A Framework for Context-Aware Multilingual Emotion Recognition with Transformers

Raquel Peña Alarcón[1]

MSc Data Science and Machine Learning

Department of Computer Science
University College London

Prof. Philip Treleaven
Dr. Omer Gunes, DeepZen

September 2024

[1] **Disclaimer:** This report is submitted as part requirement for the MSc Data Science and Machine Learning at UCL. It is substantially the result of my own work except where explicitly indicated in the text. The report will be distributed to the internal and external examiners, but thereafter may not be copied or distributed except with permission from the author.

Abstract

This thesis develops an innovative multilingual framework using advanced transformer models to redefine emotion recognition in dialogues, with a specific focus on Spanish-language contexts. Emotion recognition is the capability to accurately identify human emotions across various sources and modalities. Addressing the challenges of context-awareness and the scarcity of high-quality Spanish datasets, this investigation employs deep learning techniques to translate text into Spanish and introduces a transformer model that effectively captures current and previous contexts. This original research also integrates the GPT-4o API with in-context learning to improve dialogue comprehension.

Conducted in collaboration with DeepZen, the research is structured around three key experiments that underpin the development and validation of this framework:

Exp. 1: Multilingual Framework. This experiment presents a comprehensive multilingual framework for emotion recognition in dialogues. The framework consists of five stages, with the first two detailed in this chapter. The approach begins with data selection, followed by the translation of datasets from English to Spanish, qualitatively comparing various models, including automatic translators, neural machine translation systems, and long-sequence transformers, all evaluated in terms of linguistic naturalness, particularly for emotion recognition.

Exp. 2: Impact of Context in Emotion Recognition Conversations. The remaining three stages of the framework are developed in this experiment, beginning with an utterance-level encoder integrated with a transformer trained on Spanish corpora. This study explores the significance of context in emotion recognition within dialogues, focusing on how historical context and speaker identity, both key features of the datasets, influence classification accuracy. The analysis of emotion label transitions shows that emotions often persist or shift predictably throughout conversations, highlighting the importance of incorporating contextual information into emotion recognition models.

Exp. 3: Assessment of Pre-trained Language Model in Emotion Recognition. This experiment evaluates the effectiveness of in-context learning for emotion recognition, employing pre-trained large language models such as GPT-4o. The study builds on the translation of datasets from English to Spanish using DeepL, allowing the direct integration of translated examples into the prompts. The focus is on how these models, by utilising the context provided by these examples, can improve emotion classification in text.

A scientific paper based on this research will be submitted.

The code of this framework can be found at
https://github.com/kawaiiblitz/Context-Aware-Multi-Lingual-ERC.

Contributions to Science

The thesis presents the following original contributions to sciences:

- **Multilingual Emotion Recognition Techniques:** This study applies deep learning methods to translate text into Spanish within a comprehensive AI-driven pipeline, facilitating emotion recognition across multiple languages.

- **Evaluation of Transformers Architectures:** This work analyses and evaluates several multilingual, transformer-based models for emotion recognition based on textual features, comparing their performance across different datasets and exploring the impact of fine-tuning and context-awareness integration.

- **Incorporation of Historical and Speaker Context in Dialogue Datasets:** For context-rich datasets such as IEMOCAP and MELD, this research leverages historical context and speaker-specific context to enhance emotion recognition accuracy, focusing on the flow of conversation and individual speaking styles.

- **In-Context Learning Framework:** This study develops a context-aware prompt engineering framework that integrates in-context learning within the GPT-4o API. The framework incorporates few-shot learning techniques, using a varying number of examples in the prompts to evaluate performance. Additionally, it explores the impact of different window sizes for historical context, testing how the inclusion of more or fewer previous utterances affects the model's ability to accurately classify emotions across conversations.

Impact Statement

- **Enhancement of Multilingual Emotion Synthesis Technologies:** The integration of sophisticated context-aware mechanisms and the translation of emotional context into different languages significantly improves the fidelity and naturalness of synthesised emotional speech. Consequently, *DeepZen* is positioned to deliver unmatched multilingual emotion synthesis capabilities, enabling more engaging and relatable voice interactions. This advancement allows the company to excel in applications such as multilingual customer service, global media production, and personalised interactive experiences.

- **Reduction of Production Costs and Time:** The automation of emotional recognition and expression reduces the need for extensive human intervention in the audio production process. This advancement significantly cuts down the labour and technical costs associated with post-production and manual tuning of vocal inflections. By streamlining these processes, the technology allows for quicker production timelines and lowers overall production costs, making high-quality emotional audio content more accessible and scalable. Additionally, the use of pre-trained language models further accelerates development by reducing the time required to design task-specific AI architectures, as these models can be adapted for emotion recognition with minimal additional training.

- **Market Expansion and Penalisation:** The capability to adjust emotional responses in voice narration precisely allows for the creation of highly personalised audio products. These can be tailored for different cultural contexts, languages, and individual user needs, expanding the market reach globally. Particularly, this technology positions *DeepZen* to lead in sectors requiring high emotional and contextual sensitivity, such as mental health, education, and interactive entertainment, offering content that is nuanced and tailored to specific contexts.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The objective of this chapter is to provide an overview of this thesis by discussing the motivation behind the research problem, the objectives, the three experiments, the contributions of this study, and the structure of this thesis. The chapter begins with an introduction to the significance of emotion recognition and the role of advanced NLP architectures, specifically transformers, in enhancing this field. The chapter then outlines the primary objectives, describes the three experiments conducted, and summarises the key contributions of the study. It concludes with a detailed overview of the thesis structure, setting the stage for the subsequent chapters.

Given the recent advancements in transformer models, this thesis presents an original framework in advanced multilingual machine learning models to evaluate the accuracy of emotion recognition, focusing on Spanish translation, large language models (LLMs), and pre-trained transformers within an in-context learning paradigm. Emotion recognition aims to classify human emotions, playing a critical role in fields such as human-computer interaction to enhance user experience. Traditional methods often struggle with context-awareness in dialogues, and there is a lack of high-quality Spanish datasets. This research uses deep learning to translate text into Spanish and develops a transformer model that captures both historical and future context. The evaluation also incorporates the GPT-4o API with in-context learning to improve dialogue understanding.

In collaboration with DeepZen, this research introduces an innovative multilingual framework designed to predict human emotions in dialogues. This framework was assessed against existing models across two datasets, demonstrating a weighted F1-score of 60% across emotion categories in the datasets.

1.1 Motivation

In an increasingly globalised world, multilingualism has become more relevant than ever (Mim, 2023). Accurate emotion recognition is essential for enhancing human-computer interaction and fostering more empathetic, effective AI systems. This thesis is driven by the need to develop a robust framework for emotion recognition that capitalises on advanced AI technologies to improve both accuracy and contextual understanding, particularly within multilingual environments.

In the field of AI, emotion recognition enables empathetic and effective responses based on the user's emotional state, making it crucial in various applications. Despite advancements driven by deep learning models, emotion recognition remains an active research area due to limitations such as a predominant focus on text-only inputs and challenges in capturing speaker context. These challenges are particularly pronounced in languages like Spanish, where high-quality datasets are scarce. Incorporating historical context from conversations into AI frameworks is essential to better capture linguistic subtleties and improve the accuracy of emotion classification. This is especially relevant, as research such as Barrett (2017) highlights that emotions are not merely internal states, but are shaped by external factors, including the broader context.

A significant portion of emotion recognition research in Spanish relies on the Spanish Emotion Lexicon (SEL), a resource specifically developed for classifying emotions (Rangel, Sidorov, and Guerra, 2014). However, the limitations of this lexicon have led researchers to explore alternative methods, such as transferring emotion recognition models from English to Spanish, in order to bridge resource gaps and improve accuracy (Sidorov et al., 2013). These challenges highlight the need for more robust tools and frameworks to enhance emotion recognition in Spanish-language contexts.

The scarcity of high-quality emotion recognition datasets in Spanish has traditionally driven the need for creating and annotating new corpora (Garcia-Cuesta, Salvador, and Pãez, 2024). However, this research recognises that building new datasets is not always the most efficient or necessary approach (Deng and Ren, 2023). Instead, this study explores the potential of leveraging existing, robust multi-modal datasets originally developed in English, using advanced translation tools. By doing so, it addresses the limitations in Spanish-language resources while capitalising on the extensive annotation and validation already present in these datasets.

Furthermore, with the advent of large language models (LLMs) and pre-trained transformers, there exists a significant opportunity to enhance emotion recognition systems, particularly those tailored for the Spanish language. These models, renowned for their ability to capture semantic contexts, provide a powerful tool for improving the accuracy and depth of emotion recognition in multilingual settings. By fine-tuning models specifically trained on Spanish-language corpora, this research seeks to advance the precision and effectiveness of emotion recognition in Spanish dialogues, thereby surpassing the limitations of traditional methods.

Since the introduction of ChatGPT in late 2022, generative AI models have demonstrated significant potential in text classification, including emotion recognition. A key challenge in this area is understanding how providing context or examples within the prompt can enhance the performance of these AI models. This study addresses this challenge by evaluating the performance of GPT-4 for emotion classification when given contextual prompts and examples.

The pressing need to develop multilingual emotion recognition systems is evident. By utilising deep learning services to translate text into Spanish and other languages, this research aims to create a framework that is both multilingual and capable of understanding context. The findings of this study will contribute to the development of more effective AI systems, enhancing human-machine interactions across various applications.

1.2 Research Objectives

The primary objective of this thesis is to advance the field of emotion recognition by developing a robust and multilingual framework capable of accurately capturing and classifying human emotions in dialogues, with a particular emphasis on Spanish-language contexts. This investigation seeks to address the limitations of existing approaches through the following specific objectives:

First, this work focuses on the generation of multilingual emotion datasets. Given the scarcity of comprehensive emotion datasets in Spanish, the research will leverage advanced AI tools to translate and adapt the complexities inherent in English-language datasets, particularly those prominent in Speech Emotion Recognition (SER). The goal is to develop Spanish-language datasets that accurately reflect natural emotional expressions, thereby enhancing the precision and cultural sensitivity of emotion recognition systems.

Second, this thesis presents an innovative approach to context-aware Spanish Transformer architecture. It involves the development and fine-tuning of a Transformer-based model, utilising RoBERTa-bne for generating embeddings. The novelty of this study lies in the integration of a GRU that processes multiple contexts within feed-forward networks. This approach allows for more effective use of contextual information in dialogues, enhancing the accuracy of emotion classification and providing a deeper understanding of language in multilingual settings.

Finally, the research examines the role of advanced language models in processing multiple dialogues within the same conversation, as well as the use of in-context learning. It focuses on the application of few-shot learning techniques, integrated into the prompts of pre-trained models such as GPT-4o, to evaluate how these architectures, optimised for contextual understanding, improve the accuracy of emotion classification in Spanish-language environments.

1.3 Research Experiments

This thesis commences with a thorough examination of the challenges associated with emotion recognition in multilingual contexts, progressively refining its focus in subsequent experiments to address specific technical dimensions, including dataset translation, model architecture, and contextual awareness. The research is organised into three empirical studies:

1. **Multilingual Framework:** The first experiment introduces the initial stages of the multilingual framework for classifying emotions in dialogues. This study begins with an empirical evaluation of the translation quality provided by three distinct tools: Google Translate, a transformer-based sequence model, and an AI-driven online tool. These tools are applied to translate the most prominent English-language datasets in emotion recognition, namely IEMOCAP and MELD, into Spanish. As part of the evaluation, an analysis of these datasets in terms of category distribution was conducted, an aspect that will also be of significant importance in subsequent chapters. The results offer valuable insights into the effectiveness of these tools in handling colloquial expressions and highlight the potential for integrating

culturally specific transformer models into the framework.

2. **Impact of Context in Emotion Recognition Conversations:** The second experiment advances the subsequent stages of the multilingual framework for Spanish-translated dialogues. It applies the methodology of employing an utterance-level encoder using RoBERTa trained on Spanish corpora and compares its performance to the English-language transformer version, in order to demonstrate that language-specific transformers tend to yield better results. Contextual information from the previous step is incorporated as a critical component within the feed-forward network to classify emotions across the selected datasets. The focus is on evaluating how the integration of this specific context improves accuracy in emotion classification compared to a baseline model without contextual information.

3. **Assessment of Pre-trained Language Model in Emotion Recognition:** This experiment further explores the importance of context in emotion recognition within conversations, but from the perspective of pre-trained language models such as ChatGPT-4o. Ablation studies are conducted on the translated datasets to investigate the effects of incorporating the in-context learning and the inclusion of historical utterances in the prompt engineering of these APIs. The goal is to measure the impact of these strategies on emotion classification accuracy within dialogues.

1.4 Scientific Contributions

This research contributes to the existing literature in the following ways:

1. **Comprehensive Pipeline for Emotion Detection in Multilingual Contexts:** The existing literature has predominantly focused on isolated stages within an emotion classification framework, such as the creation of datasets evaluated on pre-existing architectures or, in some cases, the development of specific architectures trained on language-specific corpora. This research enriches the literature by contributing to the limited number of Spanish datasets while also fine-tuning an architecture that integrates technically relevant aspects, such as context awareness, within the layers of the model. Importantly, all of this is achieved within a comprehensive framework with clearly defined stages. The subsequent points will detail each of these processes.

2. **Extension of Multilingual Emotion Recognition:** Due to the positive performance of the proposed flexible framework, it can be extended to languages beyond Spanish. By leveraging translation tools such as DeepL, which supports more than 30 languages, and selecting or fine-tuning the appropriate transformer for emotion classification, the framework offers a scalable solution for multilingual emotion recognition. Additionally, both the generated translations and pre-trained language models, such as GPT, can be integrated into the process by incorporating translated examples into the prompts through in-context learning techniques, which optimises emotion classification across different languages.

3. **Detailed Exploration of Spanish Datasets for Speech Emotion Recognition:** Unlike previous approaches that have focused on either creating costly datasets or relying on manual human translations and tools like Google Translate, this study investigates dataset translation as a viable alternative to the creation of extensive new datasets. By utilising advanced NLP and AI tools, such as MarianMT and DeepL, the research demonstrates how the translation of Spanish datasets can effectively preserve the speaker's context and ensure the accurate rendering of colloquial expressions.

4. **Incorporation of Historical and Speaker Context for Dialogue Datasets:** While the use of historical and speaker-specific context has been explored in English-language datasets (Wei et al., 2023), this contribution extends these techniques to other languages by leveraging utterances generated by a transformer trained on Spanish corpora. The approach enriches the model by incorporating the history of the last five utterances to improve the accuracy of emotion recognition. This involves a deeper understanding of the conversational flow and the individual speaking style.

5. **In-Context Learning and Historical Context using Pre-trained Language Models:** This study presents a prompt engineering framework that employs strategies such as few-shot prompting to adapt pre-trained models like GPT-4 to emotion recognition tasks. Unlike traditional approaches requiring large amounts of training data and specialised models for each task, this approach enables the models to quickly adapt using only a few examples. Furthermore, this research extends previous work by incorporating historical utterances into the prompts focusing on recent dialogues to generate better context for emotion classification. While these techniques are typically applied to English datasets, this work uniquely explores their effectiveness in Spanish-language data, enhancing the models' performance in emotion recognition.

1.5 Thesis Structure

The structure of this thesis is organised as follows:

- **Chapter 2 – Background and Literature Review.** The relevant literature and the key concepts in the areas of this research are reviewed. The first part of the chapter provides a historical perspective on theories of emotion, starting with foundational ideas and progressing through significant developments in the 20th and 21st centuries. It also categorises emotional models, highlighting their strengths and limitations. The chapter then transitions to advancements in emotion recognition within human-computer interaction, focusing on the progression from traditional approaches to more sophisticated deep learning methods. A dedicated section explores multi-modal emotion recognition (MER) datasets, surveying widely-used resources and their role in advancing research. It concludes by detailing the shift to Transformer-based models, including key LLM families from OpenAI, Meta, and Google, emphasising how these models have enhanced emotion recognition.

- **Chapter 3 – Multilingual Framework.** This chapter presents a comprehensive multilingual framework for emotion recognition in dialogues. The framework consists of five stages, with the first two detailed in this chapter. The approach begins with data selection, followed by the translation of datasets from English to Spanish, comparing various models, including automatic translators, neural machine translation systems, and transformers tailored to emotion recognition. The results provide empirical insights into the handling of colloquial expressions and establish a foundation for exploring the integration of contextual information in emotion classification. The methodology and quantitative outcomes will be discussed in detail in the following chapter.

- **Chapter 4 – Impact of Context in Emotion Recognition Conversations.** This second study examines the use of a specialised utterance-level encoder integrated with a transformer trained on Spanish corpora to evaluate the role of context in emotion recognition. Additionally, it investigates how historical context and speaker identity influence the accuracy of emotion classification, with results evaluated through F1 score, precision, and recall across emotion classes. Building on the framework from the previous chapter, this study leverages the transformer's attention mechanism to incorporate contextual subtleties, enhancing the model's ability to detect emotions more effectively.

- **Chapter 5 – Assessment of Pre-trained Language Model in Emotion Recognition** The third study extends the previous findings by examining further the role of context in emotion recognition within text dialogues through an ablation study. It delves into the use of prompt engineering with the advanced pre-trained language model GPT-4o, modifying the number of in-context examples and the inclusion of historical utterances to assess their impact on model performance. Additionally, the performance of GPT-4o's smaller variant is evaluated to determine its efficacy. These investigations, applied to datasets translated into Spanish by DeepL, utilise the weighted F1 score throughout the experiments.

- **Chapter 6 – Conclusion.** This chapter concludes the thesis by summarising the key findings from the studies conducted, from the proposed multilingual framework to the assessment of pre-trained LLMs. It highlights the main achievements and acknowledges the limitations encountered. The chapter also offers recommendations for future research directions that could build upon the work presented here.

Chapter 2

Background and Literature Review

This chapter presents background information on various emotion theories in psychology and key concepts relevant to this research. The first part of the chapter provides a historical perspective on theories of emotion, starting with foundational ideas and progressing through significant developments in the 20th and 21st centuries. It also categorises emotional models, highlighting their strengths and limitations.This is followed by the transition to advancements focusing on the progression from traditional approaches to more sophisticated deep learning methods. A dedicated section explores multi-modal emotion recognition (MER) datasets, surveying widely-used resources and their role in advancing research. It concludes by detailing the shift to Transformer-based models, including key LLM families from OpenAI, Meta, and Google, emphasising how these models have enhanced emotion recognition.

2.1 Theories of Emotion: A Brief Historical Perspective

As shown in the timeline (see Fig. 2.3), the evolution of psychological theories of emotion begins with Charles Darwin's seminal work *The Expression of the Emotions in Man and Animals* (1872) (Darwin, 1993), which argued that emotions are evolutionary adaptations essential for survival and social behaviour. This idea laid the foundation for many later theories.

In the early 20th century, behaviourism, which dominated American psychology, largely ignored emotions, focusing instead on observable actions. However, by the 1960s, critiques of behaviourism's neglect of inner mental states emerged, and scholars began re-examining the importance of emotions in understanding human behaviour. Arnold (1960) bridged biology and psychology by proposing that emotions are adaptive responses to environmental challenges. Her cognitive appraisal model positioned emotions as crucial mediators between personal evaluations and actions.

Building on Darwin's ideas, Ekman and Friesen (1971) conducted cross-cultural research and identified six basic emotions: happiness, sadness, anger, surprise, fear, and disgust, arguing that these expressions are universally recognisable, providing insight into the biological roots of emotions.

Further expanding on the biological view, Mehrabian (1974) introduced the Pleasure-Arousal-Dominance (PAD) model, incorporating dimensions of emotional response in relation to environmental perception, and linking cognitive appraisal theories like Arnold's to emotional evaluation.

