

Moreover, ChatGPT exhibits a distinctive response style that may lead to clients' frustration, characterized by overly repetitive interactions that are not natural in a conversation. One example can be seen in Table 5.18, where the response repeatedly starts with the wording "I understand".

On the other side, ChatGPT approaches the clients carefully, effectively preventing those with negative emotions from further escalating their negativity. A notable example can be seen in Table 5.19

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

Table 5.13: Results of human evaluation for utterances' empathy.

Context:

Client :#PRS_ORG#

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

Client: Yes if possible

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 5.14: Example of a rephrasing in that Vicuna only corrects the original response. The client's source language is Brazilian Portuguese.

A clear illustration of the enhancements achieved by both Vicuna and ChatGPT can be found in Table 5.20. 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 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. Moreover, we see that LLMs deliver better performance when the dataset is adequate to the natural flow of a conversation, improving the results in the empathy raise.

Context:
Client: Good afternoon, I would like to know the activities in Campinas SP

Original
Agent: Hi!
Vicuna Rephrasing
Agent: Hi! I'd like to know the activities in Campinas, SP.

Table 5.15: Example of a rephrasing that translates the client utterance. The client's source language is Brazilian Portuguese.

Context:
Client: What I sent earlier was...##PHONENUMBER#-01 this is the purchase I paid for with a credit card, it has already been billed, and it shows as shipped on the #PRS.ORG# website!.
Agent: ah ok sorry. do you have another email address?
Client: I just don't understand why they haven't released it yet... another address? But is my email #PRS.ORG# the one I gave you?!

Original
Agent: do you have the invoice ?
Vicuna Rephrasing
Agent: Yes, I have the invoice.

Table 5.16: Example of a rephrasing that answers to the agent's original response. The client's source language is Brazilian Portuguese.

Context:
Client : Hello, I would like to transfer my books from the #PRS_ORG# account to #EMAIL# and remove that first email from my account.

Original
Agent: Hi, thanks for the information.I will be happy to assist you.
Vicuna Rephrasing
Agent: Hello, I understand that you would like to transfer your books from your current account (#PRS_ORG#) to a new email address (#EMAIL#) and remove the first email from your account. I will gladly help you with that.

Table 5.17: Example of a rephrasing that repeats the context. The client's source language is Brazilian Portuguese.

Client: At the moment, I don't know when I bought them; approximately 3 to 5 years ago.

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: Yes

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: I'm waiting. Thank you. Good afternoon. The message says it's empty.

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

Client: I downloaded it again, but it still has the same message.

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

Table 5.18: Example of a rephrasing in that ChatGPT repeats the structure for all answers. The client's source language is Brazilian Portuguese.

Context:

Client : I see, do you have an account? There's a 'Sign In' option, and above, there's the help button.

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

Client: I can't find the link

Original

Agent: Would be possible to see a screenshot ?

ChatGPT Rephrasing

Agent: Could you please provide a screenshot?

Table 5.19: Example of a rephrasing that enhances the carefulness used by ChatGPT. The client's source language is Brazilian Portuguese.

Context:

Client : It still doesn't open, but I'll send the information that appears to me.

Agent: Ok, please send it

Client: Visitor uploaded: information-#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 5.20: 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 Brazilian Portuguese.

6

Conclusion

Contents

6.1 Conclusions and Future Work	70
---	----

6.1 Conclusions and Future Work

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

Regarding the ERC module, we employed an LLM, like a RoBERTa-based model, to understand its capabilities within a customer service context. Our study revealed that incorporating context turns as input to the pre-trained language model led to a notable enhancement in performance, with the model achieving the best performance with 4 context turns. Notably, the model exhibited robust performance in identifying empathy, signifying its potential as an automatic empathy classifier.

Furthermore, we undertook an examination of a larger Language Model's performance in the ERC task, without subjecting it to fine-tuning specific to the task, relying solely on prompt engineering. This analysis highlighted the importance of the training process itself, even when employing very large models.

As far as the Empathy Rephrasing system is concerned, we successfully built a system capable of enhancing the empathy of the agents' responses in a customer service environment, supported by empathy's manual evaluation, since the automatic evaluation proved to be insufficient to evaluate this emotion, reinforcing the usefulness of developing a good predictor. Additionally to the empathy critical point, we were also able to improve significantly the relevance and fluency of agents' responses.

