Explainable Emotion Classification in Social Media

by

Muhammad Hammad Fahim Siddiqui

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Computer Science

e © Muhammad Hammad Fahim Siddiqui, Ottawa, Canada, 2025

Examining Committee

The following served on the Examining Committee for this thesis.

External Member: Héctor Gibrán Ceballos-Cancino
Professor, Institute for the Future of Education
Tecnológico de Monterrey

Carleton Member: Ahmed El-Roby
Associate Professor, School of Computer Science
Carleton Univeristy

Internal Member(s): Mehrdad Sabetzadeh
Professor, School of Electrical Engineering & Computer Science
University of Ottawa

Paula Branco
Assistant Professor, School of Electrical Engineering & Computer Science
University of Ottawa

Supervisor(s): Diana Inkpen
Professor, School of Electrical Engineering & Computer Science
University of Ottawa
and
Alexander Gelbukh
Instituto Politécnico Nacional, Mexico

Declaration of Authorship

I hereby certify that this thesis is entirely my own original work except where otherwise indicated. I am aware of the University of Ottawa regulations concerning plagiarism, including those regarding consequent disciplinary actions. Any use of the works of any other author, in any form, is properly acknowledged at their point of use.

Abstract

In the age of internet social media, people express their emotions openly online, making accurate emotion classification an increasingly important yet challenging task. Unlike sentiment analysis, which focuses on the polarity of opinions, emotion analysis dives deeper to identify specific emotions within text. Emotion classification has several practical applications, such as brand perception analysis, crisis management, and processing customer feedback, which are vital for businesses, governments, and organizations to engage effectively with their audiences.

This thesis explores both multi-label and multi-class emotion classification and its explainability, specifically in tweets. For multi-label classification, we utilized the SemEval 2018 task E-c dataset, which comprises 11 distinct emotions, and for multi-class classification, we employed the DAIR AI emotion dataset. Through comprehensive experiments, we demonstrated robust model performance across both classification tasks, underscoring the adaptability and effectiveness of our approach.

Our research uses instruction-based fine-tuning of large language models (LLMs) like GPT-2 and experiments with zero-shot and dynamic few-shot classification using GPT-4o, LLaMA 3 (8B), and DeepSeek R1 (Distilled Qwen 32B). Building on previous baseline performance, we further improved emotion detection and model interpretability by developing a self-explaining model that uses generative explanations and preference alignment. Notably, we constructed a novel preference alignment dataset using GPT-4o with chain-of-thought prompting, where human annotators assessed model outputs for correctness, clarity, helpfulness, and verbosity. Utilizing this dataset, we preference-aligned GPT-4o via Direct Preference Optimization (DPO) and open-source models, including LLaMA 3 (8B) and DeepSeek R1 (Distilled Qwen 32B), using Odds Ratio Preference Optimization

(ORPO).

The resulting self-explaining models achieved state-of-the-art multilabel classification performance (68.85%) on the SemEval 2018 E-c dataset and competitive accuracy (93.1%) on the DAIR AI multiclass dataset. Furthermore, the explanations generated by our models exceeded existing explainability techniques, achieving a higher sufficiency score of 63.66%, reflecting better interpretability and alignment with human expectations. We also performed detailed human evaluations of the generated explanations, demonstrating that explanations produced by our preference-aligned models significantly surpass those from pre-trained models.

Acknowledgements

I always thought that my achievements in life were mine alone—that I am a self-made man—but nothing could be farther from the truth. Behind every step I have taken, there has been a network of support and encouragement, that has carried me through challenges and helped me reach my goals. And today, I want to thank all of you.

First and foremost, I would like to express my deepest gratitude to my parents. Their tireless efforts and unwavering dedication ensured that I received the best possible education in my home country and a nurturing upbringing. Their sacrifices and support have been the foundation upon which I built this academic journey.

I am profoundly thankful to my brothers, whose support throughout this process was invaluable. By shouldering family responsibilities, they enabled me to focus fully on my studies and research without distraction.

I extend my sincere thanks to my supervisor and co-supervisor for their continuous mentorship, insightful guidance, and diligent oversight. Their knowledge transfer and dedication ensured that this thesis was completed on time and to the highest scientific standards. I am also grateful to my thesis advisory committee for their invaluable feedback, which strengthened the substance, value, and innovation of my work.

To my extended family in the United States, Pakistan and the UAE, I owe a special thanks. My family in Pakistan and UAE have always been my greatest cheerleaders, celebrating every success and milestone. My family in the United States has been a home away from home during my time in Canada, offering warmth, hospitality, and a social space that greatly contributed to my well-being.

And to that one special person who has been alongside me through a significant part of my academic journey—who always made sure I received the push, the motivation, the zeal, and the purpose to excel in life—thank you. I owe you big time.

Last but certainly not least, I thank my friends both in Pakistan and Canada. Their encouragement during difficult times and readiness to help in any capacity have been a source of strength and motivation, helping me achieve the goals I set for myself.

And to all who helped, I owe you more than just citations.

To all of you, I am forever grateful.

Table of Contents

List of Tables **xii**

List of Figures **xiv**

1 Introduction **1**

1.1 Overview 1

1.2 Motivation 2

1.2.1 Understanding public opinion 3

1.2.2 Crisis management and early warning systems 3

1.2.3 Brand perception and customer feedback 3

1.2.4 Research and personalization 4

1.3 Research Questions 5

1.4 Contributions 6

1.4.1 Transformer-based models 6

1.4.2 Instruction tuning of GPT-2 6

1.4.3 Zero-shot detection with OpenAI GPT models 7

1.4.4 Few-shot and dynamic few-shot detection with OpenAI GPT models 7

1.4.5 Explainable AI 7

1.5 Publications 8

1.6 Outline of the Thesis 9

1.7 Summary 12

2 Background and Related Work **14**

2.1 Background: Theory of Emotions . . . 14

2.1.1 Emotion theory . . . 15

2.1.2 What are emotions? . . . 16

2.1.3 The significance and complexity of emotion detection in research . . 19

2.2 Related Work . . . 27

2.2.1 Emotion classification . . . 28

2.2.2 Explainable AI . . . 46

2.3 Summary . . . 50

3 Data **51**

3.1 SemEval 2018 Affect in Tweets Dataset . . . 51

3.1.1 Dataset composition . . . 52

3.1.2 Dataset statistics . . . 52

3.2 Twitter Emotions Corpus (TEC) . . . 53

3.2.1 Dataset composition . . . 56

3.2.2 Dataset statistics . . . 56

3.3 Dair AI . . . 57

3.3.1 Dataset composition . . . 58

3.3.2 Dataset statistics . . . 59

3.4 ISEAR . . . 59

3.4.1 Dataset composition . . . 60

3.4.2 Dataset statistics . . . 61

3.5 Affective Text . . . 61

3.5.1 Dataset composition . . . 62

3.5.2 Dataset statistics . . . 62

3.6 Chosen Datasets . . . 63

3.7 Summary . . . 64

4 Methodology 66
4.1 Task Description . . . 66
4.2 Evaluation Measures . . . 67
4.2.1 Evaluation measures - Multi-class . . . 67
4.2.2 Evaluation measures - Multi-label . . . 67
4.2.3 Evaluation measure - Explanations . . . 70
4.3 Transformer Models for Emotion Classification . . . 73
4.3.1 Hyperparameter settings . . . 73
4.4 Instruction prompts . . . 74
4.4.1 Prompt structure . . . 74
4.4.2 Fine-tuning process . . . 75
4.4.3 Inference mechanism . . . 75
4.4.4 Experimental settings . . . 76
4.5 Zero-shot Classification . . . 77
4.6 Few-shot Classification . . . 77
4.6.1 Vanilla few-shot . . . 77
4.6.2 Dynamic few-shot . . . 79
4.7 Explainable AI . . . 82
4.8 Post-hoc Explainers . . . 82
4.9 Generative AI Explanations Using Pre-trained Models . . . 83
4.10 Self-explaining Model . . . 84
4.10.1 What is a self-explaining model, and why do we need one? . . . 84
4.10.2 What should explanations/output look like? . . . 85
4.10.3 How? Pre-trained LLMs . . . 87
4.10.4 Making LLAMA and DeepSeek as good as GPT-4o . . . 88
4.10.5 How? Supervised fine-tuning . . . 88
4.10.6 How? Alignment . . . 89

4.10.7 Preference dataset for alignment 90
4.10.8 Methodology used for alignment 97
4.11 Architecture Diagram 102
4.12 Summary 105

5 Evaluation 106
5.1 Classification results 107
5.1.1 Transformer-based models 107
5.1.2 Instruction tuning prompts 107
5.1.3 Zero-shot 108
5.1.4 Few-shot, dynamic few-shot and self-explaining models 108
5.1.5 Multi-label results 111
5.1.6 Multi-class results 113
5.2 Explainable AI Evaluation 114
5.2.1 Evaluating LIME, SHAP and Generative AI Explanations using the automatic measure 114
5.2.2 Self-explaining model 115
5.2.3 Automatic evaluation discussion 118
5.2.4 Human evaluation 120
5.2.5 Adherence to structure 126
5.2.6 Cost and complexity vs. performance 127
5.3 Summary 129

6 Conclusion, Limitations and Future Work 130
6.1 Conclusion 130
6.2 Summary of Contributions 132
6.2.1 Publications, datasets, and models 132
6.2.2 Answering the research questions 133
6.3 Limitations 135
6.4 Future Research Directions 137

References **139**

List of Tables

3.1 Example of multi-label text for emotion classification 53

3.2 Dataset distribution statistics in Train, Dev, Test sets. Also includes the percentage of presence of each respective emotion in the dataset. 54

3.3 Example of multi-class tweets for emotion classification from the TEC dataset 56

3.4 Details of the Twitter Emotion Corpus 57

3.5 Example of multi-class tweets for emotion classification from the Dair AI dataset 58

3.6 Details of the Dair AI Corpus 59

3.7 Example of multi-class experiences for emotion classification from the ISEAR dataset 60

3.8 Details of the ISEAR Corpus 61

3.9 Example of multi-class tweets for emotion classification from the Affective Text dataset 62

4.1 Example prompts for fine-tuning and testing setups. Similar prompts are generated for each emotion. 76

4.2 The used zero-shot, few-shot, and system prompts. 78

4.3 Examples of explanations generated through different explainable AI models 84

4.4 Output example from the self-explaining model 86

4.5 Example of annotated response pairs from our preference dataset 93

4.6 Krippendorff's Alpha values indicating reliability score across all explanation quality dimensions for both responses 95

5.1 Emotion classification results on different models for SemEval-2018 dataset. The best values are in bold. 109

5.2 Emotion classification results on different models for Dair AI emotion dataset. The best values are in bold. 110

5.3 Sufficiency metric for evaluating explanations for the models 120

5.4 Average human evaluation scores (on a 5-point Likert scale) for explanation quality across four dimensions. Evaluations were performed on 300 randomly sampled tweets from the test set. 124

List of Figures

1.1 Logical flow of thesis chapters and their interconnections 13

2.1 The Plutchik's wheel of emotions 17

3.1 Heatmap illustrating the correlation of emotions in the SemEval 2018 dataset. 55

4.1 System architecture diagram for self-explaining model 104

5.1 Human Evaluation Comparison of GPT Models 121
5.2 Human Evaluation Comparison of DeepSeek Models 122
5.3 Human Evaluation Comparison of LLaMA Models 123
5.4 Model Output Adherence to JSON-style Structure 127

Chapter 1

Introduction

1.1 Overview

In recent years, the rapid growth of social media platforms has completely changed how people communicate. Social media has created a new era where individuals, groups, and organizations can easily and quickly share their opinions, emotions, and thoughts. This shift has not only changed how we connect with each other but also how we share information and express emotions in a digital world.

With the rise of social media, people are sharing their emotions online more openly than ever. This creates a challenge: how can we effectively manage and analyze the emotional content in these large volumes of text? Emotion classification is a complex task, made even harder by the informal and evolving nature of online language, including slang, memes, and cultural references. Unlike sentiment analysis, which is widely studied area that focuses on the polarity of opinions, emotion analysis dives deeper to identify specific emotions expressed within texts.

This research is important in creating effective emotion classification models specifically designed for social media text. Accurately identifying emotions can help businesses, governments, and organizations better understand public feelings, make informed decisions, and engage more meaningfully with their audiences.