Izard (1977) advanced the Differential Emotion Theory, identifying ten primary emotions present from birth. His work emphasised the role of emotions in motivating behaviour, contrasting with Ekman's focus on expression universality and aligned more with Arnold's interest in the functional role of emotions.

In 1980, Russell (1980) proposed a study organising emotions into a circular structure defined by two dimensions: valence (positive or negative) and arousal (high or low). This model (see Fig. 2.1) offered a more dynamic view of emotions, highlighting their interconnections. Plutchik (1982) in his psycho-evolutionary theory expanded on these ideas by categorising emotions into eight primary pairs, such as joy-sadness and trust-distrust. His *"Wheel of Emotions"* (see Fig. 2.2) emphasises that these emotions are fundamental biological adaptations.

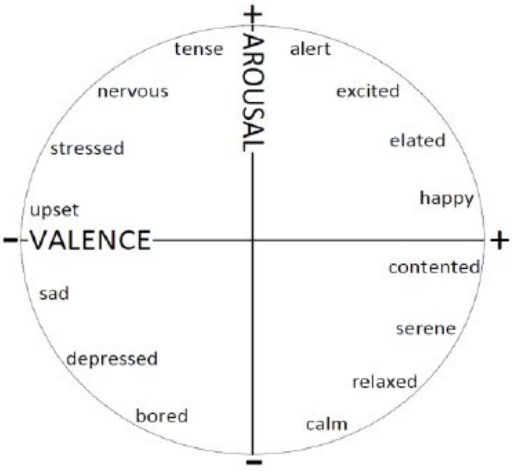


Figure 2.1: *Russell's Circumplex Model of Affect.*

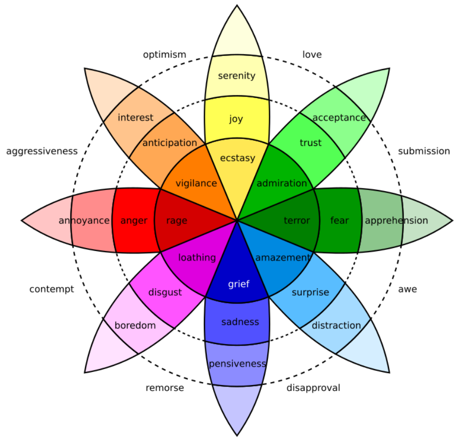


Figure 2.2: *Plutchik's Wheel of Emotions.*

In a departure from these universalist theories, Lazarus (1991) highlighted the role of personal evaluations in emotional experiences. He argued that emotions arise from how individuals assess a situation's relevance to their goals and well-being, introducing a more personalised dimension to emotional theory. Later on, Barrett (2017) challenged Ekman's universality thesis, proposing that emotions are constructed from internal and external factors, shaped by cultural influences; this constructivist approach emphasises the brain's role in generating emotional responses based on context and cultural perspectives.

Emotional models can be classified into two categories: discrete/categorical and continuous/dimensional:

- Categorical Models: Categorical models, such as those by Carroll Izard and Paul Ekman, classify emotions into simple categories.

- Dimensional Models: Dimensional models, which map emotions onto dimensions such as arousal, valence, and dominance, better capture the complexity of affect.

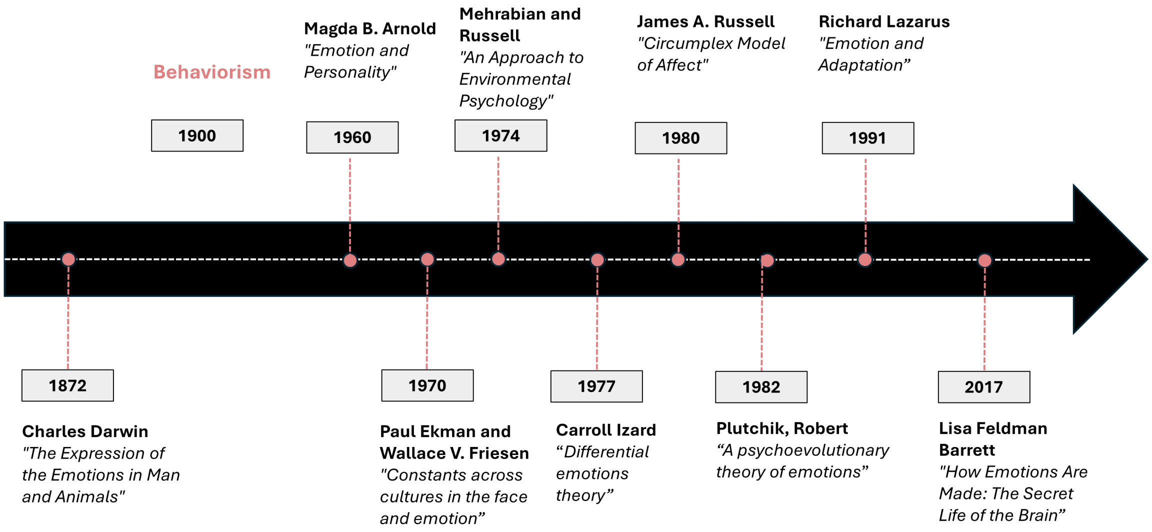

1872
Charles Darwin
"The Expression of the Emotions in Man and Animals"
Behaviorism
1900
1960
Magda B. Arnold
"Emotion and Personality"
1970
Paul Ekman and Wallace V. Friesen
"Constants across cultures in the face and emotion"
1974
Mehrabian and Russell
"An Approach to Environmental Psychology"
1977
Carroll Izard
"Differential emotions theory"
1980
James A. Russell
"Circumplex Model of Affect"
1982
Plutchik, Robert
"A psychoevolutionary theory of emotions"
1991
Richard Lazarus
"Emotion and Adaptation"
2017
Lisa Feldman Barrett
"How Emotions Are Made: The Secret Life of the Brain"

Figure 2.3: *Timeline of the historical evolution of psychological theories of emotion*

- Hybrid Models: Combine categorical and dimensional representations provide a balance between simplicity and precision.

2.2 Datasets for Emotion Recognition

Many Multimodal Emotion Recognition (MER) datasets have been extensively surveyed in the literature (Jianhua Zhang et al., 2020). According Gu, Shen, and J. Xu (2021), the currently available datasets for MER are all acted and include several widely used resources.

Introduced by Douglas-Cowie et al. (2011) the **HUMAINE** dataset aimed to create a coherent set of emotional responses from various forms of emotional content. Despite challenges such as noisy natural data and complex labelling methods, the dataset has been valuable, especially for real-world applications like call-centers.

Busso et al. (2008) presented the **IEMOCAP** dataset in 2008 to study expressive human communication through a combination of verbal and non-verbal channels. The dataset includes data from ten actors who participated in dyadic sessions, capturing both scripted and spontaneous emotional interactions. This multi-modal corpus is a key resource for studying and modelling human emotions.

The Acted Facial Expressions in the Wild **(AFEW)** dataset, proposed by Dhall et al. (2011), is a dynamic dataset that closely resembles real-world conditions, consisting of facial expressions extracted from movies. The dataset covers a wide age range and is labelled with six basic emotions plus a neutral class, making it useful for multi-modal emotion recognition experiments.

The **SEMAINE** dataset, introduced by McKeown et al. (2011), focuses on non-verbal communication and includes high-quality multimedia recordings of emotionally charged human interactions. It supports research on fluent interaction, especially in human-machine communication, with detailed labelling of emotional states.

Proposed by Ringeval et al. (2013) the **RECOLA** is a multi-modal corpus of spontaneous collaborative interactions in French, recorded during a video conference. It includes continuous annotations of emotions across arousal and valence dimensions, along with social behaviour labels, making it a valuable resource for studying affecting interactions.

A. B. Zadeh et al. (2018) introduced the CMU Multimodal Opinion Sentiment and Emotion Intensity **(CMU-MOSEI)** dataset, which is the largest dataset for sentiment analysis and emotion recognition to date. It includes over 23,000 annotated video clips with multi-modal data, providing a rich resource for studying sentiment and emotion in varied contexts.

The Multimodal EmotionLines Dataset **(MELD)**, proposed by Poria, Hazarika, et al. (2018), is an extension of the EmotionLines dataset. MELD contains around 13,000 utterances from 1,433 dialogues from the Friends TV series, labelled with sentiment and emotion. This multi-modal dataset is particularly useful for studying affective dialogue systems, providing text, audio, and visual data for each utterance.

2.3 Emotion Recognition (Text) from Deep Learning to Transformers

Textual Emotion Recognition

Textual Emotion Recognition (TER) involves understanding emotions conveyed through text. Unlike Sentiment Analysis, which broadly categorises text as positive, negative, or neutral (Medhat, Hassan, and Korashy, 2014), TER captures a more nuanced range of emotional states, such as joy, anger, and sadness. This provides deeper insights into individual experiences, particularly on platforms like Twitter, where emotions are frequently shared. TER is crucial in areas such as human-computer interaction and mental health monitoring (Peng et al., 2022). This section explores various TER approaches, including deep learning and transformer-based methods.

Deep Learning-Based Approach

In the realm of Emotion Recognition (ER), several Deep Learning (DL) architectures have been instrumental. Convolutional Neural Networks (CNNs) were originally designed for processing spatial data, such as images, leveraging their ability to capture local correlations (LeCun et al., 1998). Over time, CNNs have been adapted for sequential data, such as text, by applying character-level convolutions or using pre-trained word embeddings. This adaptation enables CNNs to capture the structure of sentences in text-based tasks, including Emotion Recognition (ER) (Kim, 2014).

Advanced architectures like ResNet and Inception have further expanded CNN capabilities by introducing residual connections and multi-scale feature extraction. ResNet uses skip connections to improve training of deeper networks, addressing issues like the vanishing gradient problem (He et al., 2016). Inception networks, on the other hand, employ multiple filter sizes to capture information at different resolutions, improving generalisation (Szegedy et al., 2015). These networks have proven effective in various NLP tasks, including ER, by extracting meaningful features from text data (Alzubaidi, J. Zhang, Humaidi, et al., 2021).

Recurrent Neural Networks (RNNs), designed to handle sequential data, are key to TER due to their capacity to model temporal dependencies. A distinguishing feature of RNNs is their use of hidden states, which serve as a form of memory, retaining information from previous inputs in the sequence, capturing the emotional flow within conversations (Hochreiter and Schmidhuber, 1997). Advanced models such as DialogueRNN (Majumder et al., 2019) have leveraged RNNs to track the flow of emotions within dialogues. Moreover, Long Short-Term Memory (LSTM) networks, a variant of RNNs, mitigate the vanishing gradient problem, enhancing the model's ability to maintain long-term contextual information.

Recent approaches have further advanced the field by incorporating additional layers of context and knowledge. DialogueGCN (Ghosal, Majumder, Poria, et al., 2019), which employs a Bi-GRU architecture, and the Knowledge-Enriched Transformer (KET) (Zhong, D. Wang, and Miao, 2019), which integrates hierarchical self-attention with commonsense knowledge. Additionally, Relational Graph Attention networks (RGAT) (Ishiwatari et al., 2020) and the Knowledge Aware Incremental Transformer (KAITML) (D. Zhang et al., 2020) represent cutting-edge developments that enhance emotion recognition by context-aware mechanisms.

These advancements underscore the evolution from traditional neural networks to complex Deep Learning models that significantly improve the accuracy and efficacy of emotion recognition in conversational texts.

Transformer Background

Transformers are a revolutionary class of deep learning models that have dramatically advanced the field of Natural Language Processing (NLP) since their introduction in 2017 with the paper "Attention Is All You Need" (Vaswani, 2017). They have been successfully adapted for use in domains such as computer vision, speech recognition, and emotion analysis (Baevski et al., 2020). Transformers are often pre-trained on extensive text corpora using objectives such as language modelling.

Core Innovation: Self-Attention Mechanism.The central innovation of Transformers lies in their self-attention mechanism. This mechanism allows the model to weigh the importance of different parts of an input sequence in parallel, rather than sequentially. It operates by computing attention scores using three vectors: query (Q), key (K), and value (V), obtained by linearly projecting the input embeddings.The self-attention mechanism has proven particularly effective in emotion recognition tasks, as it helps capture emotional cues scattered throughout the text (Zhong, D. Wang, and Miao, 2019).

Model Architecture. Transformers are built upon an encoder-decoder framework, where the encoder processes the input sequence, and the decoder generates the output sequence without access to future tokens.

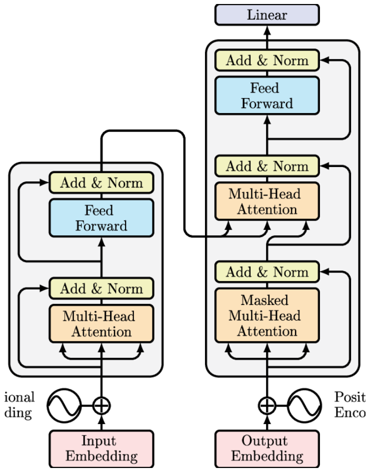


Figure 2.4: *Transformer architecture (Vaswani et al., 2017)*

1. **Positional Encoding:** Since Transformers lack the inherent positional awareness of RNNs, positional encodings are added to input embeddings. These encodings provide information about the order in the sequence, ensuring the model understands their relative positions.

2. **Multi-Head Attention:** This feature involves running multiple attention mechanisms in parallel, allowing the model to capture various aspects of the data simultaneously to manage diverse tasks.

3. **Position-Wise Feed-Forward Networks:** After the attention mechanisms, position-wise feed-forward networks are applied independently to each position in the sequence, before passing it on to the next layer.

4. **Residual Connections and Layer Normalisation:** To address challenges like the vanishing gradient problem, Transformers utilise residual connections and layer normalisation to stabilise the training process.

Transformers Approach

The development of Transformer-based models in Natural Language Processing (NLP) began with the original Transformer, introducing the revolutionary self-attention mechanism. To overcome its fixed-length context limitation, **Transformer-XL** was introduced, enhancing the model's ability to capture long-term dependencies through segment-level recurrence and relative positional encoding.

Building on the Transformer foundation, **Generative Pre-Training (GPT)** models, including GPT-2 and GPT-3, significantly scaled up the architecture. These models, which use only the decoder part of the Transformer, are optimised for text generation. However, their immense resource requirements and limitations in handling long-term dependencies spurred further innovations.

BERT (Bidirectional Encoder Representations from Transformers) advanced emotion recognition tasks by using bidirectional context understanding through Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). This allowed BERT to capture nuanced emotional cues in text, making it highly effective in tasks like sentiment analysis. Building on this, **XLM (Cross-Lingual Language Model)** enabled emotion recognition across languages by integrating Translation Language Modelling (TLM) with MLM, improving its cross-lingual capabilities.

XLNet further improved BERT's capacity for emotion recognition by introducing Permutation Language Modelling (PLM), which better captured complex emotional sequences, and leveraged Transformer-XL's strengths to address fixed-length constraints. **RoBERTa** optimised BERT's pre-training process, refining its ability to handle multitask settings, including emotion recognition, through techniques like dynamic masking (Peng et al., 2022).

Finally, **DistilBERT** provided a more efficient model, maintaining strong accuracy in emotion recognition tasks while offering faster performance, making it a lightweight solution for general-purpose NLP tasks, including the detection of emotional states in text (Peng et al., 2022).

This progression of Transformer-based models illustrates the continuous refinement of the architecture, with each iteration addressing specific challenges and pushing the boundaries of NLP capabilities, including emotion recognition.

2.4 Popular Pre-trained Large Language Model Families

The progression of Transformer-based models illustrates the continuous refinement of the architecture, with each iteration addressing specific challenges and pushing the boundaries of NLP capabilities (Brown, 2020), including emotion recognition . Building on these advancements, recent developments in large language models (LLMs) by leading organisations such as OpenAI, Meta, and Google have further expanded the potential of NLP (Minaee et al., 2024). These models have not only improved general language understanding but have also been adapted to tackle complex tasks like emotion recognition in multilingual contexts, aligning closely with the focus of this research. Figure 2.5 illustrates the key models in each family.

OpenAI's GPT Series: OpenAI's Generative Pre-trained Transformers (GPT) series has revolutionised language modelling, from GPT-1 to the more advanced GPT-4. The models have excelled in few-shot learning, particularly GPT-3, with its 175 billion parameters, which showcased abilities to handle a variety of tasks with minimal training. GPT-4 introduced multi-modal capabilities, processing both text and images, and enhanced user-aligned outputs through reinforcement learning from human feedback (RLHF). These advancements have proven beneficial for emotion recognition, especially in contexts where specific prompts guide the model.

Meta's LLaMA Models: The Large Language Model Meta AI (LLaMA) family stands out as a powerful open-source alternative. Launched in 2023, LLaMA models introduced novel elements like SwiGLU activation and rotary embeddings, enhancing multilingual dialogue capabilities. The second generation, LLaMA-2, further improved few-shot learning and prompt engineering, with RLHF improving conversational context management, making the models particularly useful for emotion recognition in dialogue-based tasks.

Google's PaLM Family: Google's Pathways Language Model (PaLM) series has advanced multilingual and reasoning tasks, with PaLM-2 offering a more compute-efficient solution. These models focus on instruction tuning and in-context learning, allowing them to perform well in emotion recognition across different languages. The Flan-PaLM models demonstrate strong abilities in few-shot learning, making them highly effective for tasks that require minimal input data to optimise responses.

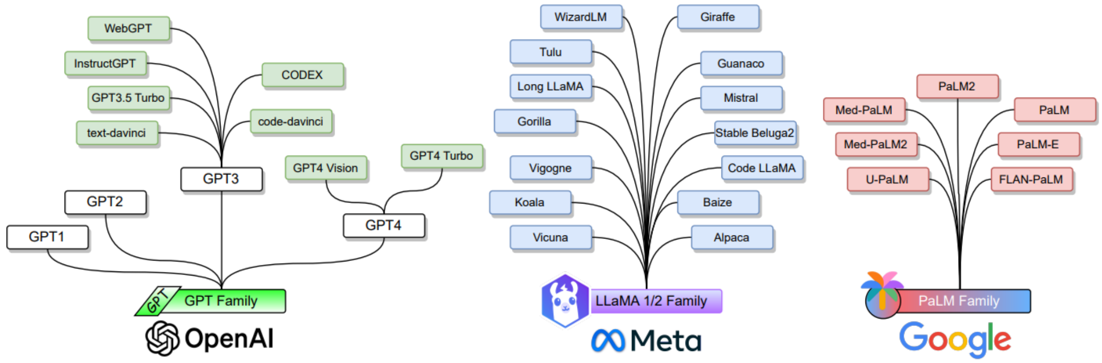


Figure 2.5: *Overview of popular LLM families from OpenAI, Meta, and Google (Minaee et al., 2024).*

Chapter 3

Multilingual Framework

This chapter presents a comprehensive multilingual framework for emotion recognition in dialogues. The framework consists of five stages, with the first two detailed in this chapter. The approach begins with data selection, followed by the translation of datasets from English to Spanish, comparing various models, including automatic translators, neural machine translation systems, and transformers tailored to emotion recognition. The results provide empirical insights into the handling of colloquial expressions and establish a foundation for exploring the integration of contextual information in emotion classification. The methodology and quantitative outcomes will be discussed in detail in the following chapter.

3.1 Introduction

The ability to understand emotions in conversational contexts is crucial for a wide range of applications, including customer service automation and mental health interventions (Tacconi et al., 2008). In multilingual environments, this task becomes increasingly complex due to linguistic diversity and cultural variations in emotional expression. This chapter introduces a comprehensive framework for emotion recognition in multilingual dialogues, employing advanced natural language processing methodologies such as long-sequence transformers and automated translation systems to address these challenges.