Future work and limitations

Concerning the ERC module, opportunities for advancement exist. While we have employed RoBERTa models, there is potential for further exploration. This may include the use of larger fine-tuned models, investigating multi-modal recognition and incorporating advanced models like GPT-4, which would necessitate the expansion of the dataset to include image representations, and optimizing loss functions to enhance emotion comprehension.

For future validation, it would be beneficial to expand our methods by incorporating a dataset encompassing additional language pairs. Furthermore, an alternative approach worth exploring involves fine-tuning a pre-trained model specifically for the Empathetic Rephrasing task, as opposed to relying solely on prompting an LLM. Finally, it would also be possible to have the adjusted dataset to be labelled again in order for the labelled emotions to be more precise.

There are two main limitations to our work that we would like to discuss:

- One limitation of our work is how we define the empathy label. Currently, the utterance in question is categorized as either empathetic or non-empathetic, with no intermediate classification being considered, contrarily to [Buechel et al., 2018], where a 7-point scale is used. Incorporating a

scale to measure the varying degrees of empathy in responses would be highly valuable for understanding the level of empathy conveyed. It would enable the identification of responses that lack empathy significantly, prompting the need for rephrasing to ensure their appropriateness in addressing client concerns effectively

- The same applies to the remaining emotions. It would be useful for the agent to easily flag the clients who are displaying strong negative emotions so they can be even more careful addressing such clients' concerns. Nonetheless, this would pose a significant challenge since it would increase the ERC task difficulty.

Bibliography

- [Alm et al., 2005] Alm, C. O., Roth, D., and Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 579–586.
- [Altman, 1990] Altman, D. G. (1990). *Practical statistics for medical research*. CRC press.
- [Amjad et al., 2023] Amjad, B., Zeeshan, M., and Beg, M. O. (2023). Emp-eval: A framework for measuring empathy in open domain dialogues.
- [Ba et al., 2016] Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.
- [Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- [Balhara et al., 2022] Balhara, S., Gupta, N., Alkhayat, A., Bharti, I., Malik, R. Q., Mahmood, S. N., and Abedi, F. (2022). A survey on deep reinforcement learning architectures, applications and emerging trends. *IET Communications*.
- [Barriere et al., 2022] Barriere, V., Tafreshi, S., Sedoc, J., and Alqahtani, S. (2022). 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, Dublin, Ireland. Association for Computational Linguistics.
- [Beqari, 2018] Beqari, E. (2018). A Very Basic Introduction to Feed-Forward Neural Networks. *Dzone.com*. <https://dzone.com/articles/the-very-basic-introduction-to-feed-forward-neural>.
- [Bitner et al., 2002] Bitner, M. J., Ostrom, A. L., and Meuter, M. L. (2002). Implementing successful self-service technologies. *Academy of management perspectives*, 16(4):96–108.
- [Brown, 2015] Brown, P. (2015). Politeness and language. In *The International Encyclopedia of the Social and Behavioural Sciences (IESBS), (2nd ed.)*, pages 326–330. Elsevier.

- [Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *CoRR*, abs/2005.14165.
- [Buechel et al., 2018] Buechel, S., Buffone, A., Slaff, B., Ungar, L. H., and Sedoc, J. (2018). Modeling empathy and distress in reaction to news stories. *CoRR*, abs/1808.10399.
- [Busso et al., 2008] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.
- [Chiang et al., 2023] Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.
- [Cho et al., 2014a] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014a). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- [Cho et al., 2014b] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014b). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- [Clark et al., 2013] Clark, C. M., Murfett, U. M., Rogers, P. S., and Ang, S. (2013). Is empathy effective for customer service? evidence from call center interactions. *Journal of Business and Technical Communication*, 27(2):123–153.
- [Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- [Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- [Dias et al., 2022] Dias, I., Rei, R., Pereira, P., and Coheur, L. (2022). Towards a sentiment-aware conversational agent. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–3.
- [dprogrammer, 2020] dprogrammer, A. (2020). Rnn, lstm amp; gru.
- [Ekman, 1999] Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