In this thesis, we explore the use of instruction-tuning, zero-shot, few-shot, and dynamic few-shot inference, and preference alignment of large language models (LLMs) for emotion classification and explainability in social media, specifically focusing on tweets. We construct a novel preference dataset using GPT-4o with chain-of-thought prompting, which allows us to create self-explaining models capable of providing accurate emotion predictions along with clear and contextually grounded explanations. We compare our generative explanations against traditional explainability methods such as LIME and SHAP, quantitatively evaluating the quality of the explanation through the sufficiency metric. Additionally, we conduct detailed human evaluations to assess the correctness, clarity, helpfulness, and verbosity of the generated LLM explanations, demonstrating that preference-aligned models significantly enhance both interpretability and classification accuracy compared to pre-trained and baseline approaches.

1.2 Motivation

Social media has become a significant part of how people communicate, with users often sharing their emotions and opinions. Understanding these emotions is valuable for various reasons, including social, commercial, and psychological purposes. Unlike sentiment analysis, which only identifies whether a message is positive, negative, or neutral, emotion detection focuses on feelings like joy, anger, or sadness. This deeper understanding is essential in many areas. Below, we will explore some key uses of emotion classification and explainability in social media.

1.2.1 Understanding public opinion

Emotion classification helps us better understand public opinion, especially in politics, social issues, and major events. As conversations on these topics unfold online, identifying the emotions behind them allows policymakers, analysts, and marketers to adapt to changing trends. This helps them engage more effectively with the public and adjust their strategies to meet new needs. For example, research on public opinion during the COVID-19 pandemic used emotion analysis to track evolving attitudes over time. By analyzing emotions like fear, hope, and anger, decision-makers gained valuable insights into how people reacted to public health measures and other interventions [106].

1.2.2 Crisis management and early warning systems

In crisis management, emotion detection is instrumental in assessing public reaction and the effectiveness of organizational responses. For example, the Integrated Crisis Mapping (ICM) model maps crises based on organizational engagement and public coping strategies, emphasizing the importance of emotional analysis in crisis situations [55]. Moreover, employing emotion detection models, such as transformer-based and dictionary-based approaches, has proven effective in monitoring social media during crises, facilitating timely and appropriate interventions [24].

1.2.3 Brand perception and customer feedback

Understanding emotions in customer feedback is essential for evaluating brand perception. Studies have shown that emotions significantly influence consumer behavior and loyalty. For example, research into the emotional aspects of brand perception highlights how positive associations can enhance customer loyalty and willingness to pay premium prices [59]. Additionally, analyzing customer feelings through emotion detection enables businesses to tailor their strategies, improving customer satisfaction and brand reputation [13].

1.2.4 Research and personalization

Emotion classification is an essential tool in various research areas. It helps researchers understand trends and patterns in emotional responses, which provides insight into human behavior. It also improves personalized experiences by adjusting content based on a user's emotions.

Emotion classification on social media is instrumental in understanding public opinion, managing crises, improving brand perception, and supporting social research. It enables decision-makers to respond more effectively by providing insights into the emotional landscape expressed online. Recently, generative AI explanations have gained heightened attention due to their potential to make the predictions of emotion classification models more transparent and interpretable. Since generative AI can produce human-readable rationales, transparency becomes crucial in ensuring reliable outputs and fostering user trust. As detailed by Kim et al. [58], hybrid deep learning models demonstrate how interpretability significantly enhances the understanding of emotional responses in practical scenarios, such as social media interactions or customer feedback analys'is. The need for explainability thus directly correlates with the effectiveness of emotion detection systems, as interpretable outputs empower stakeholders to grasp the underlying rationale, thereby augmenting the reliability and accountability of model decisions [95]. Consequently, this thesis addresses both emotion classification and explainability, presenting practical solutions that leverage generative explanations to clarify how and why models arrive at their predictions, underscoring their value in real-world applications.

1.3 Research Questions

This research studies the application of AI models in Natural Language Processing to contribute a solution for emotion classification and explainability in social media content (tweets). We aim to answer the following research questions:

1. **Enhancing Large Language Models for Emotion Classification:**
    - RQ1: How can LLMs and alignment be leveraged to improve the performance of emotion classification models, and if synthetic data augmentation can be used to address class imbalance in the datasets?
2. **Explainable AI:**
    - RQ2: How do generative AI explanations and traditional statistical post-hoc methods help interpret the decisions of black-box models for emotion classification?
    - RQ3: How can a self-explaining model for emotion classification be developed using LLMs to classify emotions and provide meaningful, concept-based explanations?
    - RQ4: How can alignment techniques improve the quality of AI-generated explanations and ensure they align with human expectations, and how can we quantitatively and qualitatively evaluate them?

To answer RQ1, we make the key contributions outlined in sections: 1.4.1, 1.4.2, 1.4.3, and 1.4.4, while for RQ2, RQ3, and RQ4, we detail the key contributions in 1.4.5.

1.4 Contributions

This thesis makes important contributions through the use of LLMs, instruction-tuning, dynamic few-shot learning, and explainability. The following are the key contributions of this research.

1.4.1 Transformer-based models

We used BERT-based transformer models with multiple attention mechanisms for emotion classification. These models fine-tuned on the datasets, improved the accuracy of emotion classification compared to traditional methods. Using their strong contextual understanding of language, we performed better than the LSTM and Bi-LSTM models from related work in detecting a range of emotions in tweets. This work will be our baseline and was done before the PhD.

1.4.2 Instruction tuning of GPT-2

A key novel contribution of this thesis is the introduction of instruction tuning for fine-tuning large language models [99], specifically GPT-2. By designing instruction prompts to guide the model, we significantly improved its accuracy in classifying emotions. This approach set a new state-of-the-art in multi-label emotion classification, surpassing previous benchmarks by a notable margin.

1.4.3 Zero-shot detection with OpenAI GPT models

In addition to instruction tuning, we conducted zero-shot classification experiments using GPT-4o. These models were tested without any task-specific training, and their performance was compared to our fine-tuned GPT-2 model. Although the zero-shot models performed well, our instruction-tuned GPT-2 model performed better, demonstrating the benefits of fine-tuning for this specific task.

1.4.4 Few-shot and dynamic few-shot detection with OpenAI GPT models

We conducted few-shot emotion classification experiments using GPT-4o, initially yielding suboptimal performance. To address this limitation, we explored a dynamic few-shot approach, in which we dynamically selected the most relevant examples based on the similarity of tweet embeddings to the target tweet. We experimented extensively with different embedding strategies and demonstrated that dynamic few-shot prompting consistently outperformed standard few-shot and zero-shot approaches. This approach is novel for the task at hand.

1.4.5 Explainable AI

1.4.5.1 Comparing generative AI explanations to statistical techniques

Another contribution of this research is the focus on explainability. We compared the effectiveness of different explainability techniques, including LIME, SHAP, and generative AI explanations using GPT-4o. These experiments provided valuable insights into how the models make predictions, offering a better understanding of the decision-making process. Using the sufficiency metric, we evaluated the quality and usefulness of these explanations, contributing to the growing field of explainable AI.

1.4.5.2 Self-explaining model

We introduce a self-explaining model for emotion classification, designed to predict emotions present in tweets and generate explanations simultaneously. Leveraging generative LLMs like GPT-4o, LLAMA, and DeepSeek, our approach provides accurate emotion predictions (both single-label and multi-label) and clear, informative, and contextually grounded explanations. We constructed a novel preference alignment dataset to improve these explanations by generating pairs of candidate outputs through GPT-4o using chain-of-thought prompting, followed by expert annotation based on correctness, clarity, helpfulness, and verbosity. The resulting annotated dataset, utilized with preference optimization techniques, represents a unique resource that significantly improves smaller models and introduces state-of-the-art self-explaining emotion classification capabilities. The self-explaining framework and the preference-alignment dataset are novel contributions, opening avenues for further research in interpretable emotion classification and the alignment of large language models with human preferences.

1.5 Publications

Throughout the PhD, peer-reviewed research papers were produced directly from the research done in this dissertation:

- Instruction Tuning of LLMs for Multi-label Emotion Classification in Social Media

Content – This paper explored the effectiveness of instruction-tuned large language models for multi-label emotion detection, setting new benchmarks on established datasets. In Proceedings of Canadian AI 2024.

- Towards Interpretable Emotion Classification: Evaluating LIME, SHAP, and Generative AI for Decision Explanations – This work critically evaluated popular post-hoc explanation methods alongside generative AI-based explanations, highlighting the trade-offs in clarity, correctness, and usability. In Proceedings of Information Visualization (IV) 2024.

- Self-explaining Emotion Classification through Preference-Aligned Large Language Models – This paper introduced the concept of self-explaining models for emotion classification and demonstrated how preference-alignment can guide LLMs to generate structured outputs with high interpretability. In submission.

Additionally, one paper was published during the PhD that builds upon work initially conducted during the author's Master's degree:

- Multi-label Emotion Classification in Texts Using Transfer Learning – while the foundation of this work 1.4.1 originated in the author's previous academic degree, it is included here as it addresses the same research problem and uses one of the datasets, contributing foundational insights relevant to this thesis. Expert Systems with Applications 2023: 213.

1.6 Outline of the Thesis

The rest of this thesis is organized as follows.

Chapter 2: Background and Related Work

- Overview of emotion theory and emotion categories.
- Review of models such as Ekman's Basic Emotions Theory and Plutchik's Wheel of Emotions.
- Discussion of the role of emotions in human behavior and the significance of emotion detection in various fields.
- Review of methods for emotion classification, including supervised learning and transformer-based models.
- Exploration of instruction-tuning prompts and zero-shot, few-shot classification using LLMs.
- Discussion of explainable AI techniques such as LIME, SHAP, and generative AI explanations.
- Discussion of self-explaining models, alignment techniques and preference alignment datasets.

Chapter 3: Data

- Datasets
  - SemEval 2018 Affect in Tweets Dataset
  - Twitter Emotions Corpus (TEC)
  - DAIR AI Dataset
  - ISEAR dataset

    - Affective Text Dataset

- Chosen datasets for this study

Chapter 4: Methodology

- Transformer models for multi-label emotion classification
- Instruction tuning and fine-tuning process
- Zero-shot classification using GPT models
- Few-shot classification using GPT models
- Dynamic few-shot classification using GPT models
- Explainable AI methods (LIME, SHAP, Generative AI)
- Alignment and preference datasets
- Preference dataset generation in our work
- Self-explaining models using preference alignment
- Architecture diagram

Chapter 5: Evaluation

- Presentation of results from multi-label and multi-class emotion classification tasks.
- Discussion of explainability experiments comparing LIME, SHAP, GPT, and preference-aligned explanations.

- Evaluation of explanation quality using the sufficiency metric.
- Evaluation of explanation quality using human judges.

Chapter 6: Conclusion, Limitations, and Future Work

- Summary of key findings and their implications.
- Limitations of our work.
- Proposal for future work, including reward modeling.

Figure 1.1 illustrates the overall structure and logical flow of the thesis, showing how each chapter builds upon the previous ones and how the core research components are interconnected.

1.7 Summary

In this chapter, we introduced the core motivations for studying emotion classification in social media, highlighting its importance for applications such as public opinion analysis, crisis detection, and personalized communication. We discussed the challenges associated with multi-label emotion classification and the need for explainable AI to enhance model transparency and user trust. The key research questions guiding this thesis were presented, along with the primary contributions made during the research, including the development of instruction-tuned and self-explaining models, the construction of a preference-aligned dataset, and a comparative evaluation of different explanation techniques.