Despite Spanish being a widely spoken language with 437 millions of speakers across different continents (See Figure 3.1) and being the second most spoken language, the resources available for performing Speech Emotion Recognition (SER) in Spanish are notably limited. This scarcity presents significant challenges, particularly in the development and training of accurate models capable of effectively recognising emotions in Spanish. Existing datasets are often insufficient, with many suffering from poor annotation quality and lacking the authentic emotional expressions necessary for reliable SER in real-world applications (Garcia-Cuesta, Salvador, and Pãez, 2024).

Creating new datasets from scratch is an arduous process, involving considerable time and effort (Deng and Ren, 2023), such as recruiting actors to simulate emotions or establishing validation processes for real or elicited scenarios. To circumvent these challenges, this work explores an alternative approach: translating existing multimodal datasets, such as IEMOCAP and MELD, from English to Spanish. By leveraging robust pre-existing datasets, this approach addresses the specific difficulties of Spanish-language emotion recognition without the need for costly new dataset creation.

The motivation to utilise the IEMOCAP and MELD datasets in this study is grounded in their extensive multimodal resources and their well-established significance within the literature on emotion recognition. IEMOCAP, with its comprehensive collection of audio, visual, and textual data (Busso et al., 2008), aligns closely with the theories discussed in the previous chapter. It supports Paul Ekman and Wallace V. Friesen's theory of universal basic emotions (Ekman and Friesen, 1971), which classifies emotions into categories such as anger, disgust, fear, happiness, sadness, and surprise. The dataset's multimodal nature, coupled with detailed annotations, also makes it a resource for examining Richard Lazarus's cognitive appraisal theory (Lazarus, 1991), providing data to study how individuals appraise and react to emotionally significant events in conversations.

Similarly, the MELD dataset, derived from the TV series "Friends" (Poria, Hazarika, et al., 2018), also adheres to the classification of emotions proposed by Paul Ekman (Ekman and Friesen, 1971), making it a valuable resource for emotion recognition research. MELD's multimodal approach, encompassing text, audio, and visual data, mirrors the complex, context-dependent nature of emotional experiences emphasised in Lisa Feldman Barrett's theory of constructed emotions (Barrett, 2017).

The quality of these translations is crucial to ensuring that emotion recognition models can operate effectively in a multilingual context. The evaluation of these translation tools will be carried out in two phases. In the first phase, an empirical evaluation will focus on verifying whether the translations adequately preserve the colloquial and emotional meaning of the original sentences. Representative sentences from the datasets will be selected, and an analysis will be conducted to determine if the translations generated by each tool maintain the context and emotional intent. This analysis is based on direct observation and manual validation, ensuring that linguistic subtleties, especially in colloquial expressions, remain intact in the Spanish translations.

The second phase of the evaluation, which will be addressed in the next chapter, will analyse the impact of these translations on the performance of the complete emotion classification pipeline. Quantitative metrics will be used to compare how each translation tool affects the accuracy of emotional predictions. The model will be trained and evaluated using the translated datasets, allowing for the identification of which tool produces the best results in terms of emotion classification.

3.2 Related work

Existing research in emotion recognition for the Spanish language primarily falls into two categories: the development of Spanish datasets or the use of existing datasets for emotion classification tasks. However, a novel approach that builds a complete pipeline from translation to emotion classification has yet to be fully explored. For instance, in the work by Troiano, Klinger, and Padó (2020), translation tools were employed to preserve emotional content using a machine translation (MT) system; nevertheless, their approach focused on back translation, where the original text is

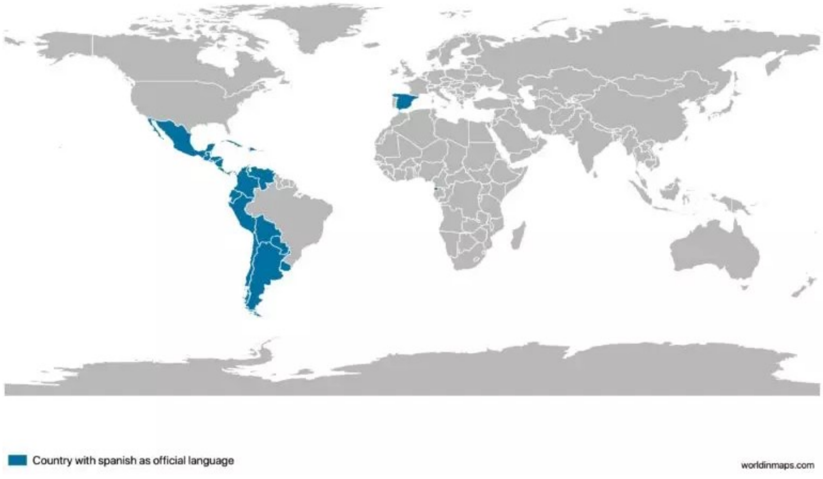


Figure 3.1: *World map of Spanish-speaking countries.*
*Source: Instituto Cervantes report (*El español: una lengua viva *2019).*

translated into another language and then back into the original language to enhance the quality of the data or correct errors.

This section begins by reviewing the Spanish datasets used for tasks related to emotion analysis or recognition. It then proceeds to examine the Neural Machine Translation (NMT) methods employed in multilingual emotion recognition tasks.

3.2.1 Spanish datasets for Speech Emotion Recognition

Spanish MEACorpus 2023 dataset: As part of the Multimodal speech-text emotion analysis, the Spanish corpus was created using audio segments sourced from public Spanish-language YouTube channels. These channels were selected to capture specific emotions such as disgust, anger, and joy in various contexts, including political speeches, sports reactions, and entertainment content. The dataset includes over 13 hours of audio from 5,129 segments, which were manually annotated with six emotion labels: disgust, anger, joy, sadness, fear, and a neutral category. The audio was recorded under diverse conditions, including noisy outdoor environments and controlled studio settings, adding complexity to the analysis. The dataset is imbalanced, with neutral and disgust emotions being the most prevalent (Pan et al., 2024).

EmoMatchSpanishDB: This is a Spanish speech emotion recognition dataset created using an elicited approach, developed to support research in automatic emotion recognition in this language. Fifty non-actors participated, expressing the six basic emotions identified by Ekman: anger, disgust, fear, happiness, sadness, and surprise, along with a neutral tone (Ekman and Friesen, 1971). A total of 4,200 audio files were recorded in a professional studio under controlled conditions, ensuring consistent, noise-free audio quality. Before each recording, an emotion induction procedure using emotionally charged images and texts was employed to ensure that participants consistently expressed the intended emotions.

NRC Word-Emotion Association Lexicon (EmoLex). Is a lexicon of English words annotated with associations to eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive) (Mohammad and Turney, 2013). The annotations were manually curated through crowdsourcing. However, the Spanish version of EmoLex shows limitations in accuracy due to the challenges inherent in directly translating emotion-laden words from English to Spanish (Liang and S. Wang, 2022).

CMU-MOSEAS: Provides a large-scale multilingual and multimodal resource for emotion and sentiment analysis. For the Spanish subset, the dataset includes approximately 10,000 sentences, each annotated with multiple labels, including sentiment (positive, negative, neutral), subjectivity, and a range of emotions such as happiness, sadness, anger, fear, and surprise. The data were gathered from diverse monologue videos on YouTube, ensuring a wide variety of speakers and topics. The annotations were aligned with the audio and visual data to facilitate comprehensive emotion recognition research in Spanish (A. Zadeh et al., 2020).

3.2.2 Neural Machine Translation Technologies

Recent advancements in Neural Machine Translation (NMT), particularly led by end-to- end transformer- based models, have greatly improved translation quality, making NMT the dominant approach in practical machine translation systems. These models excel in handling complex sentence structures, ensuring accurate translations across multiple languages. Below are key technologies utilised for multilingual translation in emotion recognition tasks (Tan et al., 2020).

Fairseq: Developed by Facebook AI Research, is a sequence-to-sequence learning framework that supports transformer-based models. It is highly customisable and has been employed to train state-of-the-art models for translation tasks, such as the WMT19 system (Ng et al., 2019). Fairseq's capability to manage complex sentence structures makes it particularly effective for translating emotional dialogues.

M2M-100: A model developed by Facebook AI, M2M-100 enables direct translation between 100 languages without the need for English as a pivot language. This model maintains emotional context, which is essential for multilingual emotion recognition (Fan et al., 2021).

Marian: Developed by the Microsoft Translator team, it is an efficient and self-contained NMT framework (Junczys-Dowmunt et al., 2018), written entirely in C++ with minimal dependencies. Widely used by various organisations and companies, Marian powers Microsoft's neural machine translation engine, offering a robust solution for translation tasks.

Opus-MT: Built upon the Marian NMT framework, Opus-MT focuses on low-resource languages and has been fine-tuned for specific language pairs. It is especially useful for translating datasets in languages such as Spanish, while preserving the emotional nuances within conversations (Tiedemann and Thottingal, 2020).

DeepL API: Is a neural machine translation service that utilises deep learning techniques to generate translations[1]. It is recognised for its advanced algorithmic approach, which effectively preserves both syntactic and semantic integrity, particularly in complex sentences (Hidalgo-Ternero, 2020).

Google Translate: Is one of the most widely used translation tools, known for its broad language support and accessibility. Google Translate's primary advantage lies in its ability to handle a wide range of languages and its ease of use, making it a convenient choice for quick translations.

DeepL has been recognised for its high-quality translations, particularly when compared to tools like Google Translate. Recent studies demonstrate that DeepL outperforms Google Translate in terms of fluency (Bhardwaj et al., 2020) and handles colloquial expressions more effectively (Hidalgo-Ternero, 2020). Therefore, both have been incorporated into the methodology to compare their impact on emotion recognition classification tasks

3.3 Methodology

The overall workflow of the translation and processing framework is summarised in Figure 3.2 below. This diagram visually represents the steps involved, providing a clear overview of how the datasets are translated and prepared for subsequent steps.

This section outlines the methodology employed for translating the IEMOCAP and MELD datasets into Spanish, a crucial step in developing a robust multilingual framework for emotion recognition in dialogues. These datasets, widely used for research in Emotion Recognition for text, include dialogues with associated metadata such as text, speaker information, and emotion labels. To facilitate the application of advanced emotion recognition models in Spanish, the translation of these datasets was performed using a combination of state-of-the-art translation tools.

The translation process involved three evaluations: Google Translate API, MarianMT from Hugging Face, and the DeepL API. Each of these tools was selected for its strengths in handling various linguistic nuances, ensuring that the translated datasets retain the emotional context and speaker-specific subtleties essential for accurate Emotion Recognition in text.

The following subsections provide a detailed description of the translation process, including the methodologies used by each translation tool, and how the translated data was subsequently processed and assessed.

3.3.1 Spanish translators

In this section, the different methodologies employed to translate each dialogue for all speakers across the IEMOCAP and MELD datasets from English to Spanish are described. A critical aspect of this translation process is the ability of each tool to accurately handle colloquial expressions in

[1] https://www.deepl.com/en/blog/how-does-deepl-work

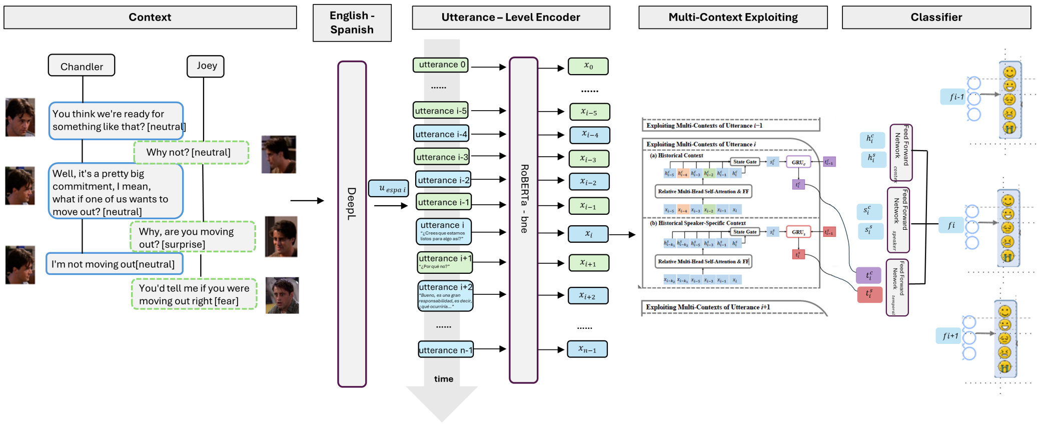

Context
English - Spanish
Utterance – Level Encoder
Multi-Context Exploiting
Classifier
Chandler
Joey
You think we're ready for something like that? [neutral]
Why not? [neutral]
Well, it's a pretty big commitment, I mean, what if one of us wants to move out? [neutral]
Why, are you moving out? [surprise]
I'm not moving out[neutral]
You'd tell me if you were moving out right [fear]
DeepL
$u_{espa\ i}$
utterance 0
......
utterance i-5
utterance i-4
utterance i-3
utterance i-2
utterance i-1
utterance i
"¿Crees que estamos listos para algo así?"
utterance i+1
"¿Por qué no?"
utterance i+2
"Bueno, es una gran responsabilidad, es decir, ¿qué ocurriría..."
......
utterance n-1
time
RoBERTa - bne
x_0
x_{i-5}
x_{i-4}
x_{i-3}
x_{i-2}
x_{i-1}
x_i
x_{i+1}
x_{i+2}
x_{n-1}
Exploiting Multi-Contexts of Utterance $i-1$
Exploiting Multi-Contexts of Utterance i
(a) Historical Context
State Gate
s_i^c
GRU$_c$
t_{i-1}^c
h_{i-5}^c h_{i-4}^c h_{i-3}^c h_{i-2}^c h_{i-1}^c h_i^c
t_i^c
Relative Multi-Head Self-Attention & FF
x_{i-5} x_{i-4} x_{i-3} x_{i-2} x_{i-1} x_i
(b) Historical Speaker-Specific Context
State Gate
s_i^s
GRU$_s$
t_{i-1}^s
$h_{i-k_1}^s$ $h_{i-k_2}^s$ h_{i-5}^s h_{i-3}^s h_{i-1}^s h_i^s
t_i^s
Relative Multi-Head Self-Attention & FF
x_{i-k_2} x_{i-k_1} x_{i-5} x_{i-3} x_{i-1} x_i
Exploiting Multi-Contexts of Utterance $i+1$
h_i^c
h_i^s
Feed Forward Network context
s_i^c
s_i^s
Feed Forward Network speaker
t_i^c
t_i^s
Feed Forward Network temporal
fi-1
fi
fi+1

Figure 3.2: *The overall architecture of the proposed Context-aware Multi-lingual framework. Utterances from different speakers are represented in coloured blocks (blue and green). The framework processes these utterances through translation, multi-context exploitation, and final classification.*

English, ensuring that the emotional nuances are preserved.

Task Definition. In this study, the initial step involves translating each utterance u_i within a conversation C_i into Spanish using a translation function T. This process is represented as:

$$T(u_i) = u_i'$$

where u_i' denotes the translated utterance in Spanish. Once all utterances u_i within the conversation C_i have been translated, the conversation is represented as a sequence of translated utterances:

$$C_i = (u_0', u_1', \ldots, u_{n-1}')$$

The next task is to predict the corresponding emotion label $y_i \in Y$ for each translated utterance u_i'. Formally, let C represent the set of conversations, S the set of speakers, and Y the set of emotion labels. For any conversation $C_i \in C$, the objective is to assign an emotion label y_i to each translated utterance u_i', structuring the task as an utterance-level sequence tagging problem. The function $S(u_i')$ maps each translated utterance to its respective speaker from the set S. This approach ensures that the model operates under real-world conditions, where only past translated utterances are available.

Translation Methodology. The translation process involves converting each dialogue from English to Spanish while maintaining the integrity of the emotions conveyed in the original utterances. The following methodologies were compared:

- **Google Translate:** In this experiment, the dialogues were translated utilising the `googletrans` Python library. The translation process entailed converting each utterance within the dialogues from the final dataset[2], ensuring that the emotional content and context were preserved throughout the translation. The implemented code is structured to read JSON files containing dialogues, translate the text of each utterance from English to Spanish, and subsequently save the translated dialogues into new JSON files.

- **MarianMT:** Developed by the Marian team and available through the Hugging Face transformers library[3], is a transformer-based model specifically designed for neural machine translation. Its fine-tuning on large parallel corpora enables it to handle multilingual tasks with precision, making it highly effective for translating dialogue datasets, based on the Marian system developed by Junczys-Dowmunt et al. (2018). The pre-trained 'opus-mt-en-es' model was utilised in this study to perform the translation tasks. The process involved loading the model and its corresponding tokeniser, translating each utterance within the dialogues, and then decoding the text into Spanish.

- **DeepL API:** The implementation involved using the DeepL API via an HTTP request interface. The process started by reading the input JSON final files, where each dialogue was translated sentence by sentence. For each utterance, a translation request was sent to the API, with a pause between requests to avoid potential throttling issues. The translated text was then updated in the corresponding dialogue structure and then saved in a new JSON file for further use in emotion recognition tasks.

3.4 Evaluation

3.4.1 Dataset

It is utilised the two well-established datasets for the experiments: the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) and the Multi-modal EmotionLines Dataset (MELD). It is important to note that the datasets were taken directly from a publicly available repository and were already pre-processed and cleaned, requiring no further processing for this study[2].

The IEMOCAP dataset is a multimodal collection that includes visual, audio, and textual data, though this study focuses on the textual modalities. Each sentence within this dataset is annotated with one of several emotion labels, including happiness, sadness, neutrality, anger, surprise, excitement, frustration, disgust, fear, and others (Busso et al., 2008).

The MELD dataset, which derived from the TV series *Friends*, offers a multi-modal collection with aligned acoustic, textual, and visual information for each utterance. In this research, it is utilised the textual data. The dataset is split into training, validation, and test sets, covering multiple speakers and emotional states. Each utterance is categorised into one of seven emotion

[2] `https://github.com/Ydongd/ERCMC/tree/master/datasets/final`

[3] `https://huggingface.co/docs/transformers/en/model_doc/marian`

labels: anger, disgust, sadness, joy, neutral, surprise, and fear (Poria, Hazarika, et al., 2018).

The dataset are divided into training and development subsets, as detailed in Table 3.1.

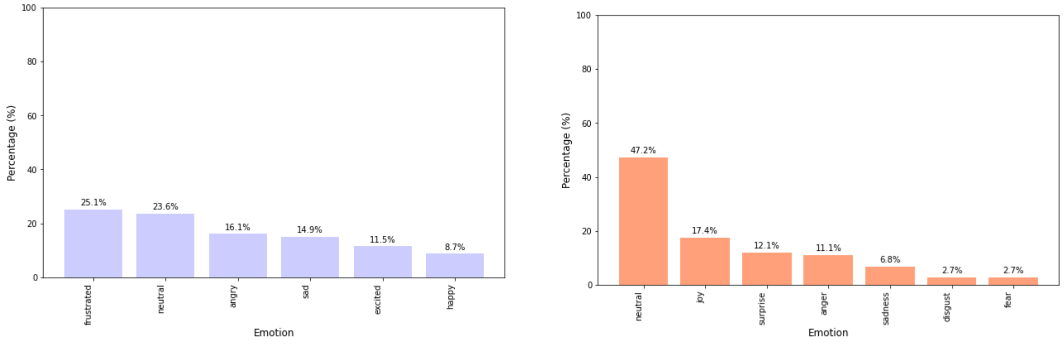
Table 3.1: Statistics of the Datasets

Dataset	Conversations			Utterances			Tokens			Classes
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	
IEMOCAP	120	120	31	5,810	5,810	1,623	207,113	25,451	66,727	6
MELD	1,038	114	280	9,989	1,109	2,610	446,521	48,561	115,285	7

Note: The token counts were calculated using OpenAI's GPT-4.

For the IEMOCAP dataset (Figure 3.3a), the distribution of labels is relatively balanced across most classes, with the exception of the 'happy' emotion, which is underrepresented, comprising only 8% of the total labels. The 'frustrated' class is the most frequent, accounting for 25% of the labels. This relatively even distribution, apart from the 'happy' class, suggests that the dataset provides a reasonable foundation for training models.