- [Elman, 1990] Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179–211. <https://www.sciencedirect.com/science/article/pii/036402139090002E>.
- [Farinha et al., 2022] Farinha, A. C., Farajian, M. A., Buchicchio, M., Fernandes, P., de Souza, J. G. C., Moniz, H., and Martins, A. F. T. (2022). Findings of the WMT 2022 Shared Task on Chat Translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT22)*, Abu Dhabi. Association for Computational Linguistics.
- [Feng et al., 2020] Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.
- [Feng et al., 2017] Feng, W., Guan, N., Li, Y., Zhang, X., and Luo, Z. (2017). Audio visual speech recognition with multimodal recurrent neural networks. In *2017 International Joint Conference on neural networks (IJCNN)*, pages 681–688. IEEE.
- [Fleiss, 1971] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- [Ghosal et al., 2019] Ghosal, D., Majumder, N., Poria, S., Chhaya, N., and Gelbukh, A. (2019). Dialoguecn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.
- [Gonçalves et al., 2022] Gonçalves, M., Buchicchio, M., Stewart, C., Moniz, H., and Lavie, A. (2022). Agent and User-Generated Content and its Impact on Customer Support MT. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 199–208.
- [Guibon et al., 2021] Guibon, G., Labeau, M., Flamein, H., Lefeuvre, L., and Clavel, C. (2021). Few-shot emotion recognition in conversation with sequential prototypical networks. *arXiv preprint arXiv:2109.09366*.
- [Hanif et al., 2010] Hanif, M., Hafeez, S., and Riaz, A. (2010). Factors affecting customer satisfaction. *International research journal of finance and economics*, 60(1):44–52.
- [He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- [Hochreiter, 1998] Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116.

- [Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- [Jones et al., 1995] Jones, T. O., Sasser, W. E., et al. (1995). Why satisfied customers defect. *Harvard business review*, 73(6):88.
- [Lafferty et al., 2001] Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
- [Lee et al., 2010] Lee, S. Y. M., Chen, Y., and Huang, C.-R. (2010). A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 45–53.
- [Li et al., 2022] Li, Q., Li, P., Ren, Z., Ren, P., and Chen, Z. (2022). Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10993–11001.
- [Li et al., 2017] Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.
- [Liang et al., 2022a] Liang, Y., Meng, F., Xu, J., Chen, Y., and Zhou, J. (2022a). Msctd: A multimodal sentiment chat translation dataset. *arXiv preprint arXiv:2202.13645*.
- [Liang et al., 2022b] Liang, Y., Meng, F., Zhang, Y., Chen, Y., Xu, J., and Zhou, J. (2022b). Emotional conversation generation with heterogeneous graph neural network. *Artificial Intelligence*, 308:103714.
- [Lin et al., 2019] Lin, Z., Madotto, A., Shin, J., Xu, P., and Fung, P. (2019). Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*.
- [Liu, 2012] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.
- [Liu et al., 2022] Liu, Y., Maier, W., Minker, W., and Ultes, S. (2022). 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.
- [Liu et al., 2019] 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*.

- [Ma et al., 2005] Ma, C., Prendinger, H., and Ishizuka, M. (2005). Emotion estimation and reasoning based on affective textual interaction. In *International conference on affective computing and intelligent interaction*, pages 622–628. Springer.
- [Majumder et al., 2020] Majumder, N., Hong, P., Peng, S., Lu, J., Ghosal, D., Gelbukh, A., Mihalcea, R., and Poria, S. (2020). Mime: Mimicking emotions for empathetic response generation. *arXiv preprint arXiv:2010.01454*.
- [Majumder et al., 2019] Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., and Cambria, E. (2019). Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.
- [Mehrabian and O'Reilly, 1980] Mehrabian, A. and O'Reilly, E. (1980). Analysis of personality measures in terms of basic dimensions of temperament. *Journal of Personality and Social Psychology*, 38(3):492.
- [Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- [Oliver, 2014] Oliver, R. L. (2014). *Satisfaction: A behavioral perspective on the consumer*. Routledge.
- [Pan and Yang, 2010] Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.
- [Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- [Pereira et al., 2022] Pereira, P., Moniz, H., and Carvalho, J. P. (2022). Deep Emotion Recognition in Textual Conversations: A Survey. *arXiv preprint arXiv:2211.09172*.
- [Pipis, 2020] Pipis, G. (2020). Character level text generation. <https://predictivehacks.com/character-based-text-generation/>.
- [Poria et al., 2018] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.
- [PS and Mahalakshmi, 2017] PS, S. and Mahalakshmi, G. (2017). Emotion models: a review. *International Journal of Control Theory and Applications*, 10(8):651–657.