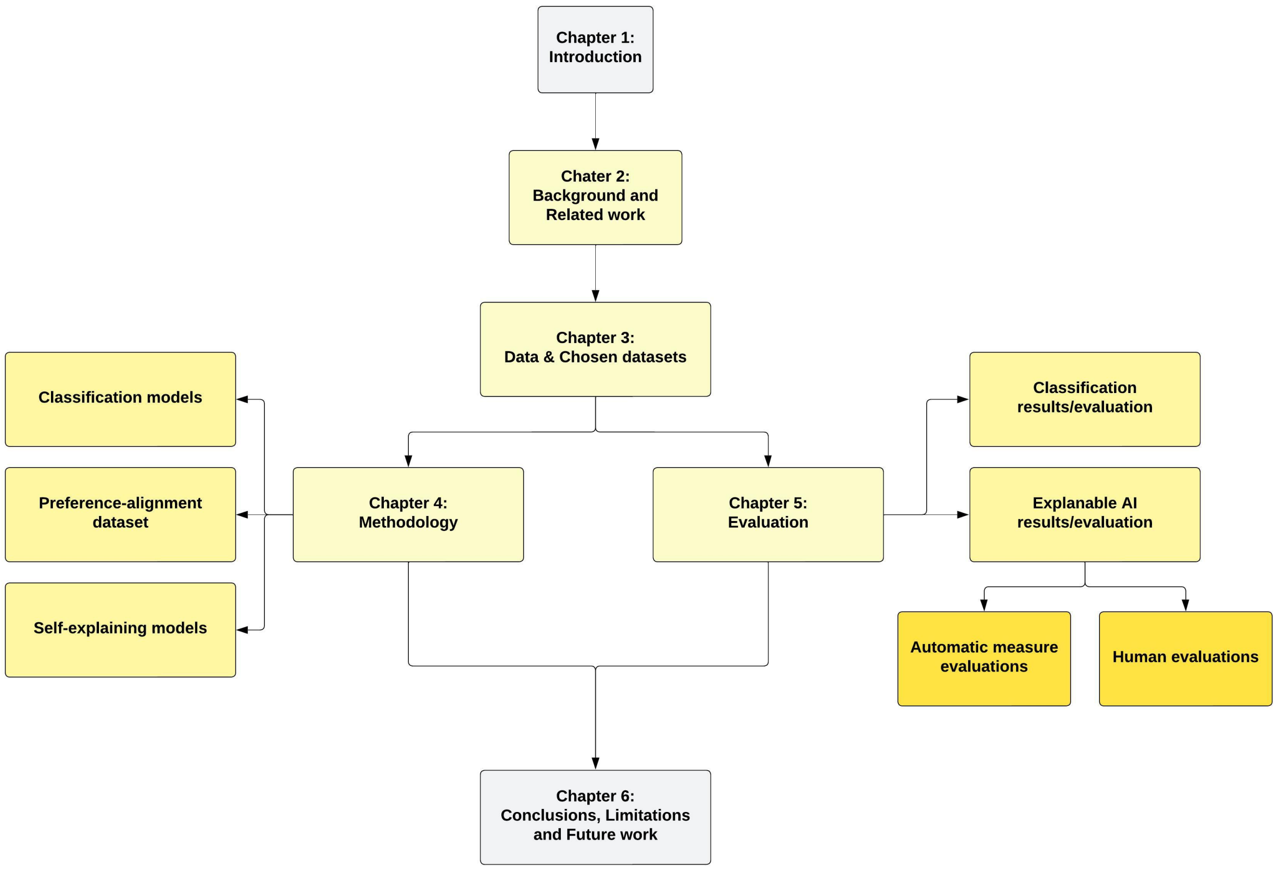

Figure 1.1: Logical flow of thesis chapters and their interconnections

In the next chapter, we will review the theoretical background on emotions, existing approaches to emotion detection, and prior work on explainability in AI, laying the groundwork for the methodologies proposed in this thesis.

Chapter 2

Background and Related Work

2.1 Background: Theory of Emotions

Emotions are psychological states containing various feelings, thoughts, and behaviors. They are crucial to human experiences and integral to personal and social interactions. The study of emotions spans multiple disciplines, including psychology, neuroscience, linguistics, and artificial intelligence. In recent years, the intersection of linguistics and natural language processing (NLP) has opened new avenues for research in emotion classification, enabling machines to understand and interpret human emotions through language.

This chapter will explore the theoretical perspectives of emotions, beginning with a detailed examination of what emotions are and how they are defined. We will go into the underlying processes and mechanisms that give rise to emotional experiences. Finally, we will investigate the causes of emotions, considering internal and external factors that influence emotional responses.

2.1.1 Emotion theory

Research on emotions has evolved through three primary approaches [117]: the categories approach, the dimensions approach, and the cognitive appraisals approach. The categories-based approach, which groups emotions into discrete categories based on similarity, is the most extensively researched area in Natural Language Processing (NLP). This approach enables researchers to understand emotions through multi-class, multi-label, and intensity measurements of emotions. Two widely used sources of emotions in the NLP space are Ekman's basic emotions [31] and Plutchik's wheel of emotions [85]. Ekman's model identifies six basic emotions, while Plutchik's wheel offers a more nuanced and comprehensive representation of emotional states, instrumental in guiding emotion classification research. Paul Ekman's theory identifies six basic emotions that are universally recognized: happiness, sadness, fear, anger, surprise, and disgust. These emotions are characterized by distinct facial expressions that are consistent across cultures.

- Happiness: Joy and contentment, expressed through smiling and laughter.
- Sadness: Feelings of loss and sorrow, shown by frowning and crying.
- Fear: Anxiety and apprehension, marked by widened eyes and a faster heartbeat.
- Anger: Hostility and irritation, displayed through frowning and a clenched jaw.
- Surprise: Shock and amazement, expressed by raised eyebrows and an open mouth.
- Disgust: Revulsion and aversion, indicated by a wrinkled nose and raised upper lip.

On the other hand, Robert Plutchik's Wheel of Emotions is a comprehensive model that illustrates the complexity and interrelation of human emotions. Plutchik proposed that there are eight primary emotions, each paired with an opposite:

- Joy vs. Sadness
- Trust vs. Disgust
- Fear vs. Anger
- Surprise vs. Anticipation

These primary emotions can combine to form more complex emotions, creating a nuanced spectrum of emotional experiences. The wheel is structured in a circular arrangement, demonstrating how emotions can blend and vary in intensity. For example:

- Joy and Trust combine to create Love.
- Fear and Surprise can lead to Alarm.
- Sadness and Disgust might result in Remorse.

The model emphasizes emotions' evolutionary purpose, suggesting that they are adaptive responses to environmental stimuli that have evolved. The Plutchiks wheel of emotion is shown in the figure 2.1.

2.1.2 What are emotions?

Emotions are complex psychological states critical in human behavior, influencing our thoughts, actions, and interactions. The study of emotions spans various disciplines, including psychology, neuroscience, philosophy, and anthropology, each providing unique insights into their nature and function. Paul Ekman's foundational work (1992) [31] is pivotal in studying emotions, particularly his theory of basic emotions. Ekman proposed that

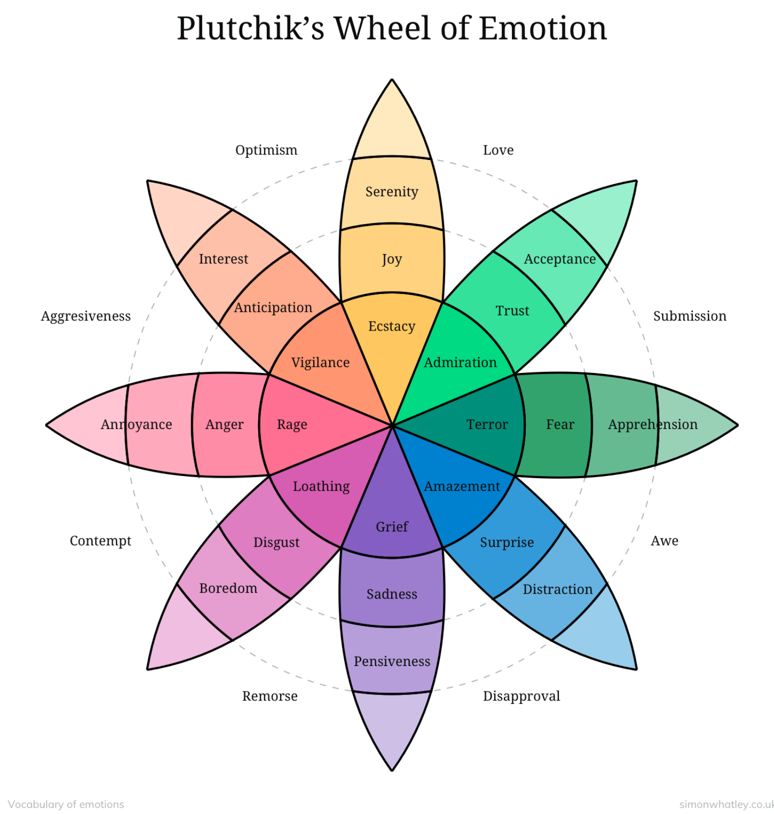

Figure 2.1: The Plutchik's wheel of emotions

certain emotions—happiness, sadness, fear, anger, surprise, and disgust—are universal and biologically innate, evidenced by distinct facial expressions recognized across cultures. Ekman's research underscores the evolutionary significance of these emotions, suggesting that they are essential for human survival by preparing individuals to respond to environmental challenges. Expanding on this biological perspective, J Prinz [86] argues that emotions are gut reactions that have evolved to serve adaptive functions. Prinz posits that emotions are not unique to humans but are shared across species, indicating a deep evolutionary root. This view supports the idea that emotions are fundamental to the human experience, providing quick, automatic responses to critical situations. Complementing these evolutionary perspectives, Mordka [75] delves into the structure and function of emotions, emphasizing their non-cognitive aspects. Mordka differentiates emotions from sensory and cognitive activities, suggesting that they are distinct states that do not necessarily involve rational thought. This perspective highlights emotions' automatic and sometimes irrational nature, which can arise independently of conscious deliberation. In contrast, Reisenzein [93] addresses the definitional challenges of emotions, noting the absence of a universally accepted definition. Reisenzein emphasizes that emotions are complex mental-behavioral processes that require precise conceptualization. He advocates for a nuanced understanding of emotions that considers their mental and physical components. Beatty [12] explores the cultural context of emotions, arguing that social and cultural factors influence their meaning and expression. Beatty's anthropological perspective reveals that emotions are biological phenomena and social constructs shaped by cultural norms and practices. This view suggests that while certain emotions may be universal, their expression and significance can vary widely across different societies. Scherer [98] contributes to understanding emotions by investigating their measurement and the methods to quantify emotional experiences. Scherer's work provides a framework for understanding the dimensions and intensity of emotions, facilitating the development of tools to measure emotional responses accurately. Frijda [**?**] offers an information-processing perspective, suggesting that emotions arise as responses to significant events. Frijda's model integrates emotions into the broader context of human cognition, proposing that emotions prepare individuals for action by signaling important environmental changes. This view aligns with the evolutionary perspective, highlighting the adaptive functions of emotions in human behavior. Finally, Deonna and Teroni [26] provide a philosophical introduction to emotions, discussing various theories explaining emotions and how they relate to other mental states. Their work highlights the complexity and diversity of emotional experiences, suggesting that emotions are not easily reducible to simple definitions or classifications. This philosophical approach encourages a deeper exploration of emotions' subjective and existential dimensions.

In conclusion, studying emotions is multifaceted, integrating insights from biology, psychology, anthropology, and philosophy. Understanding emotions requires a comprehensive approach considering their evolutionary origins, structural characteristics, cultural variations, and philosophical implications. This integrated perspective enriches our knowledge of emotions and enhances our ability to apply this understanding in various domains, from mental health to artificial intelligence.

2.1.3 The significance and complexity of emotion detection in research

2.1.3.1 The role of emotions in human behavior

Studying emotions is crucial because they play a fundamental role in human behavior, decision-making, and social interactions. The significance of emotions extends across various disciplines, including psychology, neuroscience, artificial intelligence, and marketing. This literature review highlights key reasons why understanding emotions is essential. Emotions significantly influence decision-making processes. Damasio [21] proposed the somatic marker hypothesis, which suggests that emotional processes guide (or bias) behavior and decision-making, particularly in complex and uncertain situations. Emotions serve as a heuristic, helping individuals to quickly navigate decision-making processes by attaching emotional value to different choices. Furthermore, emotions are integral to social interactions and communication. Ekman and Friesen [32] demonstrated that facial expressions of emotion are universal, indicating that emotions play a crucial role in nonverbal communication. Understanding and interpreting emotions in others is essential for effective social interactions, empathy, and relationship building. The ability to read and respond to others' emotions fosters better interpersonal relationships and enhances social cohesion. In the context of mental health, the study of emotions is vital for research and practice. Emotional dysregulation is a key feature of many psychological disorders, including depression, anxiety, and bipolar disorder. Gross [41] highlighted the importance of emotion regulation strategies in mental health, demonstrating that the ability to effectively manage emotions is crucial for psychological well-being. Understanding how to regulate emotions can lead to better therapeutic interventions and improved mental health outcomes. In the field of artificial intelligence, understanding and replicating human emotions is critical for developing more sophisticated and human-like AI systems. Picard [82] introduced the concept of affective computing, which involves the creation of systems that can recognize, interpret, and simulate human emotions. This capability is essential for improving human-computer interactions and creating more empathetic and responsive technologies. By integrating emotional intelligence into AI, we can enhance user experiences and make technology more accessible and effective. In summary, studying emotions is essential due to their profound impact on decision-making, social interactions, mental health, artificial intelligence, marketing, and cognitive processes. Understanding emotions allows researchers and practitioners to develop better interventions, technologies, and strategies across various fields, ultimately enhancing human well-being and social functioning. By exploring the multifaceted role of emotions, researchers can gain deeper insights into human behavior and develop more effective ways to support and improve emotional health and social interactions.