In contrast, the MELD dataset (Figure 3.3b) exhibits an imbalance, particularly with the 'neutral' class, which dominates the dataset at 47%. The 'fear' and 'disgust' classes are notably underrepresented, each contributing only around 3% of the total labels. This imbalance, especially the prevalence of 'neutral', may pose challenges in the classification of emotions.



(a) *IEMOCAP Dataset*

(b) *MELD Dataset*

Figure 3.3: *Comparison of emotion label distributions in the IEMOCAP and MELD datasets.*

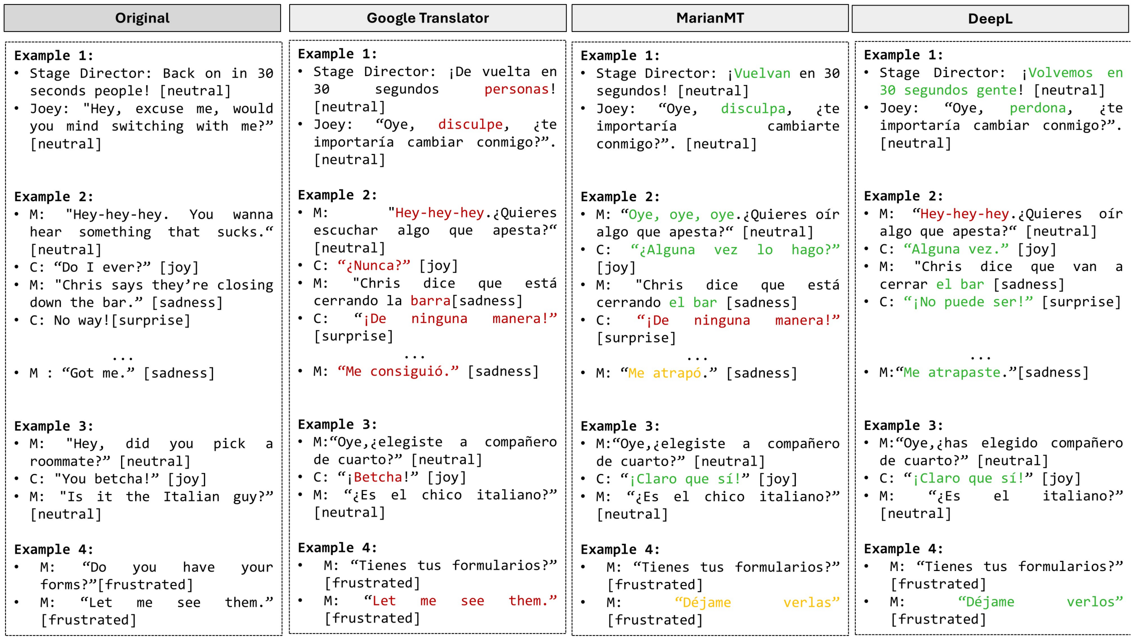
Given these imbalances, it is essential to apply evaluation metrics that account for the uneven distribution of classes. In the next chapter, the methodology and results of the emotion classification task will highlight the use of the weighted F1 score. This metric, which balances precision and recall according to class representation, is particularly suited for addressing the challenges posed by the dominant 'neutral' class in MELD.

3.4.2 Baselines

Spanish translators. This section describes the use of Google Translate as a basic baseline for translating the datasets from English to Spanish. Google Translate is a widely used translation tool that does not employ complex artificial intelligence models or advanced transformer architectures.

The purpose of using Google Translate as a baseline is to establish a clear comparison with more sophisticated models, which are expected to offer higher accuracy and consistency in challenging contexts such as emotion recognition in dialogues.

3.4.3 Results



Original	Google Translator	MarianMT	DeepL
Example 1: • Stage Director: Back on in 30 seconds people! [neutral] • Joey: "Hey, excuse me, would you mind switching with me?" [neutral]	**Example 1:** • Stage Director: ¡De vuelta en 30 segundos personas! [neutral] • Joey: "Oye, disculpe, ¿te importaría cambiar conmigo?". [neutral]	**Example 1:** • Stage Director: ¡Vuelvan en 30 segundos! [neutral] • Joey: "Oye, disculpa, ¿te importaría cambiarte conmigo?". [neutral]	**Example 1:** • Stage Director: ¡Volvemos en 30 segundos gente! [neutral] • Joey: "Oye, perdona, ¿te importaría cambiar conmigo?". [neutral]
Example 2: • M: "Hey-hey-hey. You wanna hear something that sucks." [neutral] • C: "Do I ever?" [joy] • M: "Chris says they're closing down the bar." [sadness] • C: No way![surprise] ... • M : "Got me." [sadness]	**Example 2:** • M: "Hey-hey-hey.¿Quieres escuchar algo que apesta?" [neutral] • C: "¿Nunca?" [joy] • M: "Chris dice que está cerrando la barra[sadness] • C: "¡De ninguna manera!" [surprise] ... • M: "Me consiguió." [sadness]	**Example 2:** • M: "Oye, oye, oye.¿Quieres oír algo que apesta?" [neutral] • C: "¿Alguna vez lo hago?" [joy] • M: "Chris dice que está cerrando el bar [sadness] • C: "¡De ninguna manera!" [surprise] ... • M: "Me atrapó." [sadness]	**Example 2:** • M: "Hey-hey-hey.¿Quieres oír algo que apesta?" [neutral] • C: "Alguna vez." [joy] • M: "Chris dice que van a cerrar el bar [sadness] • C: "¡No puede ser!" [surprise] ... • M:"Me atrapaste."[sadness]
Example 3: • M: "Hey, did you pick a roommate?" [neutral] • C: "You betcha!" [joy] • M: "Is it the Italian guy?" [neutral]	**Example 3:** • M:"Oye,¿elegiste a compañero de cuarto?" [neutral] • C: "¡Betcha!" [joy] • M: "¿Es el chico italiano?" [neutral]	**Example 3:** • M:"Oye,¿elegiste a compañero de cuarto?" [neutral] • C: "¡Claro que sí!" [joy] • M: "¿Es el chico italiano?" [neutral]	**Example 3:** • M:"Oye,¿has elegido compañero de cuarto?" [neutral] • C: "¡Claro que sí!" [joy] • M: "¿Es el italiano?" [neutral]
Example 4: • M: "Do you have your forms?"[frustrated] • M: "Let me see them." [frustrated]	**Example 4:** • M: "Tienes tus formularios?" [frustrated] • M: "Let me see them." [frustrated]	**Example 4:** • M: "Tienes tus formularios?" [frustrated] • M: "Déjame verlas" [frustrated]	**Example 4:** • M: "Tienes tus formularios?" [frustrated] • M: "Déjame verlos" [frustrated]

Figure 3.4: *Examples of three conversations from the original MELD dataset, translated using Google Translator, MarianMT, and DeepL. Translation issues are highlighted in red, well-translated phrases are marked in green, and neutral translations are shown in orange.*

In this section, it is compared the performance of the three translation tools by qualitatively analysing translations of conversations from the MELD dataset mainly which contains a higher occurrence of colloquial phrases and informal speech, making it particularly challenging for machine translation tools. As shown in Figure 3.4, the translations were evaluated across four example conversations. Translation issues are highlighted in red, well-translated phrases are marked in green, and neutral translations are shown in orange.

Google Translator. One of the primary issues identified with Google Translator is its difficulty in handling short, fragmented, or unconventionally structured sentences. Additionally, Google Translator struggles with certain special characters and punctuation marks. For instance, phrases like "Let me see them." or "Next." were not translated correctly due to Google Translator's inability to handle the lack of context or specific punctuation properly.

In Example 1 of Figure 3.4, we observe that phrases like "You betcha!" and "Got me" are not translated appropriately, leading to a loss of meaning and context. The expression "No way!" in Example 2, which usually conveys surprise, is translated as "De ninguna manera," which in Spanish could be interpreted as anger rather than disbelief or surprise. These issues highlight the limitations of Google Translator in preserving the nuances of conversational English, particularly when dealing with colloquial expressions.

MarianMT. Shows a significant improvement over Google Translator, effectively handling most of the issues encountered previously. As observed in the examples, MarianMT provides more accurate translations for colloquial phrases, preserving the intended meaning and context better than Google Translator.

However, MarianMT still exhibits some challenges. For instance, in Example 2, "No way!" is translated as "De ninguna manera," similar to Google Translator, which does not fully capture the intended surprise. Additionally, there are minor gender agreement issues, such as translating "forms" as feminine, which is incorrect in this context. Despite these small flaws, MarianMT generally offers a more faithful translation compared to Google Translator.

DeepL. It demonstrates the most accurate translation among the three tools, particularly in handling colloquial language and preserving the natural flow of Spanish. As seen in the examples, DeepL translates phrases like "You betcha!" and "Got me" correctly, maintaining the context and emotional tone of the original English sentences.

DeepL also effectively manages phrases like "No way!" in Example 2, translating it as "¡No puede ser!" which conveys the correct sense of surprise in Spanish. The accuracy and contextual relevance of DeepL's translations make it superior to both Google Translator and MarianMT in this qualitative analysis.

The qualitative analysis of these translations, conducted by a native Spanish speaker, reveals that while Google Translator and MarianMT are both open-source tools, DeepL, despite its associated costs, provides superior translations. DeepL's ability to handle colloquial phrases and maintain the natural context of conversations suggests that incorporating such a robust translation tool, which leverages AI and advanced linguistic algorithms, crucial for the emotion classification. These qualitative findings will be complemented by a quantitative evaluation in the next chapter, where the performance of these translation tools will be assessed in the context of emotion classification using metrics such as the weighted F1 score. This combined approach will provide a comprehensive understanding of how translation quality impacts the accuracy and effectiveness of emotion recognition systems.

3.5 Summary

This chapter has outlined the initial stages of the multilingual framework for emotion recognition, with a particular focus on the Spanish language. The study underscores the considerable challenge posed by the scarcity of Spanish datasets for Speech Emotion Recognition (SER), despite the language's widespread use globally. Existing resources frequently depend on basic dictionaries or direct translations, which do not provide the necessary depth for effective SER applications.

Nonetheless, this chapter demonstrates that generating new datasets is not always essential. By employing advanced AI-based translation tools like DeepL, it is possible to produce high-quality translations of English datasets into Spanish. This method represents a significant advancement, particularly when compared to previous approaches using tools like Google Translate. The efficacy of this approach is validated empirically through its application to established SER datasets, such as IEMOCAP and MELD, which were translated into Spanish and subsequently used to assess the performance ins emotion recognition models.

The findings of this chapter highlight that, while there are inherent challenges in addressing emotion recognition for the Spanish language, these can be effectively managed through the strategic application of translation tools and specialised language models. Additionally, this approach is not only innovative for processing data in Spanish but also holds the potential for broader application to other languages, offering a versatile and robust framework for emotion recognition in multilingual contexts.

The subsequent chapter will further explore the integration of contextual information to advance emotion recognition methodologies. Future chapters will undertake a detailed evaluation to determine if these Spanish-trained models offer superior performance in emotion classification tasks compared to their English-trained counterparts, with assessments based on metrics such as the weighted F1 score.

Chapter 4

Impact of Context in Emotion Recognition Conversations

This chapter explores the use of a specialised utterance-level encoder integrated with a transformer trained on Spanish corpora to evaluate the role of context in emotion recognition. Additionally, it investigates how historical context and speaker identity influence the accuracy of emotion classification, with results evaluated through F1 score, precision, and recall across emotion classes. Building on the framework from the previous chapter, this study leverages the transformer's attention mechanism to incorporate contextual subtleties, enhancing the model's ability to detect emotions more effectively.

4.1 Introduction

This chapter is motivated by the need to improve the accuracy of Emotion Recognition in Conversations, a critical area for the development of more human-like and effective dialogue systems. In multilingual environments, this task becomes particularly challenging due to linguistic diversity and cultural variations in emotional expression. Furthermore, the contextual nature of the data, as exemplified by the MELD and IEMOCAP datasets, highlights the importance of considering context for accurate emotion classification. In these datasets, the emotions of the characters are deeply intertwined with the conversational context, making the correct interpretation of that context essential for any Emotion Recognition in Conversations (ERC) model.

Additionally, transformer-based models, particularly those trained on large Spanish-language corpora, such as RoBERTa-bne, have shown great potential in generating robust embeddings from translated utterances. These embeddings serve as crucial representations for each utterance, capturing both semantic and emotional content, which is essential for emotion recognition tasks. Recent advances in deep learning have demonstrated that these pre-trained models, when fine-tuned for specific tasks, such as ERC, outperform traditional methods (Pan et al., 2024).

A key aspect of improving the accuracy of emotion classification lies in the integration of context, such as historical context and speaker identity, which can significantly influence how emotions are expressed and perceived in dialogues. Current research highlights the importance of incorporating multiple types of context, as emotions can shift depending on what has been said previously and by whom (Ghosal, Majumder, Mihalcea, et al., 2020). To fully understand the emotional dynamics within a conversation, it is crucial to not only consider the current content but also how emotions evolve over time, capturing both temporal dependencies and speaker-specific characteristics

(Poria, Cambria, et al., 2017).

To address these contextual challenges, this study leverages advanced transformer models, such as RoBERTa-bne, which have been pre-trained on large Spanish-language corpora and fine-tuned for ERC. These models, combined with the integration of multiple contexts, are particularly effective in capturing long-term contextual relationships and improving emotion recognition performance in multilingual environments (Luo, Phan, and Reiss, 2023; Wei et al., 2023).

The methodological approach adopted in this study is developed in two main phases. In the first phase, the integration of context into the multi-head attention mechanism within the transformer's encoder is explored; the representations generated by the transformer are then processed through a GRU (Gated Recurrent Unit) network to model the temporal evolution of emotions. In the second phase, a softmax layer is employed for the final classification of emotions, ensuring that the emotional subtleties detected in the previous stages are translated into accurate predictions. These two phases form part of the final stages of the multilingual framework illustrated in Figure 3.2. Following the approach outlined in the previous chapter, this study continues to analyse and work with the IEMOCAP and MELD datasets, which are particularly valuable due to their rich contextual nature. These datasets allow for the exploration of how different types of contexts influence emotional expression in various conversational situations. Evaluating the models on these datasets will not only help validate the effectiveness of the proposed approach but also provide valuable insights into how ERC models can be refined and improved in future research.

To assess the model's effectiveness and ensure that any potential class imbalances in the IEMOCAP and MELD datasets are properly managed, the weighted F1 score is utilised. This is particularly relevant since, in previous chapters, a classification bias due to dataset imbalance was identified, ensuring that less represented classes are given the same importance as the more frequent ones. Additionally, the effectiveness of the different translation methods discussed in the previous chapter will be evaluated using this metric.

The aim of this chapter is to demonstrate how the integration of historical, temporal, and speaker-specific contexts, along with the use of advanced transformer models in the relevant language, can improve the accuracy of emotion classification in dialogues. This approach will enable a better understanding of emotions in conversations, facilitating the development of more effective and empathetic dialogue systems.

4.2 Related work

4.2.1 Contextual Analysis in ERC

The inclusion of context in ERC has evolved significantly over the years. Initial approaches primarily relied on static text classification methods, such as bag-of-words and basic sentiment analysis, which lacked the capacity to capture the dynamic and sequential nature of conversations. The introduction of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks marked a significant advancement, allowing models to process sequential data and consider the temporal progression of emotions. For example, DialogueRNN and DialogueGCN demonstrated how these methods could model the flow of emotions within a dialogue, improving classification performance by accounting for the context provided by previous utterances (Ezzameli and Mahersia, 2023).

As deep learning techniques evolved, attention mechanisms and transformers became the tools of choice for ERC tasks. These models, exemplified by the Knowledge-Enriched Transformer (KET) and Relational Graph Attention Networks (RGAT), excel at capturing complex dependencies across entire dialogues. They also incorporate speaker-specific context, recognising that different individuals may express emotions differently, and these expressions need to be understood within the broader conversational context (Ezzameli and Mahersia, 2023).

4.2.2 Advanced Contextual Transformers

Transformers, with their self-attention mechanisms, have revolutionised the way context is utilised in ERC. By enabling models to weigh the importance of different parts of a conversation, transformers can more accurately predict emotions by considering not just the immediate utterance but the entire dialogue history. The integration of context into transformers, particularly using techniques like multi-head attention and relative positional encoding, has been shown to significantly enhance ERC performance. Studies such as Wei et al. (2023) illustrates how these models can leverage historical and speaker-specific context to improve emotion recognition accuracy, particularly in complex and nuanced conversations.

The application of context-aware transformers in ERC has also been extended to include gated recurrent unit (GRU) that refine the temporal representation of emotions within a conversation. This approach ensures that models can not only capture the immediate context but also maintain a coherent understanding of the emotional flow throughout the dialogue (Jiao et al., 2019).

4.2.3 Architectures for Spanish Emotion Recognition

Pre-trained models based on architectures like BERT and RoBERTa have proven to be highly effective for emotion recognition in Spanish text (Pan et al., 2024). These models, tailored specifically for the Spanish language, allow for more accurate emotion classification. The following sections first discuss traditional methods, followed by a review of more advanced transformer-based models for SER.

Traditional Methods:
SVM (Support Vector Machine) is a machine learning method that maps data to a high-dimensional space to construct a linear decision boundary (Vapnik, 1995). In the study by Garcia-Cuesta, Salvador, and Pãez (2024), SVM outperformed other traditional machine learning techniques like XGBOOST and Feed-Forward Deep Neural Networks (FFNN), achieving the highest F1 score for emotion recognition tasks.

XGBOOST is an optimised gradient boosting framework that uses decision trees as base learners (T. Chen and Guestrin, 2016). It is known for its speed and performance in classification tasks. In the study by Garcia-Cuesta, Salvador, and Pãez (2024), XGBOOST was evaluated; although it did not perform as well as SVM, it was still an important method tested.

RNN (Recurrent Neural Network) combined with an enhanced BiLSTM and attention mechanism, as explored by Liang and S. Wang (2022) leverages the strengths of BiLSTM for capturing temporal dependencies in emotional data. The addition of attention further improves the model's capacity to focus on key emotional features within the text, leading to better emotion classification performance.

Transformer-Based Methods:
MarIA is a family of Spanish language models that includes both RoBERTa-based encoder models and GPT-2-based generative models. Trained on a 570 GB corpus of Spanish texts from the Spanish Web Archive, these models represent some of the largest and most proficient language models in Spanish. The MarIA models were developed using data crawled by the National Library of Spain from 2009 to 2019. In study by Pan et al. (2024) RoBERTa-based MarIA model was fine-tuned for text classification and delivered the best results in emotion recognition tasks, surpassing other Spanish models in various natural language understanding tasks.

BETO is one of the first pre-trained language models specifically designed for Spanish, based on the BERT architecture. It was trained on a diverse set of Spanish texts, including data from the OPUS project and Wikipedia. It has been used in the study by Pan et al. (2024) where it demonstrated strong performance in emotion recognition when fine-tuned for this specific task.

Similarly, **BERTIN**, a RoBERTa-based model, was developed using the Spanish portion of the mc4 dataset (De la Rosa et al., 2022). Though less complex than MarIA, BERTIN has proven effective in classification tasks, though it demonstrated slightly lower accuracy in emotion recognition, also highlighted in the same study.

For resource-constrained environments, **ALBETO** and **DistilBETO** offer more efficient alternatives. ALBETO, a smaller version of BETO, and DistilBETO, a distilled version, both reduce the computational demands while maintaining reasonable performance. Although not as powerful as their larger counterparts, they provide efficient solutions where computational resources are limited, as observed in the same evaluation (Pan et al., 2024) .