- [Qian et al., 2023a] Qian, Y., Wang, B., Lin, T.-E., Zheng, Y., Zhu, Y., Zhao, D., Hou, Y., Wu, Y., and Li, Y. (2023a). 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.
- [Qian et al., 2023b] Qian, Y., Wang, B., Ma, S., Bin, W., Zhang, S., Zhao, D., Huang, K., and Hou, Y. (2023b). Think twice: A human-like two-stage conversational agent for emotional response generation. *arXiv preprint arXiv:2301.04907*.
- [Radford et al., 2018] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.
- [Raffel et al., 2019] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.
- [Rajaobelina et al., 2021] Rajaobelina, L., Brun, I., Kilani, N., and Ricard, L. (2021). Examining emotions linked to live chat services: The role of e-service quality and impact on word of mouth. *Journal of Financial Services Marketing*, pages 1–18.
- [Rashkin et al., 2018] Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L. (2018). Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.
- [Roumeliotis and Tselikas, 2023] Roumeliotis, K. I. and Tselikas, N. D. (2023). Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192.
- [Saha and Ananiadou, 2022] Saha, T. and Ananiadou, S. (2022). 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.
- [Schuller et al., 2012] Schuller, B., Valster, M., Eyben, F., Cowie, R., and Pantic, M. (2012). Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456.
- [Shen et al., 2021] Shen, W., Wu, S., Yang, Y., and Quan, X. (2021). Directed acyclic graph network for conversational emotion recognition. *arXiv preprint arXiv:2105.12907*.
- [Snell et al., 2017] Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.
- [Su et al., 2023] Su, Y., Bian, H., Fan, B., Lian, B., Zhang, C., Zhang, B., and Huang, R. (2023). Rlca: Reinforcement learning model integrating cognition and affection for empathetic response generation. *IEEE Transactions on Computational Social Systems*.

- [Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.
- [Thost and Chen, 2021] Thost, V. and Chen, J. (2021). Directed acyclic graph neural networks. *arXiv preprint arXiv:2101.07965*.
- [Touvron et al., 2023] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models.
- [Vaswani et al., 2017] 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.
- [Wang et al., 2020] Wang, Y., Zhang, J., Ma, J., Wang, S., and Xiao, J. (2020). 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.
- [Willcox, 1982] Willcox, G. (1982). The feeling wheel: A tool for expanding awareness of emotions and increasing spontaneity and intimacy. *Transactional Analysis Journal*, 12(4):274–276.
- [Wolf et al., 2019] Wolf, T., Sanh, V., Chaumond, J., and Delangue, C. (2019). Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.
- [Xu et al., 2022] Xu, Y., Cui, Y., Jiang, X., Yin, Y., Ding, J., Li, L., and Wu, D. (2022). Inconsistency-Based Multi-Task Cooperative Learning for Emotion Recognition. *IEEE Transactions on Affective Computing*, 13(4):2017–2027.
- [Yang et al., 2023a] Yang, K., Ji, S., Zhang, T., Xie, Q., and Ananiadou, S. (2023a). On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *arXiv preprint arXiv:2304.03347*.
- [Yang et al., 2023b] Yang, K., Ji, S., Zhang, T., Xie, Q., and Ananiadou, S. (2023b). On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *arXiv preprint arXiv:2304.03347*.
- [Zahiri and Choi, 2018] Zahiri, S. M. and Choi, J. D. (2018). Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aai conference on artificial intelligence*.
- [Zaranis et al., 2021] Zaranis, E., Paraskevopoulos, G., Katsamanis, A., and Potamianos, A. (2021). Empbot: a t5-based empathetic chatbot focusing on sentiments. *arXiv preprint arXiv:2111.00310*.
- [Zhao et al., 2023] Zhao, W., Zhao, Y., Lu, X., Wang, S., Tong, Y., and Qin, B. (2023). Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

[Zhong et al., 2019] Zhong, P., Wang, D., and Miao, C. (2019). Knowledge-enriched transformer for emotion detection in textual conversations. *arXiv preprint arXiv:1909.10681*.