2.1.3.2 Applications of emotion detection

Emotion detection in text and social media has gained significant attention in recent years, driven by advances in natural language processing (NLP) and machine learning. This field focuses on identifying and interpreting the emotional content of textual data, which has numerous applications across various domains. This literature review explores four key areas where emotion detection is being applied: mental health monitoring, customer feedback analysis, political sentiment analysis, and human-computer interaction. By examining recent studies, we highlight the advancements and challenges in these applications.

2.1.3.2.1 Mental Health Monitoring The application of emotion detection through social media platforms has become a crucial area of research, particularly in mental health monitoring. Recent studies have shown that machine learning (ML) techniques can effectively analyze vast quantities of user-generated content to identify vulnerabilities related to mental health conditions, including anxiety, depression, and suicidal ideation. For instance, Hu et al. explored the relationship between emotion goal dynamics and mental health, suggesting that fluctuations in emotional objectives can significantly impact psychological well-being. Their findings underscore the importance of understanding emotional states within the context of real-time interactions on social media, where such dynamics often unfold [47]. Similarly, Sijia et al. have emphasized the potential of artificial intelligence in psychological interventions, highlighting that ML can be employed to predict crises, which enables timely support and resource allocation for individuals showing signs of mental distress [100]. This capability is particularly relevant as it allows for proactive measures in mental health care, using social media data for early identification of at-risk populations.

Additionally, Kolliakou et al. conducted a time-series regression analysis revealing significant correlations between social media conversations concerning mental health and fluctuations in crisis episodes. They noted that problematic social media use can exacerbate symptoms of depression and anxiety, supported by the mechanisms of emotional contagion [61]. These insights suggest that monitoring social media discourse offers valuable indicators of community mental health trends.

Furthermore, the integration of ML in the analysis of social media data goes beyond passive observation. For example, Mastoras et al. demonstrated the feasibility of detecting depressive tendencies through analysis of typing patterns on phones and tablets, emphasizing how such technology can aid in the self-management of mental health [70]. This perspective aligns with the broader narrative surrounding remote monitoring in mental health care, where traditional methods are enhanced by innovative technologies that provide data-driven insights into emotional states.

In light of these advancements, the potential of machine learning in mental health applications on social media is evident. By leveraging data that reflects real-time emotional expression, health professionals can implement more precise interventions tailored to individual needs, thereby transforming the approach to mental health management in the digital age [100] [61] [70].

2.1.3.2.2 Customer Feedback Analysis Emotion detection is also extensively used in analyzing customer feedback to understand consumer sentiments and improve business strategies. Companies gather vast amounts of text data from product reviews, social media comments, and customer surveys. By employing emotion detection algorithms, businesses can gain insights into customer satisfaction and identify areas for improvement. Hussein et al. [49] conducted a comprehensive review of sentiment analysis techniques, emphasizing the role of emotion detection in understanding customer emotions at a granular level. Their work highlights the integration of deep learning models to enhance the accuracy of emotion classification in customer feedback. A specific application in the hospitality industry was explored by Xu et al. [119], who analyzed online reviews of hotels to detect emotions related to customer experiences. Their study demonstrated that emotion detection could identify patterns in customer satisfaction and dissatisfaction, providing actionable insights for improving service quality. Tumasjan et al. [110] analyzed customer sentiment on social media platforms such as Twitter. Their study showed how real-time emotion detection could help companies respond promptly to customer feedback, enhancing customer service and engagement. Jiang et al. [54] explored the use of emotion detection in e-commerce reviews. They developed a model to analyze both textual and contextual information, improving the accuracy of emotion analysis and providing deeper insights into customer emotions.

By understanding the emotional drivers behind customer feedback, businesses can tailor their strategies to enhance customer experiences and foster loyalty. Emotion detection allows companies to move beyond surface-level analysis of customer opinions, offering a deeper understanding of the emotional context that influences consumer behavior. This enables businesses to not only address customer concerns more effectively but also to proactively create positive emotional experiences that drive customer loyalty and long-term success.

2.1.3.2.3 Political Discourse Analysis Emotion detection plays a crucial role in political sentiment analysis, helping to gauge public opinion and predict electoral outcomes. Social media platforms, in particular, are rich sources of data reflecting public sentiment towards political events, candidates, and policies. Stieglitz et al. [103] explored the use of emotion detection in analyzing Twitter data during election campaigns. Their research demonstrated how emotions expressed in tweets could predict voter behavior and election results. They highlighted the importance of capturing emotional nuances to provide a more accurate representation of public sentiment. Further, Wang et al. [114] employed deep learning models to analyze emotional responses to political debates on social media. Their study showed that different emotions, such as anger, fear, and enthusiasm, could significantly influence political discourse and voter engagement. By analyzing tweets during and after debates, they were able to identify how specific emotional reactions correlated with shifts in public opinion and engagement levels. This research underscores the value of emotion detection in understanding the dynamics of political communication and its impact on public opinion, providing insights into how emotional rhetoric can shape voter perceptions and behaviors.

Kao and Jurafsky [57] focused on the role of emotions in political speeches. Using NLP techniques, they analyzed the emotional content of speeches by political leaders and found that emotional appeals significantly impacted public perception and support. Their analysis revealed that speeches with higher emotional content, particularly those evoking positive emotions like hope and pride, were more likely to resonate with audiences and garner support. This study highlights the strategic use of emotional rhetoric in political communication and its effectiveness in influencing public opinion.

Joyce et al. [56] extended this research by examining the emotional undertones in political tweets during major events such as presidential debates. Their study demonstrated how real-time emotion detection could track shifts in public sentiment and predict political trends. By analyzing the emotional content of tweets, they were able to observe how public sentiment evolved in response to specific statements or events during the debates. This approach provided valuable insights into the immediate emotional reactions of the public, which can be critical for understanding the impact of political events on voter attitudes.

These studies show the potential of emotion detection to provide deeper insights into political sentiment and enhance the accuracy of public opinion analysis.

2.1.3.2.4 Human-Computer Interaction Emotion detection is increasingly significant in human-computer interaction (HCI) and generative AI agents, driven by the need for systems that can understand and respond to human emotions. This capability enhances the user experience by making interactions more intuitive, empathetic, and effective. Emotion detection is crucial in HCI because it allows systems to adapt to the user's emotional state, leading to more personalized and satisfying interactions. Picard [82] introduced the concept of affective computing, which focuses on the development of systems that can recognize, interpret, and respond to human emotions (Picard, 1997). This foundational work paved the way for numerous applications in HCI.

One of the primary applications of emotion detection in HCI is in educational technology. Systems that can detect students' emotions can provide tailored feedback and support, enhancing learning outcomes. D'Mello and Graesser [29] demonstrated that affect-aware tutoring systems, which respond to learners' emotional states, can significantly improve engagement and learning efficiency. These systems can identify when a student is frustrated or confused and adjust the instructional strategies accordingly.

In healthcare, emotion detection is employed to support mental health interventions. Tao and Tan [109] developed a system that uses facial expression analysis to monitor patients' emotional states, aiding in the diagnosis and treatment of mental health conditions. Such systems can provide real-time emotional feedback to therapists, allowing for more responsive and effective treatment plans.

Moreover, emotion detection is utilized in customer service to improve user satisfaction. Virtual agents and chatbots equipped with emotion recognition capabilities can identify users' emotional cues and adjust their responses to be more empathetic and supportive. McTear et al. [71] explored the integration of emotion detection in conversational agents, highlighting its potential to enhance user experience by making interactions more natural and human-like.

Generative AI agents, particularly those based on large language models, have seen significant advancements with the integration of emotion detection. These agents can generate human-like text and engage in complex conversations, making emotion detection a valuable addition to enhance their interactions.

Large language models like GPT-3, developed by OpenAI, have shown remarkable capabilities in understanding and generating text. Integrating emotion detection allows these models to tailor their responses based on the emotional context of the conversation. Zhou et al. [128] proposed a framework for emotion-aware conversational AI, which enables generative models to adjust their tone and content according to the detected emotions. This capability is crucial for applications in customer service, where empathetic and contextually appropriate responses are essential.

Generative AI agents with emotion detection are also being explored in creative applications, such as storytelling and content generation. Ghosh et al. (2017) developed an emotion-aware storytelling system that generates narratives based on the emotional tone specified by the user [38]. This application not only enhances user engagement but also opens new avenues for personalized content creation in entertainment and education.

While the integration of emotion detection in HCI and generative AI agents presents numerous benefits, it also poses several challenges. One major challenge is the accuracy of emotion detection systems. Emotions are complex and can be expressed differently across cultures and individuals. Ensuring that systems can accurately recognize and interpret these variations is crucial for their effectiveness.

2.2 Related Work

This section will conduct a thorough literature review on emotion detection. We will examine the main methods for detecting emotions in text, highlighting the algorithms, models, and techniques developed and refined over the years. Additionally, we will explore the concept of explainable AI in the context of emotion detection, discussing its importance and the approaches used to ensure transparency and interpretability in AI systems.

The need for computational linguistics and AI models in detecting emotions is driven by the increasing reliance on digital communication and the necessity to enhance human-computer interaction. Several research studies argue for the importance and applicability of these technologies in various fields.

From a human-computer interaction lens, making computers understand human emotions is essential. Picard [82] puts forward the concept of affective computing, which emphasizes the importance of systems that can recognize and respond to human emotions to create more natural and effective interactions. This foundational work highlights how emotion detection can make machine interactions more intuitive and empathetic. Another reason emotion recognition is such an important area of research is the mass adoption of social media. Billions of people use their favorite social media platforms to communicate, express opinions, leave remarks, or socialize. The platforms are numbered but used by hundreds of millions of people. Some notable platforms include X (formerly Twitter), Facebook, Instagram, Reddit and Quora. Researchers have developed computational linguistics-based models to detect and classify emotions in social media data for many years. A study by Multidisciplinary Digital Publishing Institute (MDPI) [2] elaborates on how such models can analyze vast amounts of user-generated content to understand public sentiment and behavior. This capability is crucial for businesses and policymakers to monitor and respond to societal trends. These days, artificial intelligence is being integrated into many day-to-day tools. People interact with these AI models, such as OpenAI's ChatGPT, Google's Gemini, or Meta's Llama, to increase their productivity at work, school, or businesses [91]. Such generative AI models benefit significantly from incorporating emotion detection. Zhou et al. [128] discussed a framework for emotion-aware conversational AI, which allows these models to adjust their responses based on the detected emotions, making interactions more contextually appropriate and engaging.

2.2.1 Emotion classification

2.2.1.1 Linguistic methods

Linguistic rule-based approaches for emotion detection rely on predefined linguistic rules and patterns to identify emotional content in text. These methods involve creating rules based on syntactic, semantic, and lexical features to detect emotions effectively. One notable rule-based approach is the Affect Analysis Model (AAM), introduced by Neviarouskaya, Prendinger, and Ishizuka [76]. This model employs linguistic rules to analyze text and detect affective states. AAM is designed to recognize the effect of text messages and blogs, often using informal or garbled writing styles. The model processes sentences through stages like cue processing, abbreviation detection, and syntactic analysis, and it can handle various sentence complexities. It classifies affect into nine neutral emotion categories, using vectors of emotional words, their relationships, sentence tense, and first-person pronoun presence. Authors' evaluation across different text types shows promising accuracy (up to 77% for blogs, 70.2% for fairy tales), with superior performance compared to other systems in news headlines. Similarly, another significant contribution addressing a gap in emotion processing research by Lee et al. [64] presents a rule-based approach to detecting emotion causes in text. The authors construct a Chinese emotion cause annotated corpus and identify seven linguistic cues, formulating two rules for emotion cause detection. They develop a system based on these rules and propose a two-phase evaluation scheme for performance assessment. Experiments demonstrate promising results in detecting both emotional causes and events, laying a foundation for future research on implicit information and cause-event relationships. Rule-based linguistic methods are inherently not computational or statistical, allowing them to capture much meaning from the written text. There has been research on detecting even implicit emotions in text. Udochukwu et al. [111] introduce a rule-based pipeline approach for detecting implicit emotions in text, leveraging the OCC (Ortony, Clore, & Collins) Model. Unlike traditional methods focusing on explicit emotional expressions, this approach identifies emotions even in the absence of emotion-bearing words. Evaluated across three datasets with five emotion categories, the method consistently outperforms lexicon matching by 17–30% in F-measure and shows competitive performance compared to supervised classifiers. Notably, in formal texts, the approach achieves an average F-measure of 82.7% for "Happy," "Angry-Disgust," and "Sad," surpassing supervised baselines by nearly 17%. These results highlight the approach's effectiveness for implicit emotion detection.

Rule-based approaches offer several advantages, including transparency, interpretability, and the ability to incorporate domain-specific knowledge. These systems can be tailored to specific applications, making them highly adaptable. For instance, the AAM's comprehensive linguistic rules allow for detailed analysis of affective states. At the same time, the emotion cause detection system provides contextual insights for understanding the underlying reasons for emotional responses.