4.3 Methodology

Building on previous research that demonstrates the effectiveness of RoBERTa-based models for emotion recognition tasks in Spanish (Pan et al., 2024) and further supported by studies exploring contextual information in dialogues, which highlight RoBERTa's strong performance in such tasks (Wei et al., 2023), this methodology utilises RoBERTa-bne, a version of RoBERTa fine-tuned specifically on a Spanish-language corpus. This model functions as an encoder-decoder framework, enabling accurate prediction of emotion labels from translated Spanish data.

This section focuses on the next stages of the process from the framework illustrated in Figure 3.2. With the translated datasets now prepared, the next step is to describe the implementation of the Utterance-level encoder, the Multi-context Exploiting mechanism, and the final Classifier.

4.3.1 Utterance-Level Encoder

Given a conversation $U_i = (u_0, \ldots, u_{n-1})$, each translated utterance $T(u_i)$ undergoes several stages of processing to obtain a useful representation for emotion classification.

1. Tokenisation, Special Tokens and Embedding: Each translated utterance $T(u_i)$ is tokenised, converting into a sequence of tokens using the RoBERTa-bne model from Hugging Face[1]. During this process, the special tokens `[CLS]` and `[SEP]`are added to the beginning and end of the sequence, respectively. The `[CLS]` token is particularly crucial, as it is used to capture the global context of the utterance. Padding and attention masks are also applied to handle variable-length sequences, ensuring that the model focuses only on the relevant tokens, excluding the padding added to equalise sequence lengths.

2. Construction of the Hidden Matrix: After tokenisation, the sequence of tokens is passed through the RoBERTa-bne model, producing a hidden state matrix H'. This matrix H' contains vector representations for each token in the sequence, including the `[CLS]` token. The matrix H' can be expressed as:

$$H' = M_{\text{RoBERTa-bne}}(T(u_i))$$

where H' is the hidden state matrix and $M_{\text{RoBERTa-bne}}$ represents the RoBERTa-bne model. Note that H' reflects the processing of the translated utterance $T(u_i)$.

3. Representation Extraction: From the matrix H', the embedding corresponding to the `[CLS]` token, denoted as x'_i, is extracted. This embedding x'_i acts as a summarised and comprehensive representation of the translated utterance $T(u_i)$. This process is represented as:

$$x'_i = H'[0]$$

where $H'[0]$ corresponds to the hidden state associated with the `[CLS]` token in the sequence.

[1] `https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne`

The representation x_i' is crucial not only for capturing the global context of the translated utterance $T(u_i)$ but also for feeding into the subsequent stages of the model that involve exploiting multiple contexts.

After obtaining the encoding for each utterance, the framework extracts different contexts from this and previous encodings, which are then utilised for emotion classification via a feed-forward network. This approach unfolds through a series of stages designed to capture historical context, speaker-specific context, and temporal context, thereby enhancing the model's accuracy in dialogues.

As illustrated in Figure 4.1, two primary types of context are exploited to improve the accuracy of emotion classification:

- **Historical Context**: This context refers to the accumulated information from previous utterances in the conversation. It captures how emotions and content evolve over time, which is crucial for understanding the flow and progression of emotions as the conversation unfolds. In the model, as illustrated in Figure 4.1 (a), this context is processed using relative multi-head self-attention to focus on the relationships between the current utterance and the last 5 utterances. The output of the attention mechanism is combined with the state information and passed through a GRU to capture long-term dependencies, generating the updated representation t_i^c.

- **Speaker-Specific Context**: This context pertains to the identity of the speaker and how their individual characteristics influence the interpretation of their utterances. Different speakers may express emotions in distinct ways, and this context ensures that the model accounts for these variations. As depicted in Figure 4.1 (b), the speaker-specific context uses only the embeddings of the speaker's own utterances from the previous five dialogue turns. These embeddings are processed using relative multi-head self-attention, ensuring that the model attends to the most relevant utterances based on their relative positions within the dialogue. Afterwards, this information is passed through a GRU, which captures the evolution of speaker-specific emotions over time, generating the output t_i^s.

4.3.2 Multi-Context Integration

The **Relative Encoder** is an extension of the standard transformer encoder that incorporates relative positional information between words instead of using absolute positions. Unlike traditional encoders, which consider the fixed position of a word within a sentence, it accounts for the relative distances between tokens, making it particularly effective within a sequence.

In the Relative Encoder, two separate positional encodings are added to both the key K and the value V vectors during the self-attention calculation (Shaw, Uszkoreit, and Ashish Vaswani, 2018). Specifically, a relative positional encoding a_{ij}^K is added to the key vector K_j, and a separate relative positional encoding a_{ij}^V is added to the value vector V_j. These adjustments ensure that

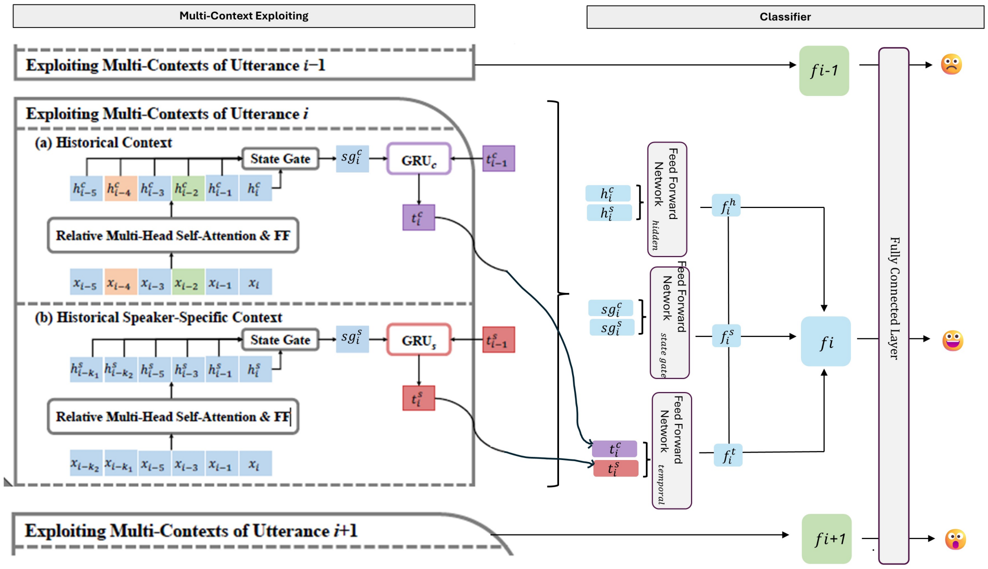

Multi-Context Exploiting
Classifier
Exploiting Multi-Contexts of Utterance i−1
Exploiting Multi-Contexts of Utterance i
(a) Historical Context
State Gate
sg_i^c
GRU_c
t_{i-1}^c
h_{i-5}^c h_{i-4}^c h_{i-3}^c h_{i-2}^c h_{i-1}^c h_i^c
t_i^c
Relative Multi-Head Self-Attention & FF
x_{i-5} x_{i-4} x_{i-3} x_{i-2} x_{i-1} x_i
(b) Historical Speaker-Specific Context
State Gate
sg_i^s
GRU_s
t_{i-1}^s
$h_{i-k_1}^s$ $h_{i-k_2}^s$ h_{i-5}^s h_{i-3}^s h_{i-1}^s h_i^s
t_i^s
Relative Multi-Head Self-Attention & FF
x_{i-k_2} x_{i-k_1} x_{i-5} x_{i-3} x_{i-1} x_i
Exploiting Multi-Contexts of Utterance i+1
h_i^c
h_i^s
Feed Forward Network hidden
sg_i^c
sg_i^s
Feed Forward Network state gate
t_i^c
t_i^s
Feed Forward Network temporal
f_i^h
f_i^s
f_i^t
fi-1
fi
fi+1
Fully Connected Layer

Figure 4.1: *Multi-context attention mechanism and classification task.*
H: Hidden states, SG: State Gate, C: Historical context, S: Speaker-specific context, T: Temporal context.

both the key and value vectors incorporate information about the relative positions of tokens.

Integration of Contexts in the Multi-Head Attention Mechanism. The process begins with the tokenisation and embedding of each utterance using the RoBERTa-bne model. After tokenisation, the embedding representations for each type of context are projected through linear transformations to obtain the queries (Q), keys (K), and values (V) required for calculating multi-head attention (Wei et al., 2023).

For the **historical context** C and **speaker-specific context** S, the attention score is calculated as follows:

$$e_{ij} = \frac{Q_i \cdot (K_j + a_{ij}^K)^T}{\sqrt{d_k}}$$

- e_{ij} represents the attention score between the i-th query and the j-th key, incorporating the relative positional encoding a_{ij}^K.
- Q_i and K_j are the query and key vectors for tokens i and j, respectively.
- a_{ij}^K is the relative positional encoding added to the key vector.
- d_k is the dimensionality of the key vectors.

Once the attention scores e_{ij} are computed for each context, the attention weights α_{ij} are calculated using softmax:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_j \exp(e_{ij})}$$

where α_{ij} represents the normalised attention weight for each token pair i, j. These attention weights determine the contribution of each token j when calculating the final context representation for token i in each context.

The final output h_i for token i is computed by summing the weighted value vectors V_j, where each value vector incorporates its relative positional encoding a_{ij}^V:

$$h_i = x_i' + \sum_j \alpha_{ij} \left(V_j + a_{ij}^V \right)$$

- x_i' serves as the representation of the translated utterance.
- V_j is the value vector for the j-th token.
- a_{ij}^V is the relative positional encoding added to the value vector.
- α_{ij} is the attention weight for the i-th and j-th token pair.

The **State Gate (SG)** processes the representations from the historical and speaker-specific contexts separately, transforming them into local states that capture relevant information from preceding utterances. For each context, the SG computes an attention-based weighted sum of the final outputs h_i calculated in the previous step, over, in this case, the last five utterances. This allows the model to focus on the most relevant historical or speaker-specific information independently for each context.

To calculate the local state sg_i for a given utterance, the State Gate first computes an intermediate score st_j for each previous h_j (where h_j represents either the historical or speaker-specific context). This is done by applying a weight matrix W, followed by a non-linear activation function:

$$st_j = \tanh \left(h_j W^T \right)$$

Here, W is a learnable weight matrix, and h_j is the hidden state at time step j for the respective context. Once the intermediate scores st_j are calculated, the attention weights β_j are obtained by applying a softmax function across the scores from the preceding five utterances (i.e., the window of interest):

$$\beta_j = \frac{\exp \left(st_j \right)}{\sum_{k=i-5}^{i-1} \exp \left(st_k \right)}$$

These weights determine how much attention is given to each hidden state within the window for each context. The final local state sg_i is then computed as a weighted sum of the hidden states:

$$sg_i = \sum_{k=i-5}^{i-1} \beta_k h_k$$

This process is carried out independently for both the historical and speaker-specific contexts, ensuring that each context's contribution to the final representation is processed separately. By integrating the SG mechanism for both historical and speaker-specific contexts, the model effectively balances the contributions from multiple contextual perspectives while maintaining sensitivity to the temporal dynamics of emotional expression in dialogues. The local states from each context are then passed through the GRU layers, which capture the distinct temporal evolution of emotions in each context.

Temporal Evolution of Contexts. The representations from the historical context t_i^c and speaker-specific context t_i^s are processed separately through GRUs to capture the temporal evolution of emotions. As illustrated in Figure 4.1, the GRU processes allow the model to capture the evolving emotional state across different utterances. This is formalised as:

$$t_i^c = \mathrm{GRU}(sg_i^c, t_{i-1}^c)$$

where t_i^c is the tracked global state for the historical context prior to u_i, and t_0^c is initialised with zero. Similarly, for the speaker-specific context:

$$t_i^s = \mathrm{GRU}(sg_i^s, t_{i-1}^s)$$

The representations of tokens in both the historical context t_i^c and the speaker-specific context t_i^s are weighted according to these attention weights, allowing the model to focus on more relevant tokens when building the overall representation for each context. These updated representations t_i^c and t_i^s are passed through a Feed Forward Network (FFN) and then used in the final emotion classification process.

4.3.3 Final Classification

Processing of Contextual Representations. The final representation F_i for each utterance is derived from three distinct components, each corresponding to different aspects of the model's processing pipeline: hidden state representations from self-attention, outputs from the State Gate, and temporal evolution captured by GRUs.

1. **Hidden State Component (f_i^h)**: This component captures the embeddings generated from the self-attention mechanism for both the historical context (h_i^c) and the speaker-specific context (h_i^s). These hidden state representations are concatenated and processed through a Feed Forward Network to obtain f_i^h:

$$f_i^h = (h_i^c \,||\, h_i^s) W_h^F$$

   - h_i^c and h_i^s are the hidden state representations from the self-attention mechanism for the historical and speaker-specific contexts, respectively.
   - W_h^F is a weight matrix learned by the FFN.

2. **State Gate Component (f_i^s)**: The State Gate processes the context-specific embeddings over the past 5 utterances. The resulting local states sg_i^c and sg_i^s for the historical and

speaker-specific contexts are concatenated and passed through a FFN to generate f_i^s:

$$f_i^s = (sg_i^c \,||\, sg_i^s)W_s^F$$

- sg_i^c and sg_i^s are the local state representations generated by the State Gate for the historical and speaker-specific contexts.
- W_s^F is the weight matrix.

3. **Temporal Component (f_i^t):** This component captures the temporal evolution of emotions across multiple utterances by using GRUs. The GRU outputs, t_i^c (historical context) and t_i^s (speaker-specific context), are concatenated and passed through a FFN to produce f_i^t:

$$f_i^t = (t_i^c \,||\, t_i^s)W_t^F$$

- t_i^c and t_i^s are the GRU outputs tracking the temporal states for the historical and speaker-specific contexts.
- W_t^F is the weight matrix.

Finally, the overall representation F_i for the utterance u_i is computed by concatenating the hidden state, local states, and temporal components:

$$F_i = f_i^h \,||\, f_i^s \,||\, f_i^t$$

This representation F_i is then used as input to the final classifier for emotion prediction.

Final Classification. Finally, the logits are passed through a softmax layer to convert the scores into probabilities for each emotional class. Logits are the scores produced by the model before they are converted into probabilities, defined as follows:

$$\text{logits} = F_i W + b$$

- F_i is the final representation obtained from the concatenation of the contexts.
- W represents the learned weights of the fully connected layer.
- b is the bias term.

The logits are then passed through a softmax layer to convert these unnormalised scores into probabilities for each emotional class:

$$P_i = \text{softmax}(\text{logits})$$

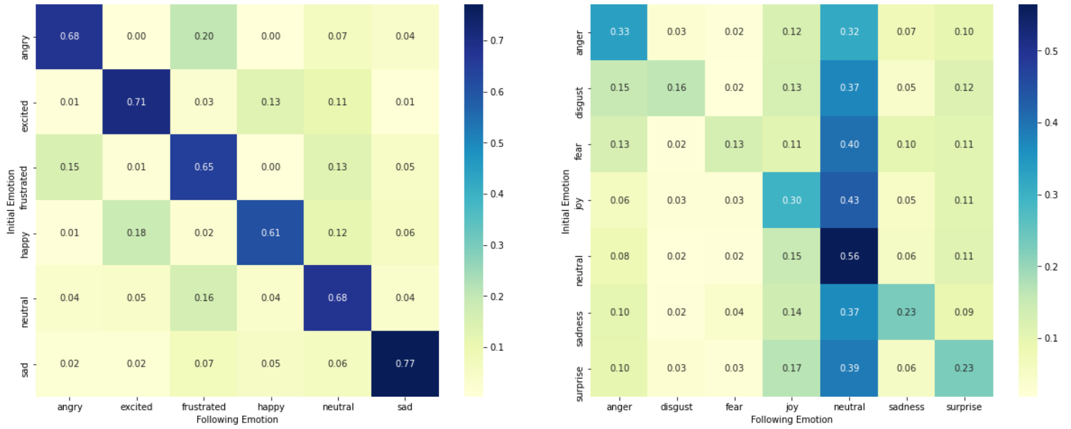
The predicted class y_i is obtained by selecting the class with the highest probability:

$$y_i = \text{argmax}(P_i)$$

4.4 Evaluation

4.4.1 Datasets

Within the domain of emotion recognition research in dialogues, it is crucial to not only ascertain the individual emotions of speakers but also to understand how these emotional states evolve and interact throughout the course of interactions. This complexity is profoundly illustrated through the examination of intra-speaker emotional transitions, which demonstrate how a speaker's emotions can dynamically influence or be influenced by subsequent emotional exchanges within the dialogue (Ghosal, Majumder, Mihalcea, et al., 2020). The graphical representations in Figure 4.2 illustrates the frequency with which specific emotions sequentially manifest within a single speaker, highlighting the pivotal role of the speaker's historical emotional context and the resultant interactive dynamics.



(a) *IEMOCAP Dataset*

(b) *MELD Dataset*

Figure 4.2: *Heatmap of intra-speaker label transition in the datasets. The colour bar represents the normalised number of transitions such that each row in the matrix adds up to 1.*

The charts for the IEMOCAP and MELD datasets, which encompass both bi-party and multi-party dialogues, offer intriguing insights into comparative emotion dynamics. The IEMOCAP dataset demonstrates a pronounced tendency towards emotional repetition, indicative of high consistency and dependency of labels within dialogues, whereas the MELD dataset exhibits more subdued patterns of these characteristics, reflecting its focus on more naturalistic interactions as opposed to the intentionally provoked emotions of IEMOCAP.

This disparity underscores the importance of considering the unique contextual frameworks of each dataset when developing and evaluating emotion recognition models. Consequently, the methodology outlined in the attention mechanisms section, which captures these nuanced label behaviours, is crucial. Such behaviours, often subtle, are not easily identified through simpler analytical approaches.

A closer examination of emotional transitions across consecutive utterances further highlights patterns of persistence or predictable shifts in emotional states during conversations. These findings reaffirm the validity of incorporating historical, temporal, and speaker-specific contexts into the design of this framework, as addressed in the methodology.

As shown in Table 3.1, the datasets were partitioned into train, test, and dev sets. The test set was used to generate all the evaluation results presented in this study, providing a reliable assessment of the model's performance.

4.4.2 Metrics

Weighted F1 Score: The weighted F1 score is used to evaluate the model's performance in this experiment, particularly designed to address class imbalances present in datasets like MELD, where the 'neutral' class is predominant. The formula is:

$$F1_{\text{weighted}} = \sum_{i=1}^{N} \left(\frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \right) \cdot w_i$$

where N is the total number of classes, Precision_i and Recall_i are the precision and recall for class i, and w_i is the weight for class i.

Precision and Recall: These metrics are computed for the model with the highest F1 performance to assess its accuracy in classifying emotional categories. Precision measures the proportion of true positive predictions among all positive predictions made by the model. Recall, in contrast, measures the model's ability to correctly identify all true positive instances in the dataset. The formulas for precision and recall are as follows:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}$$

where TP_i is the number of true positives for class i, FP_i is the number of false positives, and FN_i is the number of false negatives.

In addition, a **confusion matrix** is employed to visually assess how well the model differentiates between emotional classes, offering a clear overview of the model's predictive performance across various categories by comparing predicted and actual values.

Loss Function: During the training, the CrossEntropyLoss function is applied to the logits produced by the model, which is particularly effective for classification tasks. This function measures the divergence between the predicted probability distributions and the actual label distributions, defined as:

$$\text{CrossEntropyLoss} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

where C is the number of classes, y_i is the binary indicator (0 or 1) if class label i is the correct classification for the observation, and $\hat{y}_i$ is the predicted probability for class i. This loss is averaged over all examples in a batch, with the epoch's loss reported as a cumulative average across all batches. By minimising this divergence, the function guides the model towards more accurate predictions, enhancing its ability to differentiate between emotional classes effectively.