However, these approaches also face challenges, such as the difficulty in capturing the full complexity of human emotions and the reliance on extensive rule creation, which can be time-consuming and labor-intensive. Additionally, rule-based systems may struggle with the variability and richness of natural language, requiring continuous updates to remain effective. Using rule-based models, researchers started integrating these approaches with machine learning techniques to leverage the strengths of both methodologies. Hybrid models that combine rule-based and data-driven approaches can enhance the robustness and accuracy of emotion detection systems.

Supervised learning is a type of machine learning where the model is trained on a labeled dataset, meaning that each training example is paired with an output label. Supervised learning aims to learn a mapping from inputs to outputs based on the training data, enabling the model to predict the output labels for new, unseen data accurately. This section will discuss the existing supervised learning methods available for emotion classification.

2.2.1.2 Deep Learning and Transfer Learning methods

With advancements in machine learning and AI, many researchers have used a wide range of techniques and algorithms to solve the task of multi-label emotion classification. Fei, Zhang, Ren, and Ji's paper [35], "Latent Emotion Memory for Multi-Label Emotion Classification," addresses the task of identifying multiple emotions within a sentence. Treating this as a multi-label classification task encounters limitations in capturing the prior emotion distribution in a sentence and effectively incorporating context information related to the respective emotions. Fei et al. propose the Latent Emotion Memory network (LEM). This model aims to learn latent emotion distribution autonomously, overcoming the need for external knowledge. The latent emotion information is then effectively integrated into the classification network, addressing the identified limitations. The significance of this research lies in its potential to advance the state-of-the-art in multi-label emotion classification. The proposed LEM model introduces a novel approach to overcome existing methodological challenges, as demonstrated through superior performance compared to strong baselines in experimental evaluations on benchmark datasets.

Ameer et al., [5] address the growing interest in the supervised classification problem of Multi-label Emotion Classification in their research "Multi-label Emotion Classification using Content-Based Features in Twitter," particularly on its applications in diverse fields such as E-learning, marketing, education, and healthcare. The authors specifically focus on content-based methods, leveraging words and character n-grams to demonstrate their effectiveness in developing and evaluating Multi-label Emotion Classification models using Twitter data. While existing studies explore different methodologies, using content-based features, such as words and character n-grams, is a notable approach for this classification task. They contribute to the literature by proposing a content-based method, emphasizing the efficacy of content-based word unigrams in outperforming other content-based features. The achieved results, presented through metrics like Multi-label Accuracy, Micro-F1, Macro-F1, Exact Match, and Hamming Loss, provide valuable insights into the effectiveness of their proposed approach. This research extends the understanding of feature-based techniques in Multi-label Emotion Classification within the context of Twitter data.

Another research using a similar dataset on the problem is Jabreel and Moreno's paper [50], "A Deep Learning-Based Approach for Multi-Label Emotion Classification in Tweets", which contributes to the evolving landscape of sentiment and emotion analysis in social media. The authors emphasize the significance of detecting and analyzing these expressions across multiple labels. Prior sentiment and emotion analysis research often focused on single-label classification, neglecting the co-existence of multiple emotion labels in one instance. This gap is notable in the literature. The authors address this limitation by proposing a novel deep learning-based system designed specifically for the multi-label emotion classification problem in the context of Twitter data. Their approach involves transforming the problem into a binary classification. The system's performance surpasses state-of-the-art models, as evidenced by its accuracy score of 59% on the SemEval2018 Task 1: E-c multi-label emotion classification problem.

Some research is more focused on using external knowledge systems. Ying et al.'s paper [121], "Improving Multi-label Emotion Classification by Integrating both General and Domain Knowledge," addresses the challenges posed by social media text, which possesses distinct characteristics compared to general language. Previous research has underscored the effectiveness of deep learning models in tasks like sentiment analysis and question answering. However, the unique characteristics of social media language warrant specialized approaches. The authors propose a method that uses general knowledge acquired from deep language models with domain-specific knowledge to enhance emotion classification performance. Their work recognizes the significance of domain knowledge in tasks related to social media content. Experiments conducted on Twitter data demonstrate that even when a deep language model is fine-tuned with target domain data, further improvements can be achieved by integrating domain knowledge. This finding emphasizes the role of incorporating domain-specific insights for optimal performance in emotion classification, especially in applications tied to social media.

Ameer et al. [7] address the scarcity of research in multi-label emotion classification for code-mixed text, specifically English and Roman Urdu. While multi-label emotion classification has gained attention for its applications in various domains, existing benchmark corpora primarily focus on monolingual English text. Prior research has recognized the importance of benchmark corpora for evaluating multi-label emotion classification methods. However, the lack of exploration in the context of code-mixed text, prevalent in social media platforms like Facebook and Twitter among the South Asian community, creates a significant research gap. The authors introduce a substantial benchmark corpus comprising 11,914 code-mixed SMS messages manually annotated for 12 emotions to bridge this gap. Their study also compares classical machine learning, deep learning, and transfer learning-based methods on the proposed corpus. The results highlight the effectiveness of classical machine learning methods, particularly utilizing word uni-gram features. This research contributes a valuable benchmark corpus for code-mixed text and provides insights into the optimal methodologies for multi-label emotion classification in this linguistic context. The free availability of the proposed corpus further encourages research in under-resourced languages and code-mixed text.

The work "Multi-modal Multi-label Emotion Recognition with Heterogeneous Hierarchical Message Passing," by Zhang et al. [125] addresses a crucial aspect of affective-computing multi-modal emotion recognition. While this area has garnered significant attention, existing studies predominantly focus on binary classification and complete time series data. The emphasis on multi-label scenarios, considering label-to-label, feature-to-label, and modality-to-label dependencies, is relatively limited. The authors introduce a novel approach utilizing a heterogeneous hierarchical message-passing network to address this gap. This model is designed to capture the dependencies present in multi-modal, multi-label emotion recognition scenarios. Additionally, the authors contribute a new multi-modal, multi-label emotion dataset based on partial time-series content, showcasing the robust generalization capabilities of their proposed model. This research is significant in advancing the understanding and methodologies for multi-modal emotion recognition, particularly in scenarios involving multiple labels. The proposed model and dataset offer valuable contributions to the broader field of affective computing.

Similarly, another research done for emotions classification was by Ashraf et al. [10] to address the shortage of multi-label emotion datasets in a language as complex as Urdu (a widely spoken language in South Asia) in their paper "Multi-label Emotion Classification of Urdu Tweets." The authors introduced the first dataset for Urdu, comprising 6,043 tweets annotated with six basic emotions in the Urdu Nastalíq script, and use it further in their research. Existing research has recognized the challenge posed by Urdu's morphological and syntactic structure for emotion detection. The unique characteristics of Urdu script and language make multi-label emotion classification particularly complex. The paper contributes baseline classifiers, including machine learning algorithms (RF, J48, SMO, AdaBoostM1, and Bagging), deep-learning algorithms (1D-CNN, LSTM, LSTM with CNN features), and a transformer-based baseline (BERT). Various text representations, such as stylometric-based features, pre-trained word embeddings, and n-grams, are combined to tackle Urdu's multi-label emotion detection problem. This research is significant in providing annotation guidelines, dataset characteristics, and insights into diverse methodologies for emotion classification in Urdu tweets. The evaluation metrics, including micro-averaged F1, macro-averaged F1, accuracy, Hamming loss, and exact match, offer a comprehensive assessment of the tested methods. The paper contributes to the specific domain of Urdu emotion classification and the broader field of multi-label emotion analysis.

Our research "Multi-label Emotion Classification in Texts Using Transfer Learning," [6] focuses particularly on the detection of multiple emotions in short text segments—a multi-label classification problem. While previous work in emotion detection often focused on deep neural networks such as CNNs and RNNs like LSTMs, this study introduces novel contributions by utilizing multiple attention mechanisms and Transformer Networks (e.g., XLNet, DistilBERT, and RoBERTa) for multi-label emotion classification. Existing literature has acknowledged the difficulty of extracting emotions from short social media posts, emphasizing the need for advanced techniques. However, prior studies may not have fully explored the potential of specialized attention networks for each emotion and the application of pre-trained transformers through transfer learning. The proposed multiple attention mechanism in this study is innovative, shedding light on the specific contribution of each word to individual emotions—an aspect not thoroughly investigated before. The research compares the use of LSTMs and the fine-tuning of Transformer Networks for transfer learning, incorporating both single-attention and multiple-attention networks. The experimental results demonstrate the effectiveness of the proposed models, showcasing their ability to outperform the current state-of-the-art accuracy in the SemEval-2018 Task-1C dataset. Notably, the RoBERTa-MA model achieves 62.4% accuracy, surpassing the state-of-the-art by 3.6%. This establishes the proposed models, incorporating multiple attention mechanisms and Transformer Networks, as the current state of the art in multi-label emotion classification for English and Chinese datasets.

Existing literature has recognized the need to consider correlations among emotions in multi-label emotion classification tasks. However, limited attention has been given to leveraging task-specific information and exploring datasets for low-resource languages. To address these issues, Lin et al.'s paper [66] ("Multi-label Emotion Classification Based on Adversarial Multi-task Learning") proposes a novel approach involving multi-task multi-label emotion classification. The model comprises a general representation module, an emotion representation module, and an adversarial classifier. The incorporation of emotion descriptors facilitates capturing the correlation among different emotions, while adversarial training prevents an excessive injection of emotion-relevant information into the shared layer. The research not only contributes a new methodology but also introduces datasets for Indonesian and English, with 4207 and 26,019 samples, respectively, fostering future research in Indonesian multi-label emotion recognition resources. The proposed approach outperforms state-of-the-art baselines across various evaluation metrics, demonstrating macro-average F1 scores of 50.21%, 41.33%, and 40.24% on the Chinese, English, and Indonesian datasets, respectively. The open availability of codes and resources on GitHub further promotes transparency and reproducibility in multi-label emotion classification. The research addresses two critical challenges in multi-label emotion classification: the underexplored correlation among different emotions and the scarcity of public multi-label emotion datasets for low-resource languages.

The rise in suicides has prompted a need for prompt intervention and early diagnosis, making the detection of at-risk individuals crucial. Existing literature recognizes the connection between depression and suicide ideation. However, effective detection remains a significant challenge. Ghosh et al.'s paper [39], "A Multitask Framework to Detect Depression, Sentiment and Multi-label Emotion from Suicide Notes," addresses the critical issue of suicide prevention by focusing on three closely related tasks: depression detection, sentiment citation, and multi-label emotion analysis. The authors extend the CEASE corpus, a standard emotion-annotated dataset for suicide notes in English, by adding 2539 sentences from 120 new notes. The corpus is annotated with depression labels and multi-label emotion classes, while weak supervision is employed for sentiment labeling. The proposed multitask framework integrates a knowledge module incorporating external features from SenticNet's IsaCore and AffectiveSpace vector spaces into the learning process. The system simultaneously models emotion recognition as the primary task and depression detection and sentiment classification as secondary tasks. Experimental results demonstrate the effectiveness of the proposed multitask system, achieving the highest cross-validation MR of 56.47%. Importantly, the multitask models consistently outperform their single-task counterparts, emphasizing that jointly learning secondary tasks (depression detection and sentiment classification) enhances the performance of the primary task (emotion recognition). This research not only contributes to the understanding of mental health indicators in suicide notes but also offers a novel approach to improving detection accuracy through multitask learning.

2.2.1.3 Instruction Tuning prompts

Instruction prompt-based fine-tuning of Language Models (LLMs) represents a powerful and versatile Natural Language Processing (NLP) approach that holds significant promise across various tasks, including multi-label emotion classification. Researchers can guide LLMs to specialize in specific tasks or objectives by providing explicit instructions or prompts during the fine-tuning process. This technique allows for targeted adaptation, making the models more adept at understanding nuanced linguistic nuances associated with multi-label emotion classification. In emotion classification, where identifying multiple emotions in the text is a complex task, instruction prompt-based fine-tuning enables the LLMs to capture and learn intricate patterns, context, and dependencies related to different emotions. The adaptability of this approach makes it a valuable tool in solving a spectrum of NLP challenges, offering a tailored and efficient means to enhance the performance of LLMs across diverse linguistic tasks.