4.4.3 Baselines

As mentioned, this study culminates with the results of the multi-framework using context awareness, and as such, three baselines will be evaluated:

1. **Spanish Translator: Google Translator.** While in the previous chapter, the narrative and consistency of translations were qualitatively evaluated by a native Spanish speaker through the analysis of some dialogues, this chapter revisits the evaluation quantitatively using the weighted average F1-score. The results of Google Translator will be compared with MarianMT and DeepL to ascertain whether the use of more advanced AI technologies leads to more accurate translations.

2. **Utterance-Level Embeddings: RoBERTa Base Model.** A baseline is established using the RoBERTa-base model, trained on English, and compared with RoBERTa-base-bne, trained specifically on Spanish. The models' effectiveness will be assessed by comparing their weighted F1 scores in emotion recognition tasks. This comparison aims to show that a model trained on Spanish (RoBERTa-bne) is more suited to emotion classification in Spanish than one trained on English (RoBERTa base). This comparison is not intended to discredit the English model, but rather to emphasise the necessity of using models adapted to the specific language of the task.

3. **Context-Awareness: Without Any Context (RAW).** The accuracy of classifying the classes in the IEMOCAP and MELD datasets will be assessed by comparing the results of not using any context versus incorporating historical and speaker context into the multi-head attention mechanism.

4.4.4 Results

In this subsection, the results of the comparative analysis between translation methods and the performance of language-specific transformers for emotion recognition in conversations are presented; the implications of these findings are discussed, and possible explanations for the observed outcomes are explored.

To begin, the performance of three translation methods: Google Translate, MarianMT, and DeepL was evaluated without the integration of any contextual information. Table 4.1 summarises the weighted average F1 scores achieved by each method.

Although the qualitative analysis demonstrated that DeepL outperformed both MarianMT and Google Translate, Table Tables 4.1 shows that, in the absence of context, this superiority is not evident. There is little difference between the methods, and in the case of MELD, Google Translate

Table 4.1: Performance by Translation Method

Translation Method	IEMOCAP (Weighted F1)	MELD (Weighted F1)
Baseline: Google Translate	50.22	57.30
MarianMT	49.70	54.04
DeepL	51.53	55.89

even slightly surpasses DeepL. However, a crucial consideration is the incorporation of additional information, such as context, which will be discussed in the following table.

In addition to comparing the different translation methods, Table 4.2 presents the impact of context integration on the performance of each method. The weighted average F1 scores illustrate how incorporating contextual information influences the accuracy of emotion classification.

Table 4.2: Performance with and without Context Integration

		Weighted F1	
Translation Method	Context Integration	IEMOCAP	MELD
Baseline: Google Translate	RAW	50.22	57.30
Google Translate	Context C & S	57.92	57.23
MarianMT	RAW	49.70	54.04
MarianMT	Context C & S	58.64	56.40
DeepL	RAW	51.53	55.89
DeepL	Context C & S	**59.60***	**58.03***

As shown in Table 4.2 the integration of both historical and speaker-specific context (labelled "Context C & S") led to a notable improvement in performance across all translation methods. DeepL, in particular, demonstrated the most significant gains, with its weighted F1 score rising to 59.60 for IEMOCAP and 58.03 for MELD, representing a marked enhancement over its raw performance. These observations are consistent with the qualitative findings discussed previously, where DeepL outperformed both MarianMT and Google Translate in preserving emotional nuances in more complex, context-rich sentences. In IEMOCAP, context integration leads to substantial improvements due to the dataset's complexity, whereas in MELD, the gains are smaller. The hypothesis behind this will be discussed further.

Table 4.3 presents a comparison of performance between language-specific models with and without context integration. The results show that RoBERTa-base-bne, trained in Spanish, consistently outperforms the English-based RoBERTa-base when applied to Spanish translations, particularly with the inclusion of context (historical and speaker-specific). This advantage is especially evident in IEMOCAP, where the RoBERTa-bne model achieves a weighted F1 score of 59.60, highlighting the benefit of using a model aligned with the language of the dataset.

While MELD also also exhibits improvements with the RoBERTa-bne model, the performance gap between the Spanish and English-trained models is smaller, with only a 1.14-point difference.

Table 4.3: Performance Comparison Between Language-Specific Models with and without Context Integration

		Weighted F1	
Translation Method	Model Configuration	IEMOCAP	MELD
Baseline: MarianMT +	RoBERTa English RAW	45.79	54.17
MarianMT +	RoBERTa-bne Spanish RAW	49.70	54.04
MarianMT +	RoBERTa English Context C & S	53.00	55.10
MarianMT +	RoBERTa-bne Spanish Context C & S	58.64	56.40
Baseline: DeepL +	RoBERTa English RAW	47.90	55.22
DeepL +	RoBERTa-bne Spanish RAW	51.53	55.89
DeepL +	RoBERTa English Context C & S	52.10	56.89
DeepL +	RoBERTa-bne Spanish Context C & S	**59.60***	**58.03***

This may be due to the informal and culturally specific nature of MELD, which could favour the general understanding of language offered by the English-based RoBERTa. Finally, DeepL consistently outperforms other translation tools when combined with RoBERTa-bne, reinforcing the importance of high-quality translations in preserving emotional details, especially in complex and context-rich datasets.

The performance metrics underscore the model's ability to manage class imbalances, with the weighted F1-score reflecting its overall effectiveness. These metrics are evaluated for the DeepL RoBERTa-bne model with context integration. In the IEMOCAP dataset (see Table 4.4), the model performs consistently well across most emotions, particularly for well-represented classes like *'sad'*, *'neutral'*, and *'frustrated'*, where it achieves high F1-scores. Despite the *'happy'* class being less frequent, the model still demonstrates satisfactory performance.

In contrast, the MELD dataset highlights stronger performance for the *'neutral'* class, which is the most frequent, reflecting the model's accuracy in predicting dominant emotions (see Table 4.5). However, performance drops significantly for less frequent emotions such as *'disgust'* and *'fear'*. This disparity indicates that while the model is well-tuned for common classes, it struggles with less represented labels (see also Figure 3.3). When comparing the two datasets, IEMOCAP demonstrates more consistent performance across classes, whereas MELD's apparent success is heavily influenced by the prevalence of the *'neutral'* class, underscoring the importance of examining performance metrics on a class-by-class basis.

Upon examining the confusion matrices Fig 4.3, the *'frustrated'* class in the IEMOCAP dataset shows strong predictive performance, which aligns with its high frequency, as discussed earlier. However, it is often confused with *'excited'* and *'neutral'*, likely due to textual similarities. As observed in the previous analysis, the *'happy'* class remains challenging to predict, reflecting its under-representation in the dataset. In the MELD dataset, the *'neutral'* class, representing 47% of the labels, achieves a significant number of true positives, but also exhibits notable confusion with *'sadness'* and *'anger'*. Less frequent emotions, such as *'fear'* and *'disgust'*, perform poorly, consistent with the earlier discussion on class imbalance.

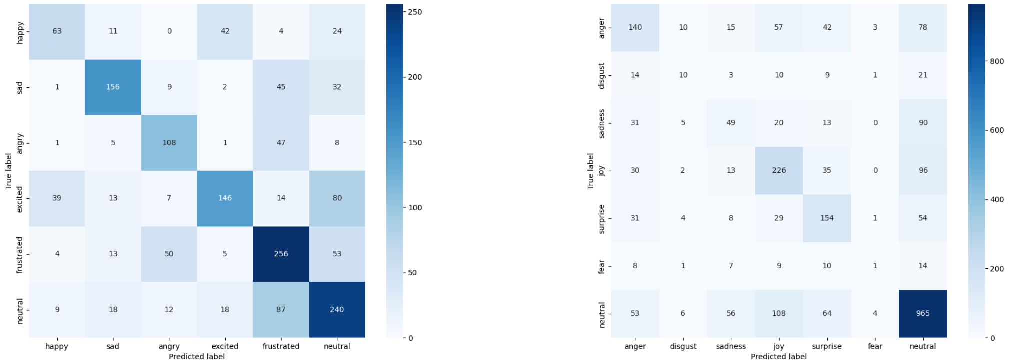
Table 4.4: IEMOCAP Performance Metrics

Emotion	Precision	Recall	F1-score
Happy	0.54	0.44	0.48
Sad	0.72	0.64	0.68
Angry	0.58	0.64	0.61
Excited	0.68	0.49	0.57
Frustrated	0.57	0.67	0.61
Neutral	0.55	0.62	0.58
Accuracy	0.60 (1623 instances)		
Macro Avg	0.61	0.58	0.59
Weighted Avg	0.61	0.60	0.60

Table 4.5: MELD Performance Metrics

Emotion	Precision	Recall	F1-score
Anger	0.46	0.41	0.43
Disgust	0.26	0.15	0.19
Sadness	0.32	0.24	0.27
Joy	0.49	0.56	0.52
Surprise	0.47	0.55	0.51
Fear	0.10	0.02	0.03
Neutral	0.73	0.77	0.75
Accuracy	0.59 (2610 instances)		
Macro Avg	0.41	0.38	0.39
Weighted Avg	0.57	0.59	0.58

Overall, the model demonstrates strong performance, particularly for well-represented classes in both datasets, providing reliable classification for the dominant emotions. However, addressing class imbalance is an area identified for future work, where potential adjustments could further enhance model performance.



(a) *IEMOCAP Dataset*

(b) *MELD Dataset*

Figure 4.3: *Confusion matrix for the test dataset using the DeepL RoBERTa-bne model with context integration.*

Configuration. The epochs and hyper-parameters were configured following the recommendations by Wei et al. (2023). For the IEMOCAP dataset, training is conducted with a batch size of 1 conversation, 4 gradient accumulation steps, and 20 epochs, using the AdamW optimiser with a learning rate of 3e-5 and a dropout rate of 0.1. For the MELD dataset, training involves a batch size of 2 conversations, 4 gradient accumulation steps, and 10 epochs, with the AdamW optimiser set to a learning rate of 1e-5 and a dropout rate of 0.1. Both models are trained on a Tesla T4 GPU, with the best checkpoint selected based on validation performance.

For the MELD dataset, the model demonstrated a steady reduction in Cross-Entropy loss, validating the choice of 10 epochs as effective for learning. In contrast, the IEMOCAP dataset exhibited fluctuating loss during intermediate epochs, eventually stabilising, which may indicate the dataset's complexity or suggest the need for further fine-tuning. For a detailed breakdown, refer to the Cross-Entropy graphs in Appendix A.1.

4.5 Summary

This chapter examines the following stages of the multilingual framework, evaluating the effectiveness of transformer-based models specifically trained on Spanish-language corpora. The RoBERTa-bne model, a variant from the MarIA family, plays a crucial role in generating embeddings that encapsulate both the semantic and emotional content of translated utterances, achieving strong results.

The analysis also highlights the importance of addressing class imbalance, especially in the MELD dataset, and using appropriate metrics, such as the weighted F1 score, to ensure a fair evaluation of the model's performance across different emotion classes.

Technically, the use of relative encoders for capturing positional dependencies and GRU networks for modelling temporal evolution allows the model to effectively integrate multiple contexts, including historical, temporal, and speaker-specific contexts, enhancing the accuracy of emotion classification. When combined with RoBERTa-bne and high-quality translation tools like DeepL, this approach achieved weighted F1 scores nearing 60% on both IEMOCAP and MELD datasets, demonstrating its success in emotion recognition in conversations.

Chapter 5

Assessment of Pre-trained Language Model in Emotion Recognition

This chapter continues the previous findings by examining further the role of context in emotion recognition within text dialogues through an ablation study. It delves into the use of prompt engineering with the advanced pre-trained language model GPT-4o, modifying the number of in-context examples and the inclusion of historical utterances to assess their impact on model performance. Additionally, the performance of GPT-4o's smaller variant is evaluated to determine its efficacy. These investigations, applied to datasets translated into Spanish by DeepL, utilise the weighted F1 score throughout the experiments.

5.1 Introduction

In the landscape of Natural Language Processing (NLP), the evolution of Large Language Models (LLMs) such as ChatGPT has introduced transformative capabilities in diverse applications, including the nuanced task of emotion recognition. Previous research has established the efficacy of foundational models like ChatGPT in discerning basic sentiments, primarily distinguishing between positive and negative emotions (Leung and Z. Xu, 2024). However, the focus of these models predominantly on the English language poses substantial challenges, restricting their utility for non-English speakers and curtailing the inclusivity of technological advancements in emotional AI interactions across diverse linguistic demographics (Y. Xu et al., 2024).

To address these limitations, this study leverages the advanced capabilities of GPT-4o (omni), released in May 2024, alongside its smaller, more recent variant introduced in July 2024, which offers improved processing speeds and cost efficiency. GPT-4o was selected for its cutting-edge generative AI capabilities and its extensive multilingual support, covering over 80 languages, making it an ideal candidate for emotion classification tasks in diverse linguistic contexts.

In contrast to traditional NLP approaches that require large amounts of task-specific fine-tuning data, humans can often learn and perform new language tasks with just a few examples or basic instructions (Brown, 2020). This gap in adaptability between human cognition and NLP systems highlights the motivation for approaches like few-shot learning, where models can be prompted with minimal examples.

This study focuses on emotion classification using Spanish-translated datasets such as IEMOCAP and MELD, with translations facilitated by DeepL. A custom Python script was developed to interact with the OpenAI API, allowing GPT-4o to be instructed via prompts to complete emotion classification tasks on the test datasets.

At the core of this experiment is *prompt engineering*, a technique for guiding pre-trained language models by designing prompts that reflect pre-training objectives (Liu et al., 2023). Unlike fine-tuning, prompt engineering allows models to adapt without altering their weights, Brown (2020) introduced *in-context learning*, where models perform tasks based on instructions within the input, without additional training. *Few-shot learning*, a form of in-context learning, provides the model with a limited set of examples to handle new tasks. Figure 5.1 contrasts in-context learning with traditional fine-tuning.

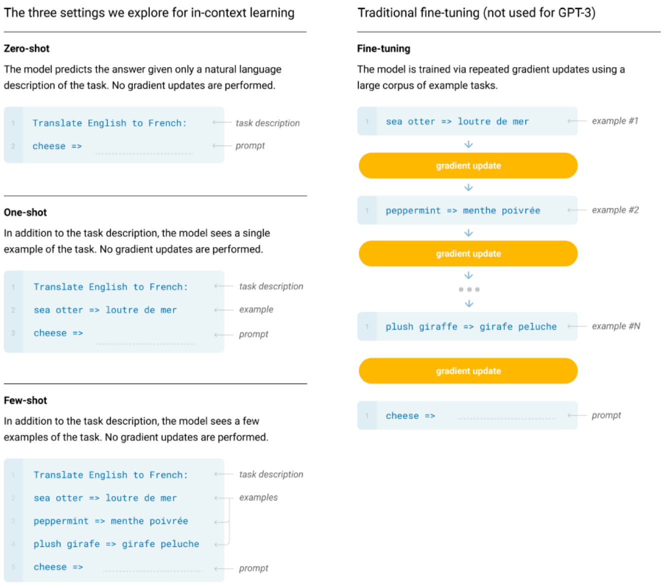

The three settings we explore for in-context learning
Zero-shot
The model predicts the answer given only a natural language description of the task. No gradient updates are performed.
Translate English to French:
cheese =>
task description
prompt
One-shot
In addition to the task description, the model sees a single example of the task. No gradient updates are performed.
Translate English to French:
sea otter => loutre de mer
cheese =>
task description
example
prompt
Few-shot
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
task description
examples
prompt
Traditional fine-tuning (not used for GPT-3)
Fine-tuning
The model is trained via repeated gradient updates using a large corpus of example tasks.
sea otter => loutre de mer
example #1
gradient update
peppermint => menthe poivrée
example #2
gradient update
plush giraffe => girafe peluche
example #N
gradient update
cheese =>
prompt

Figure 5.1: *An illustration of in-context learning and Traditional fine-tuning (Brown, 2020).*

This final experiment incorporates prompt engineering techniques, including in-context learning, to examine the model's performance across various contexts and configurations involving historical utterances. By simulating more realistic linguistic environments, this approach tests the model's adaptability without requiring extensive retraining. The study is structured around three key investigations:

1. The model's performance is evaluated using a limited number of examples in the prompt (few-shot learning) to guide emotion classification in new dialogues.

2. Historical utterances from previous conversations are incorporated into the prompt to enhance the model's understanding of emotional context in multi-turn dialogues.

3. A smaller, more resource-efficient version of the model is tested, applying both few-shot and context-aware learning techniques.

The methodology for these ablation studies and their combinations will be elaborated upon in the subsequent sections. The evaluation will primarily utilise the weighted F1 score across different configurations. Also a more detailed analysis will be provided, including class-wise metrics such as precision, recall, and F1 score.

5.2 Related work

This section reviews studies on GPT models in other languages and emotion classification. Previous research has assessed ChatGPT's general capabilities in NLP tasks, including sentiment analysis, text summarisation, and reasoning. Additionally, the performance of GPT models in machine translation has been explored (Hendy et al., 2023), while other works have identified systematic errors in ChatGPT (Borji, 2023), laying the foundation for investigating GPT's potential in multilingual emotion classification.

5.2.1 LLMs for Emotion Classification

LLMs such as GPT have demonstrated strong performance in emotion classification tasks. Recent studies have examined ChatGPT-4's zero-shot capabilities for emotion prediction, particularly in distinguishing between positive and negative states (Leung and Z. Xu, 2024). However, these studies revealed challenges in classifying more specific emotions, particularly with the multimodal MELD dataset. For instance, emotions like disgust were frequently misclassified as depression. This research builds on these findings but takes a different approach. Rather than focusing on label variations, this study evaluates a broader range of emotions and investigates how in-context learning can improve the model's ability to capture nuanced emotional states.

Furthermore, previous work has highlighted the sensitivity of GPT models to variations in emotion labels. For example, Wake et al. (2023) found that GPT-3.5 Turbo exhibited inconsistencies when slight variations in labels, such as "happy" versus "happiness," were introduced. This issue was observed across multiple datasets, including IEMOCAP, GoEmotions, and DailyDialog, indicating that GPT models, despite their strong performance in emotion classification, remain susceptible to biases stemming from subtle differences in input data. Additionally, the study here assesses whether these biases persist in the latest version of GPT, focusing on how they may impact emotion classification accuracy.

5.2.2 Prompt-Based Approach

LM-BFF, a technique introduced by Gao, Fisch, and D. Chen (2020), fine-tunes language models using a few examples with automatically searched prompts, showing great potential for text classification. PET, proposed by Schick and Schütze (2020) reformulates input examples as cloze-style phrases to assist language models in understanding tasks.

Pre-trained models have become increasingly prominent in emotion classification tasks, with GPT-4 being utilised to generate and assess new emotional labels in textual data (Arreerard and Piao, 2024). In the cited study, specific instructions were provided via prompts to guide the model's classification process. Drawing on Ekman's theory of basic emotions, the model was tasked with recognising both fundamental and more nuanced emotional states. Using a Python script to interact with the OpenAI API, GPT-4 classified emotions based on the newly crafted labels, showcasing its adaptability to diverse emotional categories without requiring extensive retraining. This methodology highlights the flexibility of pre-trained models in capturing complex emotional distinctions, while minimising the need for additional fine-tuning.

The capabilities of Large Language Models (LLMs), such as GPT-3, have also been extensively explored for tasks like machine translation (MT). In their study, Raunak, Awadalla, and Menezes (2023) investigate the role of translation quality and context manipulation in in-context learning for translations, focusing on general language processing tasks. Their approach proposes a strategy to enhance zero-shot performance by incorporating specific prompts. In contrast, while their study centres on translation quality and context manipulation, the proposed study here focuses on emotion recognition in a multilingual setting, a more challenging task due to the subtlety and complexity of emotional classification, especially when dealing with subtle emotional labels, in-context learning, and multi-turn dialogue history.