To understand the concept better, the paper by Yuan et al. [123], contributes to the ongoing exploration of instruction-based fine-tuning for large language models (LLMs), particularly in the realm of software engineering tasks. Prior research has extensively examined instructed LLMs across various natural language processing (NLP) tasks and specific domains. However, a notable gap exists in evaluating instructed LLMs within the software engineering domain, specifically beyond the NL-to-Code task. While recent efforts have touched on the application of instructed models, such as ChatGPT, in software engineering tasks, commercial models often lack transparency and reproducibility. In this context, the authors aim to address this knowledge gap by evaluating 10 open-source instructed LLMs on four key code comprehension and generation tasks, providing insights into their performance across diverse challenges in the software development lifecycle. The study investigates the effectiveness of instructed LLMs in zero-shot, few-shot, and fine-tuning settings, shedding light on their competitiveness, sensitivity to input length, and the impact of different shot selection strategies. The findings offer valuable insights into the performance of instructed LLMs in code-related tasks and present practical implications for model recommendations, performance trade-offs, and future directions in this evolving field.

The paper by Ye et al. [120], explores the enhancement of instruction-following capabilities in Large Language Models (LLMs) through the use of a Task-Agnostic Prefix Prompt (TAPP) during inference. While prior research has extensively delved into the fine-tuning of LLMs for specific tasks and instruction-based adaptation, this study takes a novel approach by introducing a fixed prompt, TAPP, which is appended to every input for zero-shot generalization, irrespective of the target task. The authors observe significant improvements in both base LLMs and instruction-tuned models, with average improvements of 34.58% and 12.26%, respectively, indicating the broad applicability of TAPP in enhancing instruction-following abilities. This innovative approach suggests that a simple fixed prompt with task-agnostic heuristics can substantially activate and improve the models' instruction-following capabilities during inference. The hypothesis posited by the authors suggests that TAPP assists language models in focusing more on the instructions of the target task during inference, contributing to improved estimation of the output distribution. This study fills a notable gap in understanding how fixed prompts can influence instruction-following in LLMs, shedding light on practical strategies for model improvement during inference.

There are certain challenges of hierarchical text classification (HTC) within multi-label classification. The paper by Wang et al [116]. discusses them. The authors recognize the complexity introduced by intricate label hierarchies and leverage pre-trained language models (PLMs) through fine-tuning to enhance HTC performance. Notably, they identify a significant gap between the sophisticated label hierarchies in classification tasks and the masked language model (MLM) pretraining tasks of PLMs. The authors introduce HPT, a Hierarchy-aware Prompt Tuning method that operates from a multi-label MLM perspective to address this. The approach involves constructing a dynamic virtual template and utilizing label words as soft prompts to incorporate label hierarchy knowledge. Additionally, a zero-bounded multi-label cross-entropy loss is introduced to align the objectives of HTC and MLM. Through extensive experiments, the paper demonstrates that HPT achieves state-of-the-art performance on three popular HTC datasets and excels in handling challenges such as imbalance and low-resource scenarios. This innovative method contributes significantly to advancing HTC techniques, showcasing the effectiveness of hierarchy-aware prompt tuning in leveraging the potential of PLMs for hierarchical text classification tasks.

The challenges continue in the form of the difficulties associated with prompt-based fine-tuning of Pre-trained Language Models (PLMs) in the context of few-shot text classification. The paper by Wang et al. [113]. The authors acknowledge that PLMs lack familiarity with prompt-style expressions during pre-training, hindering their few-shot learning performance on downstream tasks. They propose the Unified Prompt Tuning (UPT) framework to enhance few-shot text classification for BERT-style models by explicitly capturing prompting semantics from non-target Natural Language Processing (NLP) datasets. UPT introduces the paradigm Prompt-Options-Verbalizer, facilitating joint prompt learning across different NLP tasks and compelling PLMs to acquire task-invariant prompting knowledge. The authors also incorporate a self-supervised task called Knowledge-enhanced

Selective Masked Language Modeling to improve PLM generalization abilities for accurate adaptation to unseen tasks. Through multi-task learning across various tasks, UPT enables practical prompt tuning for dissimilar target tasks, particularly in low-resource settings. Experimental results across a range of NLP tasks demonstrate that UPT consistently outperforms existing state-of-the-art methods for prompt-based fine-tuning.

Another challenge is building a text classifier from a neural language model without access to the LM's parameters, gradients, or hidden representations. Hou et al. [46] discusses them in the paper. In the context of black-box classifier training, where the cost of training and inference in large-scale LMs is a concern, existing methods are computationally inefficient. These approaches often specialize LMs to the target task by exploring a vast space of prompts using zeroth-order optimization methods. In contrast, PromptBoosting introduces a query-efficient procedure that achieves state-of-the-art performance in black-box few-shot classification tasks. The method obtains a small pool of prompts through a gradient-free approach and constructs a large pool of weak learners by pairing these prompts with different elements of the LM's output distribution. The weak learners are then ensembled using the AdaBoost algorithm, requiring only a small number of forward and no backward passes. PromptBoosting demonstrates competitive performance, matching or outperforming full fine-tuning in both few-shot and standard learning paradigms while achieving a training speed that is ten times faster than existing black-box methods.

The paper by Zhang et al. [126] addresses the challenge of data availability. While current meta-learning methods have demonstrated success in various few-shot situations, they often require substantial data for constructing multiple few-shot tasks during meta-training, which may not be practical in real-world few-shot scenarios. On the other hand, prompt-tuning has emerged as an effective few-shot learner by bridging the gap between pre-training and downstream tasks. The authors propose a novel Prompt-Based Meta-

Learning (PBML) model, combining the strengths of both meta-learning and prompt-tuning methodologies. PBML assigns label word learning to base learners and template learning to meta-learners. Experimental results showcase state-of-the-art performance on four text classification datasets under few-shot settings, highlighting improved accuracy and robustness. The method also demonstrates effectiveness in low-resource scenarios, addressing the data-intensive nature of traditional meta-learning. Visualization is employed to interpret and verify the convergence improvements facilitated by the meta-learning framework.

Prompt-tuning involves inserting prompt text into the input, transforming the classification task into a masked language modeling task. The paper by Ni and Kao [78] explores the paradigm of "pre-train, prompt, and predict," which has shown notable achievements in few-shot learning compared to traditional "pre-train, fine-tune" paradigms. Knowledgeable prompt-tuning (KPT) enhances this approach by integrating external knowledge into the verbalizer, expanding the label word space using word embeddings and knowledge graphs. However, KPT may suffer from unreasonable label words affecting accuracy. The paper introduces KPT++ as a refined version of KPT, incorporating prompt grammar refinement (PGR) and probability distribution refinement (PDR) to enhance the knowledgeable verbalizer. Experimental results on few-shot text classification tasks demonstrate that KPT++ outperforms the state-of-the-art method KPT and other baseline methods, with ablation experiments and case studies supporting the effectiveness of PGR and PDR refining methods.

While vision-language pre-training models (VL-PTMs) have been extensively studied for visual tasks like image classification, their application to language tasks like text classification is less explored. The paper by Wen et al. [118] explores the deployment of large-scale pre-trained models in the prompt-tuning paradigm, demonstrating promising performance in few-shot learning. The authors introduce Visual Prompt Tuning (VPT), a novel method for deploying VL-PTMs in few-shot text classification. VPT aligns input samples and category names through text encoders, incorporating visual information learned by image encoders. Experimental results show that VPT significantly improves under both zero-shot and few-shot settings, even outperforming recent prompt-tuning methods on five public text classification datasets.

The paper by Dan et al. [22] addresses the challenge of zero-shot text classification (ZSTC), where there is a lack of labeled data for unseen classes during training. While existing studies have focused on knowledge transfer from seen to unseen classes, the authors propose a prompt-based method to enhance semantic understanding for each class, specifically addressing semantic gaps or low similarities. The technique involves generating discriminative words for class descriptions using prompt inserting (PIN) and learning a prompt matching (POM) model to assess the text's compatibility with class descriptions. Experimental results on three benchmark datasets demonstrate the effectiveness of the proposed approach, achieving state-of-the-art performance on unseen classes while maintaining competitive strength with existing ZSTC approaches for seen classes.

The challenges of multi-label text classification, mainly focusing on extracting and leveraging correlations among labels, are discussed in the paper by Song et al. [102]. The authors introduce the Label Prompt Multi-label Text Classification model (LP-MTC), inspired by prompt learning in pre-trained language models. LP-MTC utilizes a set of templates for multi-label text classification, integrating labels into the input of the pre-trained language model, and optimizing jointly through Masked Language Models (MLM). This approach aims to capture correlations among labels and semantic information between labels and text using self-attention, improving model performance. Empirical experiments on multiple datasets demonstrate the effectiveness of LP-MTC, showing a 3.4% improvement in micro-F1 on average over four public datasets compared to BERT.

2.2.1.4 Zero-shot classification

Zero-shot emotion classification with Large Language Models (LLMs) involves categorizing text into predefined emotion categories without explicit training on labeled data for each category. Instead, it leverages the pre-trained semantic understanding of LLMs, like GPT models, to infer emotional content. The model generates a probability distribution over predefined emotions by providing a prompt or description of the target emotion category alongside the input text, indicating the likelihood of each being expressed. This approach minimizes the need for extensive labeled data, making it beneficial in scenarios with limited emotion-specific training data.

The effectiveness of zero-shot emotion classification depends on the quality of the pre-trained model and the appropriateness of prompts or auxiliary training techniques. While offering a promising solution for situations where labeled emotion data is scarce, successful implementation relies on the LLM's ability to generalize across diverse emotional expressions and contexts. This section highlights some literature reviews of the research already published in emotion classification by leveraging large language models.

Zero-shot learning, facilitated by Large Language Models (LLMs), has emerged as a promising approach in Natural Language Processing (NLP), offering an efficient alternative to traditional supervised training methods. In their work, De Langhe et al. [25] explore the landscape of zero-shot text classification within the Dutch language domain. Through a comprehensive examination of methodologies, resources, and challenges, the authors present centralized benchmark results across various Dutch NLP tasks, from social media sentiment and emotion detection to news topic classification and event coreference resolution. Their findings reveal the efficacy of task-specific fine-tuning over zero-shot approaches, except notably in emotion detection. Moreover, they observe the superiority of large generative models in zero-shot settings, with prompting outperforming NLI models and MLM approaches. The study highlights important considerations such as evaluation streamlining, parameter efficiency, and prompt optimization, underscoring both the potential and limitations of zero-shot learning in practical applications.

Due to growing concerns about mental health, the role of AI in aiding detection and diagnosis has gained prominence. In their study, Jain et al. [52] explore the potential of AI language models, particularly GPT-2 and GPT-Neo-125M, in recognizing emotions such as stress, depression, and suicidality. This research marks a significant step in understanding how advanced AI tools can contribute to mental health analysis, addressing the pressing need for efficient and accurate detection methods. By evaluating the performance of these models in mental health classification tasks, the study sheds light on the evolving landscape of mental health care, highlighting the promising role of AI in leveraging textual data for improved diagnosis and support.

In the domain of textual emotion classification, where labels vary across domains and applications, traditional supervised learning methods face challenges due to the need for predefined labels. To address this issue, Plaza-del-Arco et al. [84] explore the paradigm of zero-shot learning within natural language inference, investigating how different prompt formulations influence model sensitivity across diverse corpora. In their paper, the authors investigate the sensitivity of natural language inference-based zero-shot learning classifiers to prompt formulations for emotion classification in textual data. Through experiments on various emotion datasets from different language registers, such as tweets, events, and blogs, they elucidate the significance of selecting appropriate prompts tailored to the characteristics of the corpus. Their findings highlight the importance of prompt selection and propose using prompt ensembles to enhance robustness across corpora, achieving performance levels comparable to the best individual prompt for each dataset.

In the contemporary digital landscape, online resources play an integral role in our daily lives, with businesses and individuals alike relying on user-generated reviews and comments to inform decision-making processes. Within these texts, sentiment analysis is vital for discerning positive, negative, and neutral sentiments and crucial for understanding user perspectives. However, texts often convey various emotions alongside sentiments, adding complexity to sentiment analysis tasks. In their paper, Tesfagergish et al. [36] propose a two-stage emotion detection methodology within the sentiment analysis framework. Their approach leverages unsupervised zero-shot learning with a sentence transformer to detect 34 emotions, followed by supervised ensemble learning for sentiment classification. Achieving a commendable accuracy of 87.3% on the English SemEval 2017 dataset, their hybrid semi-supervised method demonstrates promise in accurately discerning sentiments while considering emotional nuances within textual data.