In the study by Ansari, Saxena, Ahmad, et al. (2024), in-context learning with GPT-3.5 retrieves similar context windows from the training set to provide examples that assist the model in predicting emotions. Their approach focuses on leveraging these contextual examples to support a general emotion prediction tasks. In contrast, the approach proposed here utilises context windows to capture relevant information from prior multi-turn dialogues, with a more refined focus on emotion classification in multilingual settings. Additionally, it aims to explore variations in emotional labels and incorporate historical dialogue context to enhance the model's performance in Spanish, leveraging the latest version of GPT-4.

5.3 Methodology

In this study, the capabilities of the GPT-4o model for emotion classification tasks are extended by adopting prompt engineering techniques, such as Zero-shot and Few-shot. The impact of label variations on model performance is also explored, inspired by findings from the study by Wake et al. (2023) with ChatGPT-3.5 Turbo. The methodology involved systematically testing the model's response to different contexts and historical utterances.

The experiments were conducted using a custom Python script, which interfaced with the OpenAI API to submit prompts translated from English to Spanish via DeepL to the GPT-4o model. The script also retrieved and processed classification outputs. Performance for each configuration was evaluated using the weighted F1-score, accompanied by a detailed analysis of the model's ability to classify emotions in the test dataset.

5.3.1 Few-Shot Learning

Few-Shot Learning was implemented by providing the GPT-4o model with a minimal set of labelled dialogues as reference Spanish examples. Specifically, seven dialogues were selected for each dataset to guide the model in classifying new text inputs. Let $D_i = \{u_1, u_2, \ldots, u_m\}$ represent an example dialogue, where u_j denotes the j-th utterance in the dialogue, and Class_j is its corresponding emotion label. In the Few-Shot Learning setup, the model was prompted with some example dialogues:

$$\text{Examples}_{\text{Few-Shot}} = \{(u_1, \text{Class}_1), (u_2, \text{Class}_2), \ldots, (u_m, \text{Class}_m)\}$$

where each example is a separate dialogue and does not imply any temporal sequence between them.

The model then classified a new utterance based on these isolated Spanish examples. After presenting the labelled dialogue examples, the model was instructed to classify new texts using the following prompt: *"Classify the text into one of the above classes. Provide only the class name as the output."*

For instance, one of the dialogues provided to the model included:

Text: Dios mío. ¿Qué vas a hacer? [RISAS],
Class: excited,
Text: No sé. Yo también lo solicité. Ellos no...
Class: sad

This approach was evaluated on unseen data to assess its generalisation capabilities.

5.3.2 Historical Context Integration

In this study, the model's ability to classify emotions is tested using multi-turn dialogues in Spanish, where both the number of conversations and the inclusion of historical utterances are varied. Two configurations are explored:

Five conversations with five turns each: The model is provided with 5 conversations in Spanish, each containing up to 5 utterances between participants. This setup offers a substantial amount of contextual information within each conversation, allowing the model to better understand the emotional progression. Each utterance in the dialogue is accompanied by its corresponding emotion label, which is explicitly provided to the model.

Three conversations with five turns each: In this setup, the model is given 3 conversations in Spanish, also consisting of up to 5 utterances. This configuration tests the model's performance with slightly fewer contextual cues but still includes the historical context of the previous exchanges. As in the first configuration, emotion labels are assigned to each utterance to assist the model's classification.

In both configurations, historical utterances from earlier turns in the conversation are included to enrich the model's understanding of the evolving emotional context. These utterances, along with their corresponding emotion labels, provide a more comprehensive context of the interaction, enabling the model to classify subsequent utterances with greater accuracy. The inclusion of the labels and previous utterances is structured to ensure the model can fully utilise this information. Examples of the configuration with five conversations and five context turns can be found in Appendix B in Table B.1 and Table B.2.

The **context-aware prompt** used in this experiment is designed to incorporate both the historical utterances and their corresponding emotion classes, which enhances the model's understanding of the evolving emotional dynamics in the conversation.

For three historical utterances, the context is structured as follows:

$$Context = \{(u_{t-2}, Class_{t-2}), (u_{t-1}, Class_{t-1}), (u_t, Class_t)\}$$

For five historical utterances, the context expands to include more prior exchanges:

$$Context = \{(u_{t-4}, Class_{t-4}), (u_{t-3}, Class_{t-3}), ..., (u_t, Class_t)\}$$

The objective is to classify the subsequent utterance u_{t+1} by incorporating the Spanish historical context. The model receives the following instruction for classification: *"These are examples of how texts are classified into emotions. Now, given the following context and new text, classify the new text into one of the above classes."* This instruction guides the model to base its classification on the provided examples and context.

To manage this dynamically, a sliding window approach is implemented. This mechanism selects the most recent 3 or 5 utterances from each conversation, including both the text and the corresponding emotion labels, to provide relevant contextual information for classifying the next utterance. This approach ensures that the model is always presented with the appropriate conversation history, simulating a more realistic, multi-turn dialogue interaction.

In alignment with the findings from the study *Bias in Emotion Recognition with ChatGPT*, the sensitivity of GPT-4o to variations in label names is tested. Specifically, the experiment examines whether altering the label "happy" to "happiness" in the IEMOCAP dataset affects the model's classification outcomes. This is conducted to better understand how label semantics influence the model's internal decision-making process.

5.3.3 Evaluation of GPT-4o Mini

Additionally, a smaller variant of GPT-4o, named GPT-4o Mini is assessed. This variant is evaluated using the same In-Context Learning and historical context configurations to determine its performance relative to the full-sized GPT-4o model.

5.4 Evaluation

5.4.1 Dataset

Given recent studies utilizing ChatGPT-4 (Leung and Z. Xu, 2024) the model has demonstrated robust capabilities in emotion classification, particularly when adopting Ekman's model of six universal emotions and Plutchik's wheel format. This renders the choice of the IEMOCAP and MELD datasets, which are rooted in these psychological theories, particularly apt for this study.

In this evaluation, a few labelled examples from the training set are provided as prompts to the GPT-4o API, incorporating both a few examples and some conversations with 5 utterances. The model was subsequently evaluated on the test set. The datasets are partitioned in the same manner as detailed in Table 3.1, ensuring consistency in the methodology and allowing for a fair comparison of results with previous models.

5.4.2 Metrics

Building upon the observations noted in previous chapters, particularly regarding the presence of the *neutral* state in the MELD dataset, it becomes essential to employ a classification metric such as the weighted F1-score. For the model that achieves the highest weighted F1-score, performance is further evaluated by examining precision, recall, and F1-score for each class, as well as calculating the global metrics of weighted and macro F1-scores.

The weighted F1-score accounts for the proportion of each class in the dataset, giving more weight to the larger classes In contrast, the macro F1-score calculates the F1-score independently for each class and then averages them, treating all classes equally regardless of their size.

5.4.3 Baselines

This assessment of pre-trained models are compared using the weighted F1 score against the following baselines:

1. **RoBERTa-bne Spanish.** The results from the various configurations described in the methodology will be benchmarked against the best-performing model from the previous chapter. This model utilised DeepL's translation combined with the RoBERTa-bne transformer, which was enhanced through context-awareness techniques in the attention mechanism.

2. **Examples and Context-Awareness:** The datasets are evaluated by comparing the model's performance without any examples or context in the prompt to the results when incorporating in-context examples and context-awareness within both the prompt and the programming logic.

5.4.4 Results

In this subsection, the outcomes of the comparative analysis are presented and discussed, focusing on the findings from the final model using RoBERTa-bne, translated with DeepL, as the baseline. These results are compared against the in-context ablation studies outlined in the Methodology.

1. No Prompt Engineering: A baseline classification using GPT-4o in a zero-shot setting.
2. Few Customised Dialogues: Incorporating a few personalised dialogues specific to each dataset within the GPT-4o prompt.
3. Three Conversations with Five Turns: In this setup, the prompt includes three conversations, each with five turns. Additionally, a sliding window mechanism is implemented programmatically, using three historical utterances as context in the GPT-4o prompt.
4. Six Conversations with Five Turns: The prompt is extended to include six conversations, each with five turns, with the sliding window mechanism set to three historical utterances in the GPT-4o prompt.
5. Six Conversations with Five Turns and Five Historical Utterances: This configuration builds on the previous setup but extends the sliding window to include five historical utterances within the GPT-4o prompt.
6. Six Conversations with Five Turns using GPT-4o Mini: The same setup as the previous study, but using the GPT-4o Mini variant, with a sliding window of five historical utterances in the prompt.
7. Label Change from 'Happy' to 'Happiness' in IEMOCAP: This study investigates the impact of changing the label from 'Happy' to 'Happiness' in the IEMOCAP dataset. The model is tested with six conversations, each containing five turns, alongside a sliding window mechanism of five historical utterances in the GPT-4o prompt.

In Table 5.1 it can be observed the results of experiments 1 through 5 and compare them against each other. To begin with, in both datasets, the inclusion of isolated examples (Few-Shot) within the GPT-4o prompt, treated as a subset of each dataset, does not significantly enhance the precision of emotion classification. However, for IEMOCAP, the inclusion of context within the utterances proves to be highly beneficial, resulting in a 10-point increase in the weighted F1 score. For MELD, the gap is less pronounced, as it does not require extensive contextual examples to achieve a solid performance, maintaining a weighted F1 score above 61.

When compared to the previous best-performing model using transformers, the RoBERTa-bne model, IEMOCAP performs better with this model than with the in-context GPT-4o approach, particularly when considering more conversations and contextual information **59.59** vs 53.12, respectively. On the other hand, for MELD, the use of prompt engineering with GPT-4o outperforms the RoBERTa-bne model, achieving a metric of **63.25** compared to 58.03.

The underlying reason why MELD does not require extensive examples to obtain a good F1 score, and why its results remain stable, could be attributed to the fact that these interactions are more natural and everyday in nature, with emotions that are more explicit and direct. As a result, the

Table 5.1: Performance Comparison Between DeepL Spanish Translations with RoBERTa-bne and GPT-4o

		Weighted F1	
Translation Method	Model Configuration	IEMOCAP	MELD
Baseline: DeepL Spanish +	RoBERTa BNE Spanish C & S	**59.59***	58.03
DeepL Spanish +	GPT-4o Zero-Shot	39.87	61.09
DeepL Spanish +	GPT-4o Few-Shot	41.00	61.35
DeepL Spanish +	GPT-4o + 3 context convos + 3 hist utte	51.59	62.44
DeepL Spanish +	GPT-4o + 6 context convos + 3 hist utte	52.70	62.89
DeepL Spanish +	GPT-4o + 6 context convos + 5 hist utte	53.12	**63.25***

model can classify emotions with minimal information. In contrast, IEMOCAP features scripted and fictional dialogues, with more complex emotions influenced by the conversational context, making the model more reliant on prior interactions.

In Table 5.2 and Table 5.3, the performance metrics for the IEMOCAP and MELD datasets are presented, showing the model's effectiveness across each emotion label.

Table 5.2: IEMOCAP Performance Metrics (GPT-4o In-Context + 6 convos + 5 hist utte)

Emotion	Precision	Recall	F1-score
Angry	0.55	0.42	0.47
Excited	0.74	0.43	0.54
Frustrated	0.49	0.63	0.55
Happy	0.45	0.31	0.37
Neutral	0.44	0.60	0.50
Sad	0.72	0.60	0.66
Accuracy	0.53 (1623 instances)		
Macro Avg	0.57	0.50	0.52
Weighted Avg	0.56	0.53	0.53

Table 5.3: MELD Performance Metrics (GPT-4o In-Context + 6 convos + 5 hist utte)

Emotion	Precision	Recall	F1-score
Anger	0.53	0.63	0.58
Disgust	0.39	0.38	0.39
Fear	0.26	0.40	0.31
Joy	0.56	0.57	0.57
Neutral	0.78	0.70	0.74
Sadness	0.59	0.45	0.51
Surprise	0.45	0.58	0.51
Accuracy	0.63 (2610 instances)		
Macro Avg	0.51	0.53	0.52
Weighted Avg	0.65	0.63	0.63

For the IEMOCAP dataset, the model continues to face challenges in accurately identifying the 'happiness' emotion, which is reflected in a lower F1-score. Despite these challenges, the model generally performs consistently across other emotions, showing stable metrics. This suggests that while the model can manage the dataset's complexity to some extent, there are still specific areas, particularly with less frequent emotions, where performance could be improved.

In the case of the MELD dataset, the model shows difficulty in classifying 'disgust' and 'fear,' which are emotions that appear less frequently and may be less explicitly expressed in the dialogues. Although the model struggles with these negative emotions, it performs well in identifying more common and straightforward emotions like 'neutral'. This discrepancy may indicate that the examples provided during prompt engineering were not sufficient to fully capture the nuances of these less frequent emotions. In both datasets, the model exhibits similar challenges when dealing with underrepresented emotions, reflecting the same difficulties previously encountered with the DeepL RoBERTa-bne model using context.

Given that prompt engineering relies on the examples provided rather than a full training process, these results suggest that increasing the number and variety of example could further enhance the model's performance.

The cost-effectiveness of GPT-4o Mini is an important consideration for this emotion classification. With an input cost of $0.150 per million tokens compared to GPT-4o's $5.00, GPT-4o Mini is 97% cheaper for input tokens and 96% cheaper for output tokens. However, this cost reduction comes with a slight decrease in emotion classification performance. As shown in Table 5.4, using GPT-4o Mini results in a sacrifice of approximately 5 points in the weighted F1 score for the IEMOCAP dataset and 3 points for the MELD dataset. While these differences may seem small, they could impact the overall precision depending on the specific requirements of the emotion classification task. Nonetheless, given the substantial advantages in terms of cost and processing speed, GPT-4o Mini remains a viable alternative.

Table 5.4: Performance Comparison Between GPT-4o and GPT-4o Mini

		Weighted F1	
Translation Method	Model Configuration	IEMOCAP	MELD
DeepL Spanish +	GPT-4o 6 context convos + 5 hist utterances	53.12	63.25
DeepL Spanish +	GPT-4o Mini 6 context convos + 5 hist utterances	47.30	60.32

As the final ablation study, the label "happy" is changed to "happiness" in the IEMOCAP dataset. This adjustment was motivated by the sensitivity to label variations identified by Wake et al. (2023) in their study with GPT-3.5-turbo. Despite this modification, the weighted F1 score remained at 52.64, suggesting that the sensitivity observed in earlier versions has likely been mitigated in this latest version, indicating that the previously identified bias has been addressed.

Upon closer examination of the Table 5.5, it is clear that the performance for the "happiness" label is nearly identical to that of the "happy" label (0.31 vs. 0.33 in F1 score). The minor variations observed in other classes likely reflect the probabilistic nature of this large language model (LLM) rather than the change in the class label itself. Furthermore, the Spanish language context of the dataset does not appear to have influenced this outcome.

Table 5.5: Performance Metrics for the IEMOCAP Dataset with ”Happiness” Label

Emotion	Precision	Recall	F1-score	Support
Angry	0.51	0.42	0.46	170
Excited	0.73	0.43	0.55	299
Frustrated	0.50	0.62	0.55	381
Happiness	0.46	0.26	0.33	144
Neutral	0.43	0.59	0.50	384
Sad	0.71	0.62	0.66	245
Accuracy		0.53 (1623 instances)		
Macro Avg	0.56	0.49	0.51	1623
Weighted Avg	0.55	0.53	0.53	1623

5.5 Summary

This chapter provides a thorough evaluation of the GPT-4o model for emotion recognition within Spanish text dialogues, highlighting the significance of contextual prompts in enhancing accuracy, particularly with datasets such as IEMOCAP. The ablation studies conducted illustrate the impact of various prompt engineering techniques, including in-context learning and the incorporation of historical utterances.

The findings reveal that while GPT-4o demonstrates consistent performance, its accuracy can be further enhanced by adapting the model to the specific characteristics of the dataset, including linguistic features and emotion labels. Furthermore, the assessment of the smaller, cost-effective GPT-4o Mini variant indicates a minimal trade-off in performance, underscoring its viability due to significant cost advantages. The study also confirms that biases identified in earlier versions, such as GPT-3.5 Turbo, have been effectively addressed in this latest iteration.

These results suggest that GPT-4o is a practical alternative for working with datasets translated into Spanish, where it has proven effective in capturing colloquial expressions and linguistic nuances, as evidenced by the MELD dataset. Although cost considerations remain pertinent, the use of a pre-trained LLM like GPT-4o can streamline the process and yield substantial performance improvements, positioning it as a robust candidate for emotion recognition tasks, particularly in comparison to the intricate process of fine-tuning transformers for more sophisticated context-awareness.

Chapter 6

Conclusion

This chapter concludes the thesis by summarising the key findings from the studies conducted, from the proposed multilingual framework to the assessment of pre-trained LLMs. It highlights the main achievements and acknowledges the limitations encountered. The chapter also offers recommendations for future research directions that could build upon the work presented here.

6.1 Summary

This research addresses the challenges associated with emotion recognition in Spanish, particularly due to the scarcity of available datasets. Rather than creating new datasets from scratch, which is both time-consuming and expensive, this study opts for the careful translation of existing datasets, such as IEMOCAP and MELD, which are originally in English and widely recognised for their universal emotion classification based on the psychological theories. The translation was conducted using three different methods, ranging from traditional approaches, such as Google Translator, to more advanced Neural Machine Translation systems like MarianMT and DeepL. Among these, DeepL proved to be the most effective tool for preserving colloquial nuances.

While other studies have explored the use of context within transformers, these efforts have focused exclusively on English. This multilingual framework introduces a novel application of the RoBERTa-bne transformer, an advanced model specifically trained on a Spanish corpus. It incorporates a GRU mechanism to capture temporal dependencies, a relative encoder to handle the positional relationships between words in sequences, and an attention mechanism that includes historical context and speaker-specific information to enhance the accuracy of emotion classification. These elements are particularly important for accurately capturing the emotional subtleties in dialogues.

Recognising emotions based solely on text is a challenging task due to issues such as sarcasm or the absence of explicit emotional expressions. Despite these challenges, the proposed framework achieved promising results, reaching a F1 score close to 60%. In the pursuit of reducing reliance on large language models and avoiding the need for massive Spanish datasets or models with millions of parameters, more accessible yet advanced models like GPT-4o were also explored. This model was evaluated using prompt engineering techniques that manage context in various configurations of Spanish dialogues. The best results were obtained when dialogues with context and historical context were considered. The results should be analysed separately: in the case of MELD, a more natural dataset, GPT-4o effectively captured the inherent meaning and emotion in Spanish, achieving higher metrics than the RoBERTa-bne transformer, with an F1 score exceeding 63%. In contrast, the proposed framework performed better in IEMOCAP. Similarly, the mini version of GPT-4o performed well in MELD, offering a cost-effective option without a significant loss in accuracy.

In conclusion, the multilingual framework with contextual awareness presented encouraging results, suggesting its extension to other languages in text dialogues when sufficient data is not available. However, if budget allows, the use of GPT-4o with context-based prompts, or its mini version to reduce costs without significantly sacrificing accuracy, is also a viable option. The combination of advanced AI tools and thoughtful integration of context promises to extend these methodologies to other languages, thereby broadening the applicability and impact of emotion recognition technologies in multilingual contexts.

6.2 Limitations

While this research achieved promising results, it is important to acknowledge certain limitations. Firstly, the evaluation was primarily based on static text inputs, which may not fully capture the dynamic nature of emotions as they occur in real-world interactions. Emotions can change rapidly, and the absence of real-time feedback mechanisms in the current approach may limit the model's ability to adapt to these fluctuations.

Another limitation concerns the datasets used. For instance, the IEMOCAP dataset consists of simulated dialogues which, while useful for controlled studies, may not fully represent the complexity of emotions in natural conversations.

Furthermore, the use of advanced tools such as DeepL and GPT-4o was crucial for achieving high accuracy but comes with significant costs. The limitation on the number of conversations processed in the case of GPT-4o during prompt engineering was influenced by financial considerations, while the technical constraint of implementing pauses between API requests to avoid service blocks affected the processing time for translations.