In recent years, GPT-4 with Vision (GPT-4V) has garnered attention for its impressive visual capabilities across various tasks, yet its performance in emotion recognition remains underexplored. To address this gap, Zheng Lian et al. [65] present a quantitative evaluation of GPT-4V on 21 benchmark datasets spanning six tasks under the umbrella of Generalized Emotion Recognition (GER): visual sentiment analysis, tweet sentiment analysis, micro-expression recognition, facial emotion recognition, dynamic facial emotion recognition, and multimodal emotion recognition. Through systematic experimentation, the authors observe GPT-4V's robust visual understanding abilities in GER tasks, alongside its capacity to integrate multimodal cues and leverage temporal information critical for emotion recognition. However, the study notes GPT-4V's limitation in recognizing micro-expressions that demand specialized knowledge. This paper serves as the inaugural quantitative assessment of GPT-4V for GER tasks, and the authors have made the code open-source, encouraging further research to expand the evaluation scope by incorporating additional tasks and datasets.

In Multi-modal Emotion Recognition (MER), identifying human emotions from diverse modalities presents a complex challenge, especially when encountering unseen emotion labels. To address this challenge, Qi et al. [87] propose a versatile zero-shot MER framework in their paper. This framework is designed to refine emotion label embeddings, enhancing discrimination between labels and capturing inter-label relationships. It integrates prior knowledge into an effective graph space, generating tailored label embeddings. Additionally, it disentangles features of each modality into egocentric and altruistic components using adversarial learning and hierarchically fuses them through a hybrid co-attention mechanism—moreover, an emotion-guided decoder leverages label-modal dependencies to generate adaptive multimodal representations. Extensive experiments across various multimodal combinations on four datasets, including visual-acoustic and visual-textual inputs, underscore the superiority of the proposed framework over existing methods.

2.2.2 Explainable AI

Explainable AI (XAI) is increasingly recognized as crucial in developing machine learning models. The primary goal of XAI is to make the decision-making processes of AI models transparent and understandable to human users. This is particularly important in emotion and sentiment classification, where the implications of model decisions can significantly impact user experience and trust. According to Doshi-Velez and Kim [30], the need for explainability in AI arises from ensuring accountability, fairness, and transparency in automated systems. Additionally, Ribeiro et al., [94] emphasize that understanding model predictions helps diagnose errors and improve model performance. In emotion and sentiment classification, explainable AI enables stakeholders to comprehend how and why certain emotional states or sentiments are detected, fostering trust and facilitating better human-computer interaction.

Cambria et al. [28] provides a comprehensive overview of sentiment analysis techniques and the corresponding XAI methodologies. The authors argue that the need for explainability grows correspondingly as AI models become increasingly complex. They emphasize that understanding how and why AI models make specific decisions is crucial for building trust and ensuring accountability. This paper discusses various XAI techniques, such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), which help elucidate the inner workings of sentiment analysis models. By making the models' decision processes transparent, these techniques enhance user trust and facilitate the debugging and improvement of AI systems. The authors highlight that users and developers are left in the dark about the reliability and fairness of the AI's decisions without explainability. This lack of transparency can lead to mistrust and reluctance to adopt AI technologies, particularly in sensitive applications where understanding the reasoning behind decisions is critical. Therefore, the study underscores the need for ongoing research into XAI methodologies to ensure that sentiment analysis systems are accurate but also transparent and trustworthy. Building on this foundation, Introduction to Explainable AI for Sentiment Analysis explores the challenges of explainability in AI, particularly sentiment analysis. The article highlights the growing demand for models that not only perform well but also provide explanations for their predictions. It discusses the use of attention mechanisms in neural networks, where the model highlights the parts of the input text that most influenced its predictions, providing a form of intrinsic explainability. This approach aligns with the previous study's findings, emphasizing that explainable AI techniques in sentiment analysis improve user trust and facilitate better decision-making.

Due to emotions' nuanced nature, emotion classification involves more complexities than sentiment analysis. Explainable AI methods are, therefore, essential in this domain to provide clear insights into the model's interpretability and decision-making processes.

A Review on Emotion Detection by Using Deep Learning Techniques [19] reviews various deep learning approaches for emotion detection, emphasizing the role of XAI in making these models interpretable. The paper highlights techniques such as gradient-based methods and attention mechanisms, which help visualize the contribution of different input features to the model's predictions. This study reinforces that incorporating XAI techniques in emotion detection models enhances their transparency and helps debug and improve model performance.

Similarly, Understanding the Prediction Mechanism of Sentiments by XAI Methods [101] focuses on using XAI methods to explain sentiment predictions in the context of online reviews. This research underscores the importance of understanding prediction mechanisms to build trust in AI systems. The study employs SHAP and LIME to provide local explanations for individual predictions, helping users understand how specific features influence sentiment analysis outcomes. This aligns with previous findings by showing that applying XAI methods reveals insights into how different textual features contribute to sentiment scores, aiding in refining AI models and increasing user confidence. What are the Current State-of-the-Art Explainable AI Methods Used in Emotion Classifiers? (2024) discusses the latest advancements in XAI methods specifically applied to emotion classifiers, emphasizing the need for explainability in complex emotion detection models. The article highlights the development of Explainable Multimodal Emotion Recognition models, which comprehensively combine visual, textual, and auditory data to explain the emotion detection process. This study bridges the gap between sentiment analysis and emotion classification, underscoring the broader applicability and necessity of XAI techniques across different aspects of emotion and sentiment detection.

The literature emphasizes that explanations can improve user trust and facilitate better human-AI collaboration [127] [16]. Recent evaluations of various explanation methods have underscored their utility in providing insights into decision-making processes. They have shown how traditional statistical and feature attribution methods are insufficient for adequate explanations [33]. Additionally, developing self-explaining architectures aims to incorporate interpretability into neural text classifiers inherently, [89]. Recent advancements in AI, particularly with the emergence of LLMs, have transformed various applications, but making these models behave according to task-specific needs remains challenging. Preference alignment algorithms have gained attention as a means to align AI models with human preferences, ensuring that AI outputs are more relevant and acceptable to users [72]. Various preference alignment techniques, such as reinforcement learning from human feedback (RLHF), have been instrumental in fine-tuning models for applications ranging from content moderation to personalized recommendations [72]. These algorithms enable models to learn from user interactions, enhancing their performance in real-world scenarios [72]. Despite advancements in LLMs, uncontrolled generation remains a significant challenge, often leading to hallucinations—instances where the model generates false or nonsensical information. This phenomenon can undermine the reliability of AI systems, particularly in applications requiring factual accuracy, such as news generation or medical advice [14] [63]. The literature indicates that hallucinations can arise from various factors, including insufficient training data and the inherent complexity of language generation tasks [14] [63]. Addressing this issue is critical to ensuring the safe deployment of LLMs in sensitive applications. Preference alignment enhances the relevance of AI outputs and can be leveraged to create self-explaining models with controlled quality explanations. By aligning model behavior with user preferences, developers can ensure that the explanations generated by AI systems are not only accurate but also tailored to the user's context and needs [72]. This approach can significantly improve user trust and satisfaction, as users are likelier to engage with systems that provide clear and relevant explanations for their outputs [72]. Integrating preference alignment in self-explaining models represents a promising direction for future research, aiming to bridge the gap between complex AI systems and user understanding.

2.3 Summary

In this chapter, we explored the theoretical foundations of emotion, including psychological models and taxonomies such as Ekman's Basic Emotions and Plutchik's Wheel of Emotions, to provide a grounding for understanding how emotions manifest in textual data. We examined the cognitive structure of emotions and their relevance to computational tasks. Additionally, we reviewed the significance and complexity of emotion detection in natural language, outlining key applications in mental health monitoring, customer feedback, and political discourse analysis.

The second half of the chapter presented a comprehensive review of existing methods for emotion classification and explainability, including linguistic techniques, deep learning approaches, and the emerging role of large language models.

In the next chapter, we will introduce the datasets used in this research, describe their composition and statistics, and explain the rationale for selecting them for the multi-label and multi-class emotion classification tasks.

Chapter 3

Data

This chapter presents the available dataset and the methodology we will employ to train and evaluate emotion classification models capable of handling multi-label and multi-class classification tasks.

3.1 SemEval 2018 Affect in Tweets Dataset

The "Affect in Tweets" dataset is a widely used resource in multi-label emotion classification. Specifically, this dataset is from the SemEval 2018 Task 1 E-c, a shared task dedicated to exploring affective content in tweets. Released in 2018, the dataset is comprehensive, providing a range of tweets annotated with 11 distinct emotion categories for each instance. The emotions covered in the dataset include but are not limited to joy, sadness, anger, surprise, and fear, reflecting the human emotions expressed in social media. Including multiple emotion labels for each tweet in the dataset makes it suitable for multi-label emotion classification.

3.1.1 Dataset composition

The dataset comprises tweets, each annotated with one or more emotion labels from a predefined set of 11 distinct emotions. The emotions present in the dataset are:

1. anger (also includes annoyance, rage)
2. anticipation (also includes interest, vigilance)
3. disgust (also includes disinterest, dislike, loathing)
4. fear (also includes apprehension, anxiety, terror)
5. joy (also includes serenity, ecstasy)
6. love (also includes affection)
7. optimism (also includes hopefulness, confidence)
8. pessimism (also includes cynicism, no confidence)
9. sadness (also includes pensiveness, grief)
10. surprise (also includes distraction, amazement)
11. trust (also includes acceptance, liking, admiration)
12. neutral or no emotion

The multi-label nature of the dataset is particularly relevant, as it mirrors real-world scenarios where tweets often express multiple emotions simultaneously. Table 3.1 shows some multi-label emotion examples from the dataset.

3.1.2 Dataset statistics

In table 3.2, we describe the dataset partition sizes and statistics about emotions and their respective populations. Figure 3.1 shows a heatmap that visualizes how various emotions correlate in the SemEval 2018 dataset. We observe strong positive Pearson correlations

ID	Tweet	Emotions
T1	My roommate: it's okay that we can't spell because we have autocorrect. #terrible #firstworld probs	anger, disgust
T2	@FaithHill I remember it well #happy #afraid #Positive	happy, joy, love, optimism
T3	@wabermes The @RavalliRepublic had a good one but then the reporter quit.	neutral

Table 3.1: Example of multi-label text for emotion classification

between several emotion pairs, such as (anger, disgust), (joy, love), (joy, optimism), (pessimism, sadness), and (trust, optimism). These positive correlations indicate that these emotions often co-occur in tweets, reflecting patterns of emotional expression that models can leverage for improved training and interpretation.

Conversely, negative correlations in the heatmap represent inverse relationships between emotions. A negative correlation means that when one emotion is present, the other is less likely to appear simultaneously. For example, the negative correlation between (anger, joy) suggests that tweets expressing anger rarely express joy at the same time. It is important to note that a negative correlation does not imply the absence of any relationship but rather an inverse association between the emotions.

3.2 Twitter Emotions Corpus (TEC)

Saif Mohammad introduced the TEC (Twitter Emotion Corpus) dataset in his 2012 paper titled "#Emotional Tweets" [73]. The TEC dataset was specifically designed for emotion classification in tweets, an important task given the increasing use of social media platforms

Dataset	**Train**	**Dev**	**Test**	**Total**
SemEval-2018	6,838	886	3,259	10,983
Emotions	**%**			
Anger	36.1			
Anticipation	13.9			
Disgust	36.6			
Fear	16.8			
Joy	39.3			
Love	12.3			
Optimism	31.3			
Pessimism	11.6			
Sadness	29.4			
Surprise	5.2			
Trust	5.0			
Neutral	2.7			

Table 3.2: Dataset distribution statistics in Train, Dev, Test sets. Also includes the percentage of presence of each respective emotion in the dataset.