6.3 Future Work

This research paves the way for various avenues of future exploration and development. One promising direction is the application of the multilingual framework to other languages. This would involve assessing the framework's effectiveness from the initial translation of English datasets to the classification of emotions in dialogues translated into different linguistic contexts, thereby offering a broader understanding of the framework's versatility and robustness.

To address the under-representation of certain negative connotation classes within the MELD dataset, new examples could be generated using generative AI techniques, such as Generative Adversarial Networks (GANs). GANs could generate synthetic emotional dialogue samples for under-represented emotions like anger, disgust, or fear; the generator creates these examples, and the discriminator evaluates their authenticity. These synthetic examples could then be used to enrich both the multilingual transformer model and the in-context learning prompts, ensuring more balanced data representation and potentially improving classification accuracy for these challenging emotion categories.

Further, there is significant potential in exploring alternative large language models, such as Meta's LLaMa 3.1 or Google's Gemini, to refine the process of emotion recognition. Comparing the performance of these models with that of GPT-4o, particularly in multilingual contexts, could provide valuable insights into their respective strengths and weaknesses in multilingual emotion recognition tasks.

Given that GPT-4o functions as a probabilistic model, another avenue for future work could involve running the same experiments multiple times and averaging the outcomes to gain a more comprehensive understanding of the model's variability in performance.

Moreover, expanding this multilingual framework to incorporate other modalities, such as audio and video, represents a crucial next step. Future research could focus on integrating voice input, allowing users to express emotions not only through text but also through spoken interactions. This would enhance the system's ability to interpret non-verbal cues and discern deeper emotional states, especially given GPT-4o's capability to process and generate text, audio, and visual data. Such an expansion would significantly enhance the framework's applicability in real-world multimodal settings.

Bibliography

Alzubaidi, L., J. Zhang, A.J. Humaidi, et al. (2021). “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions”. In: *Journal of Big Data* 8, pp. 1–74.

Ansari, Mohammed Abbas, Chandni Saxena, Tanvir Ahmad, et al. (2024). “JMI at SemEval 2024 Task 3: Two-step approach for multimodal ECAC using in-context learning with GPT and instruction-tuned Llama models”. In: *arXiv preprint arXiv:2403.04798*.

Arnold, MB (1960). “Emotion and personality”. In: *Vol. I/Psychological aspects*.

Arreerard, Ratchakrit and Scott Piao (June 2024). “Exploring GPT-4 for Fine-Grained Emotion Classification”. English. In: The 7th Healthcare Text Analytics Conference 2024, HealTAC2024 ; Conference date: 13-06-2024 Through 14-06-2024. URL: `https://healtac2024.github.io/`.

Baevski, Alexei et al. (2020). “wav2vec 2.0: A framework for self-supervised learning of speech representations”. In: *Advances in neural information processing systems* 33, pp. 12449–12460.

Barrett, Lisa Feldman (2017). *How emotions are made: The secret life of the brain*. Pan Macmillan.

Bhardwaj, Shivendra et al. (2020). “Human or Neural Translation?” In: *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6553–6564.

Borji, Ali (2023). “A categorical archive of chatgpt failures”. In: *arXiv preprint arXiv:2302.03494*.

Brown, Tom B (2020). “Language models are few-shot learners”. In: *arXiv preprint arXiv:2005.14165*.

Busso, Carlos et al. (2008). “IEMOCAP: Interactive emotional dyadic motion capture database”. In: *Language resources and evaluation* 42, pp. 335–359.

Chen, Tianqi and Carlos Guestrin (2016). “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785–794.

Darwin, Charles (1993). “The expression of the emotions in man and animals (1872)”. In: *The Portable Darwin*, pp. 364–393.

De la Rosa, Javier et al. (2022). “Bertin: Efficient pre-training of a spanish language model using perplexity sampling”. In: *arXiv preprint arXiv:2207.06814*.

Deng, Jiawen and Fuji Ren (2023). “A Survey of Textual Emotion Recognition and Its Challenges”. In: *IEEE Transactions on Affective Computing* 14.1, pp. 49–67. DOI: `10.1109/TAFFC.2021.3053275`.

Dhall, Abhinav et al. (2011). “Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark”. In: *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, pp. 2106–2112.

Douglas-Cowie, Ellen et al. (2011). “The HUMAINE database”. In: *Emotion-oriented systems: The Humaine handbook*, pp. 243–284.

Ekman, Paul and Wallace V Friesen (1971). "Constants across cultures in the face and emotion." In: *Journal of personality and social psychology* 17.2, p. 124.

El español: una lengua viva (2019). `https://cvc.cervantes.es/lengua/espanol_lengua_viva/pdf/espanol_lengua_viva_2019.pdf`. Instituto Cervantes.

Ezzameli, Kaouther and Hela Mahersia (2023). "Emotion recognition from unimodal to multimodal analysis: A review". In: *Information Fusion* 99, p. 101847.

Fan, Angela et al. (2021). "Beyond english-centric multilingual machine translation". In: *Journal of Machine Learning Research* 22.107, pp. 1–48.

Gao, Tianyu, Adam Fisch, and Danqi Chen (2020). "Making pre-trained language models better few-shot learners". In: *arXiv preprint arXiv:2012.15723.*

Garcia-Cuesta, Esteban, Antonio Barba Salvador, and Diego Gachet Pãez (2024). "EmoMatchSpanishDB: study of speech emotion recognition machine learning models in a new Spanish elicited database". In: *Multimedia Tools and Applications* 83.5, pp. 13093–13112.

Ghosal, Deepanway, Navonil Majumder, Rada Mihalcea, et al. (2020). "Utterance-level dialogue understanding: An empirical study". In: *arXiv preprint arXiv:2009.13902.*

Ghosal, Deepanway, Navonil Majumder, Soujanya Poria, et al. (2019). "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation". In: *arXiv preprint arXiv:1908.11540.*

Gu, Xin, Yinghua Shen, and Jie Xu (2021). "Multimodal emotion recognition in deep learning: a survey". In: *2021 International Conference on Culture-oriented Science & Technology (ICCST).* IEEE, pp. 77–82.

He, Kaiming et al. (2016). "Deep residual learning for image recognition". In: *CVPR*, pp. 770–778.

Hendy, Amr et al. (2023). "How good are gpt models at machine translation? a comprehensive evaluation". In: *arXiv preprint arXiv:2302.09210.*

Hidalgo-Ternero, Carlos Manuel (2020). "Google Translate vs. DeepL: analysing neural machine translation performance under the challenge of phraseological variation". In: *Análisis multidisciplinar del fenómeno de la variación fraseológica en traducción e interpretación / Multidisciplinary Analysis of the Phenomenon of Phraseological Variation in Translation and Interpreting.* Ed. by Pedro Mogorrón Huerta. MonTI Special Issue 6, pp. 154–177.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Ishiwatari, Taichi et al. (2020). "Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations". In: *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 7360–7370.

Izard, Carroll E (1977). "Human Emotions,(1977)". In: *DOI: http://dx. doi. org/10.1007/978-1-4899-2209-0.*

Jiao, Wenxiang et al. (2019). "Higru: Hierarchical gated recurrent units for utterance-level emotion recognition". In: *arXiv preprint arXiv:1904.04446*.

Junczys-Dowmunt, Marcin et al. (2018). "Marian: Fast neural machine translation in C++". In: *arXiv preprint arXiv:1804.00344*.

Kim, Yoon (2014). "Convolutional neural networks for sentence classification". In: *EMNLP*, pp. 1746–1751.

Lazarus, Richard S (1991). *Emotion and adaptation.* Vol. 557. Oxford University Press.

LeCun, Yann et al. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

Leung, Clement and Zhifei Xu (2024). "Large Language Models for Emotion Evolution Prediction". In: *International Conference on Computational Science and Its Applications.* Springer, pp. 3–19.

Liang, Lin and Shasha Wang (2022). "Spanish Emotion Recognition Method Based on Cross-Cultural Perspective". In: *Frontiers in psychology* 13, p. 849083.

Liu, Pengfei et al. (2023). "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing". In: *ACM Computing Surveys* 55.9, pp. 1–35.

Luo, Jiachen, Huy Phan, and Joshua Reiss (2023). "Fine-tuned RoBERTa Model with a CNN-LSTM Network for Conversational Emotion Recognition". In.

Majumder, Navonil et al. (2019). "Dialoguernn: An attentive rnn for emotion detection in conversations". In: *Proceedings of the AAAI conference on artificial intelligence.* Vol. 33. 01, pp. 6818–6825.

McKeown, Gary et al. (2011). "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent". In: *IEEE transactions on affective computing* 3.1, pp. 5–17.

Medhat, Walaa, Ahmed Hassan, and Hoda Korashy (2014). "Sentiment analysis algorithms and applications: A survey". In: *Ain Shams engineering journal* 5.4, pp. 1093–1113.

Mehrabian, Albert (1974). "An approach to environmental psychology". In: *Massachusetts Institute of Technology.*

Mim, Afsana Haque (2023). "The Importance of Multilingualism in a Globalized World: Understanding multilingualism". In: *Inverge Journal of Social Sciences* 2.2, pp. 84–91.

Minaee, Shervin et al. (2024). "Large language models: A survey". In: *arXiv preprint arXiv:2402.06196.*

Mohammad, Saif M. and Peter D. Turney (2013). "Crowdsourcing a Word-Emotion Association Lexicon". In: *Computational Intelligence* 29.3, pp. 436–465.

Ng, Nathan et al. (2019). "Facebook FAIR's WMT19 news translation task submission". In: *arXiv preprint arXiv:1907.06616.*

Pan, Ronghao et al. (2024). "Spanish MEACorpus 2023: A multimodal speech–text corpus for emotion analysis in Spanish from natural environments". In: *Computer Standards & Interfaces* 90, p. 103856.

Peng, Sancheng et al. (2022). "A survey on deep learning for textual emotion analysis in social networks". In: *Digital Communications and Networks* 8.5, pp. 745–762.

Plutchik, Robert (1982). *A psychoevolutionary theory of emotions.*

Poria, Soujanya, Erik Cambria, et al. (2017). "Context-dependent sentiment analysis in user-generated videos". In: *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 873–883.

Poria, Soujanya, Devamanyu Hazarika, et al. (2018). "Meld: A multimodal multi-party dataset for emotion recognition in conversations". In: *arXiv preprint arXiv:1810.02508.*

Rangel, Ismael Díaz, Grigori Sidorov, and Sergio Suárez Guerra (2014). "Creación y evaluación de un diccionario marcado con emociones y ponderado para el español". In: *Onomazein* 29, pp. 31–46.

Raunak, Vikas, Hany Hassan Awadalla, and Arul Menezes (2023). "Dissecting in-context learning of translations in gpts". In: *arXiv preprint arXiv:2310.15987.*

Ringeval, Fabien et al. (2013). "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions". In: *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG).* IEEE, pp. 1–8.

Russell, James A (1980). "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6, p. 1161.

Schick, Timo and Hinrich Schütze (2020). "Exploiting cloze questions for few shot text classification and natural language inference". In: *arXiv preprint arXiv:2001.07676.*

Shaw, Peter, Jakob Uszkoreit, and Ashish Vaswani (2018). "Self-attention with relative position representations". In: *arXiv preprint arXiv:1803.02155.*

Sidorov, Grigori et al. (2013). "Empirical study of machine learning based approach for opinion mining in tweets". In: *Advances in Artificial Intelligence: 11th Mexican International Conference on Artificial Intelligence, MICAI 2012, San Luis Potosí, Mexico, October 27–November 4, 2012. Revised Selected Papers, Part I 11.* Springer, pp. 1–14.

Szegedy, Christian et al. (2015). "Going deeper with convolutions". In: *CVPR*, pp. 1–9.

Tacconi, David et al. (2008). "Activity and emotion recognition to support early diagnosis of psychiatric diseases". In: *2008 second international conference on pervasive computing technologies for healthcare.* IEEE, pp. 100–102.

Tan, Zhixing et al. (2020). "Neural machine translation: A review of methods, resources, and tools". In: *AI Open* 1, pp. 5–21.

Tiedemann, Jörg and Santhosh Thottingal (2020). "OPUS-MT–building open translation services for the world". In: *Proceedings of the 22nd annual conference of the European Association for Machine Translation*, pp. 479–480.

Troiano, Enrica, Roman Klinger, and Sebastian Padó (2020). "Lost in back-translation: Emotion preservation in neural machine translation". In: *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4340–4354.

Vapnik, Vladimir (1995). "Support-vector networks". In: *Machine learning* 20, pp. 273–297.

Vaswani, A (2017). "Attention is all you need". In: *Advances in Neural Information Processing Systems*.

Wake, Naoki et al. (2023). "Bias in Emotion Recognition with ChatGPT". In: *arXiv preprint arXiv:2310.11753*.

Wei, Yinyi et al. (2023). "Exploiting pseudo future contexts for emotion recognition in conversations". In: *International Conference on Advanced Data Mining and Applications*. Springer, pp. 309–323.

Xu, Yuemei et al. (2024). "A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias". In: *arXiv preprint arXiv:2404.00929*.

Zadeh, Amir et al. (2020). "CMU-MOSEAS: A multimodal language dataset for Spanish, Portuguese, German and French". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*. Vol. 2020. NIH Public Access, p. 1801.

Zadeh, AmirAli Bagher et al. (2018). "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246.

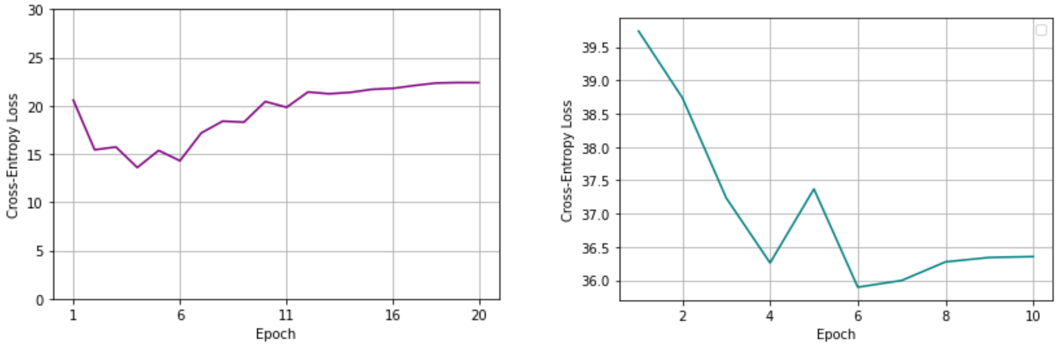
Zhang, Duzhen et al. (2020). "Knowledge aware emotion recognition in textual conversations via multi-task incremental transformer". In: *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4429–4440.

Zhang, Jianhua et al. (2020). "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review". In: *Information Fusion* 59, pp. 103–126.

Zhong, Peixiang, Di Wang, and Chunyan Miao (2019). "Knowledge-enriched transformer for emotion detection in textual conversations". In: *arXiv preprint arXiv:1909.10681*.

Appendix A

Cross Entropy Loss Metric



(a) *IEMOCAP Dataset*

(b) *MELD Dataset*

Figure A.1: *Cross-Entropy Loss across epochs on the dev set for the DeepL translation with RoBERTa-bne and context integration. The loss is calculated as the accumulated sum of the average loss per batch.*

For the MELD dataset, the model demonstrated a steady reduction in Cross-Entropy loss. In contrast, the IEMOCAP dataset exhibited fluctuating loss during intermediate epochs, eventually stabilizing, which may indicate the dataset's complexity.

Appendix B

Prompt for GPT4-o

Examples of the configuration with five conversations and five context turns added in the prompt.

Table B.1: Examples of Text Classification Emotions from IEMOCAP

Class	Text
Conversation 1	
excited	Oye, qué hora es? Esto se acerca a la medianoche, ¿verdad? Dios, esto es genial, ¿no?
excited	Mira la noche que tuvimos. No cambiaría esto por nada.
excited	Sabes, en realidad quería ir un poco más lejos en la costa y alejarme de todas las luces y la gente, pero tenía miedo de que te lo perdieras. ¿Qué tal te va?
excited	Es, es, nah, es solo Paul? Ni siquiera puedo decirlo.
Conversation 2	
sad	Así que te vas mañana.
sad	Sí. Acaban de llamar.
sad	Esto es realmente injusto.
sad	Lo sé.
Conversation 3	
angry	Eso está fuera de control.
angry	No entiendo por qué es tan complicado para la gente cuando llega aquí. Es un simple formulario. Solo necesito una identificación.
frustrated	¿Cuánto tiempo llevas trabajando aquí?
Conversation 4	
excited	Tú... No puedo creerlo. Estoy tan feliz por ti. Esto es exactamente lo que querías. Es tu sueño hecho realidad.
happy	Gracias, señor.
excited	El tipo exacto, lo apruebo totalmente. Es un tipo maravilloso. Lo pasamos muy bien. Sabe beber. Eso es genial.
happy	[RISAS]
Conversation 5	
neutral	Eso es útil.
frustrated	Creo que ya no puedo hacer esto, ha pasado mucho tiempo y no es como si no lo intentara.
neutral	Bueno, supongo que no lo estás intentando lo suficiente.
frustrated	Lo intento. Han pasado tres años.
Conversation 6	
frustrated	Oh. No seas tan grandilocuente. Solo porque resulta que no quieres uno en este momento.
angry	No seas estúpido.
angry	De verdad, Amanda.
angry	¿Cómo?
frustrated	Nada.

Table B.2: Examples of Text Classification Emotions from MELD

Class	Text
Conversation 1	
disgust	¿Qué te pasa?
neutral	Nada!
fear	Bueno, me dio un dolor cegador en el estómago cuando estaba levantando pesas, luego me desmayé y no he podido levantarme desde entonces.
neutral	Pero no creo que sea nada serio.
surprise	Esto suena como una hernia. ¡Tienes que ir al médico!
fear	No puede ser! Kay mira, si tengo que ir al médico por algo va a ser por esta cosa que me sale del estómago!
Conversation 2	
joy	¡Me encanta tu casa! ¿De dónde es este tipo?
neutral	Uh eso es un artefacto indio del siglo XVIII de Calcuta.
surprise	¡Vaya! Así que sois más que dinosaurios.
neutral	Mucho más.
Conversation 3	
neutral	Ross, ¿puedo hablar contigo un minuto?
neutral	Sí, por favor! Entonces, ¿qué pasa?
sadness	Uh, bueno... Joey y yo rompimos.
surprise	¡Dios mío, qué ha pasado?
sadness	Joey es un gran tipo, pero somos... tan diferentes! Quiero decir, durante tu discurso no paraba de reírse del homo erectus!
anger	¡Sabía que era él!
sadness	De todos modos, creo que es lo mejor.
Conversation 4	
disgust	¿Qué asco! ¿Qué ha sido eso? Algo ha explotado!
neutral	Solo ha roto aguas. Calmate, quieres?
surprise	Romper aguas, ¿qué quieres decir? ¿Qué es eso de romper aguas?
neutral	Respira, respira, respira.
Conversation 5	
anger	¡Por favor!
anger	¿Me tomas el pelo?
anger	Herí a tres hombres enormes, le hice sangrar la nariz a un tipo... no estoy orgulloso de ello, pero de verdad que lo estoy.
joy	Y todo es gracias a ti, maravilloso, increíble tú.
fear	Creo que tienes conmoción cerebral.
Conversation 6	
joy	¿Qué sí! Quieres una pista? Eh? ”Si quiero” ”Si quiero”.
neutral	Ok, estoy sintiendo que esto es algún tipo de juego de palabras, porque eres de color rosa con alegría apenas controlada.
joy	David va a proponerle matrimonio a Phoebe.
surprise	¿Por qué? ¿Por qué?