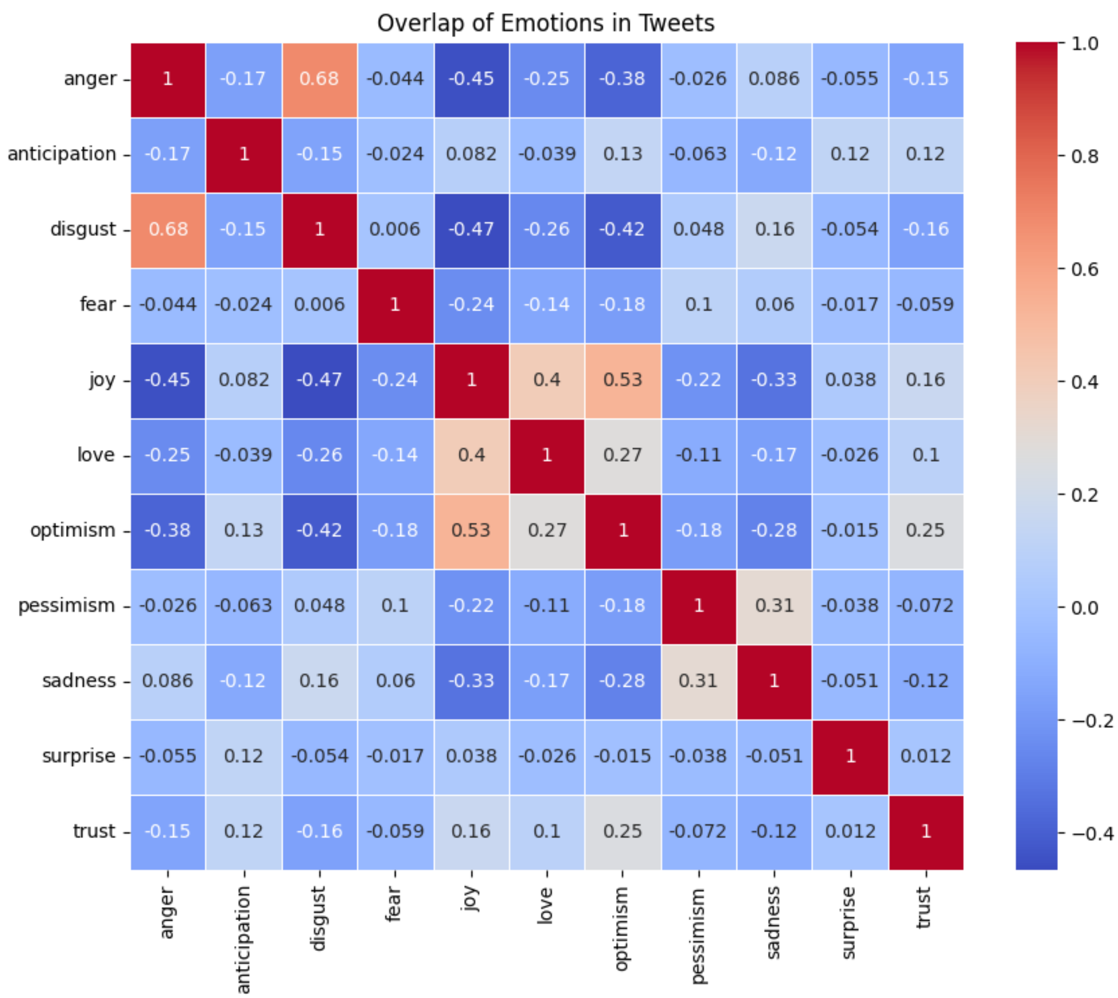

Figure 3.1: Heatmap illustrating the correlation of emotions in the SemEval 2018 dataset.

like Twitter for sharing personal thoughts and emotions. Table 3.3 shows some examples from the dataset.

ID	Tweet	Emotions
T1	Losing people everyday and I'm sick of cryin nd bein sad is the next emotion !	anger
T2	this year, for the first time in my life, I will not have a christmas tree in my home	sadness
T3	Enjoy this Day Beautiful World and Planet Earth's wonderful people!!	joy

Table 3.3: Example of multi-class tweets for emotion classification from the TEC dataset

3.2.1 Dataset composition

The TEC dataset comprises a large collection of tweets that are annotated with one of six basic Ekman emotions: anger, disgust, fear, joy, sadness and surprise. These emotions were selected based on their significance in psychological studies and their relevance to social media content. The tweets in the TEC dataset were annotated using hashtags associated with these emotions (e.g., #happy, #sad). Mohammad's approach leveraged these emotion-related hashtags to create a labeled dataset.

3.2.2 Dataset statistics

The dataset statistics of the TEC dataset are shown in Table 3.4. This distribution reveals a significant class imbalance, with the "joy" category having the most tweets, nearly twice as many as "sadness" and "surprise," and more than ten times the number of tweets labeled as "disgust." Class imbalance is a common issue in sentiment and emotion analysis, particularly in datasets like TEC where certain emotions are more frequently expressed in social media. Also, the original dataset, as provided in the paper, does not come with predefined training, validation, and test sets. This lack of predefined splits requires us to

Emotion	# of instances	% of instances
Anger	1,555	7.4
Disgust	761	3.6
Fear	2,816	13.4
Joy	8,240	39.1
Sadness	3,830	18.2
Surprise	3,849	18.3
Total Tweets	21,051	100.0
# of tweeters	19059	100.0

Table 3.4: Details of the Twitter Emotion Corpus

create our own datasets for conducting experiments. We will employ stratified sampling techniques to ensure that the training, validation, and test sets are representative of the original distribution. This approach will help maintain the proportion of each emotion in all subsets, thereby allowing the models to learn effectively across all emotion categories.

3.3 Dair AI

The Dair AI dataset, introduced in the paper "CARER: Contextualized Affect Representations for Emotion Recognition" by Saravia et al. published in 2018 [96], is a comprehensive resource for emotion classification. It is derived from Twitter data and contains two splits: Split 1 contains 416,809 rows with 6 distinct emotions, while Split 2 includes 20,000 rows annotated with the same emotion labels divided into the train (16000), validation (2000), and test (2000) sets. Originally, they worked on 8 emotions in their paper but the released dataset only contains the 6 basic emotions. Table 3.5 shows some examples from

ID	Tweet	Emotions
T1	ive probably mentioned this before but i really do feel proud of myself for actually keeping up with my new years resolution of monthly and weekly goals	joy
T2	i believe that i am much more sensitive to other peoples feelings and tend to be more compassionate	love
T3	i pretty much waddled out of the hospital feeling weird lightheaded but ok	surprise

Table 3.5: Example of multi-class tweets for emotion classification from the Dair AI dataset

the dataset.

3.3.1 Dataset composition

The dataset is constructed using a collection of English tweets gathered via the Twitter API. The dataset employs 339 hashtags associated with eight basic emotions—anger, anticipation, disgust, fear, joy, sadness, surprise, and trust—utilizing these as noisy labels for distant supervision, as outlined by Go et al. [40]. To ensure data quality, preprocessing steps from Abdul-Mageed and Ungar [1] are applied, with the hashtag appearing in the final position of each tweet serving as the ground truth. The publicly released version of the dataset focuses on six basic Ekman emotions—' sadness,' 'joy,' 'love,' 'anger,' 'fear,' and 'and surprise'—making it particularly suitable for educational and research applications.

Dataset	Train	Dev	Test	Total
Dair AI emotion	16,000	2,000	2,000	20,000
Emotions	**Count (%)**	**Count (%)**	**Count (%)**	**Count (%)**
Sadness	4666 (29.1)	550 (27.5)	581 (29.05)	5797 (29.0)
Joy	5362 (33.5)	704 (35.2)	695 (34.8)	6761 (33.8)
Love	1304 (8.15)	178 (8.9)	159 (7.95)	1641 (8.2)
Anger	2159 (13.5)	275 (13.8)	275 (13.75)	2709 (13.5)
Fear	1937 (12.1)	224 (11.2)	212 (10.6)	2373 (11.9)
Surprise	572 (3.6)	81 (4.05)	66 (3.3)	719 (3.6)

Table 3.6: Details of the Dair AI Corpus

3.3.2 Dataset statistics

The dataset statistics of the Dair AI dataset are shown in 3.6. The distribution reveals a significant class imbalance, with Joy and Sadness being the most frequent emotions dominating the dataset. In contrast, emotions like Anger and Fear are moderately represented, while Love and Surprise are notably less frequent. The provided dataset is partitioned into train, validation, and test sets, which will be used in our experiments.

3.4 ISEAR

Klaus R. Scherer and Harald Wallbott compiled the ISEAR (International Survey on Emotion Antecedents and Reactions) dataset over several years during the 1990s.

ID	Tweet	Emotions
E1	i feel awful about it too because it s my job to get him in a position to succeed and it just didn t happen here	sadness
E2	ive probably mentioned this before but i really do feel proud of myself for actually keeping up with my new years resolution of monthly and weekly goals	joy
E3	i beleive that i am much more sensitive to other peoples feelings and tend to be more compassionate	love

Table 3.7: Example of multi-class experiences for emotion classification from the ISEAR dataset

3.4.1 Dataset composition

This large-scale project involved psychologists worldwide and collected data from nearly 3,000 student respondents, both psychologists and non-psychologists, across 37 countries spanning all five continents. Participants were asked to describe situations in which they had experienced each of the seven major emotions: joy, fear, anger, sadness, disgust, shame, and guilt. The survey also gathered information on how participants appraised these situations and their emotional reactions. The resulting dataset comprehensively collects emotional experiences, capturing nuanced human emotions in diverse cultural contexts. Table 3.7 shows some examples from the dataset.

Emotion	**# of instances**	**% of instances**
Anger	1096	14.29
Sadness	1096	14.29
Disgust	1096	14.29
Shame	1096	14.29
Fear	1095	14.28
Joy	1094	14.27
Guilt	10963	14.25
Total Experiences	7,666	100.0

Table 3.8: Details of the ISEAR Corpus

3.4.2 Dataset statistics

The dataset statistics of the ISEAR dataset are shown in 3.8. The ISEAR dataset statistics indicate that the dataset is well-balanced across all emotion categories. Each emotion—anger, sadness, disgust, shame, fear, joy, and guilt—has nearly the same number of instances, with counts ranging from approximately 1,093 to 1,096 entries per emotion. This uniform distribution ensures that the dataset does not suffer from class imbalance, a common issue in many emotion recognition datasets.

3.5 Affective Text

The Affective Text dataset, created by Strapparava and Mihalcea (2007) [105], is a collection of news headlines comprising 1,250 instances. This dataset is primarily designed to classify emotions and valence in news headlines. The annotation framework is based on

ID	Tweet	[anger, disgust, fear, joy, sadness, surprise]
H1	Mortar assault leaves at least 18 dead	[22, 2, 60, 0, 64, 0]
H2	Nigeria hostage feared dead is freed	[18, 0, 52, 66, 20, 65]
H3	Bombers kill shoppers	[66, 39, 94, 0, 86, 0]

Table 3.9: Example of multi-class tweets for emotion classification from the Affective Text dataset

Ekman's basic emotions, supplemented with valence scoring.

3.5.1 Dataset composition

The Affective Text dataset comprises news headlines provided in an SGML formatted file, where each line contains a single headline and its unique identifier. Separate annotation files for emotions and valence indicators complement the dataset. Each line in the emotion annotations file corresponds to a headline's identifier and lists six emotion scores—anger, disgust, fear, joy, sadness, and surprise—ranging from 0 to 100.

3.5.2 Dataset statistics

The SemEval-2007 Task 14: Affective Text dataset contains 1,250 news headlines, split into 1,000 for testing and 250 for training. Each headline is annotated with six basic emotions: anger, disgust, fear, joy, sadness, and surprise, with intensity scores from 0 to 100. Additionally, the dataset includes valence classification (positive, negative, or neutral). While the exact emotion distribution is not provided, joy, sadness, and surprise are generally more common, with anger, fear, and disgust being less frequent. This dataset is important for studying emotion detection in short-form text.

3.6 Chosen Datasets

This chapter reviewed several datasets commonly used for emotion classification tasks. After careful consideration, we selected two datasets for our research: the SemEval 2018 E-c dataset for the multi-label emotion classification task and the DAIR AI dataset for the multi-class emotion classification task.

The SemEval 2018 E-c dataset was chosen because it is the most widely researched dataset in the multi-label emotion classification domain. Released as part of a shared task, it provides a strong foundation for benchmarking models and comparing results with previous research.

For the multi-class emotion classification task, we selected the DAIR AI dataset. Unlike older datasets, DAIR AI contains more recent social media data, making it a valuable resource for studying current trends and emotional expressions online. Additionally, it has verified state-of-the-art results on Huggingface[1], and is one of the most downloaded and widely used datasets in the multi-class emotion classification category.

Importantly, the DAIR AI emotion categories are a subset of the emotion labels found in the SemEval 2018 E-c dataset. This overlap allows us to build a comprehensive self-explaining emotion classification model that can output the most prominent emotion (based on DAIR AI's six basic emotions), the set of multiple emotions present (based on SemEval's more fine-grained labels), and a natural language explanation supporting those predictions.

[1] https://huggingface.co/datasets/dair-ai/emotion

In this setup, the multiple emotions can be interpreted as more nuanced derivatives of the six basic DAIR AI emotions. This alignment between datasets enables consistent evaluation and structured output generation across both multi-class and multi-label emotion recognition tasks.

These datasets, when used together, provide a strong and compatible foundation for building and evaluating an interpretable, multi-task framework for emotion classification in social media.

3.7 Summary

In this chapter, we presented and analyzed the datasets considered for emotion classification in this thesis. We provided detailed composition and statistics for five publicly available emotion-labeled datasets: SemEval 2018 Affect in Tweets, Twitter Emotion Corpus (TEC), DAIR AI, ISEAR, and Affective Text. Among these, we selected SemEval 2018 for the multi-label classification task due to its richness in emotion diversity and prior use in shared tasks, and DAIR AI for the multi-class classification task because of its recency, balanced label distribution, and widespread use in research.

We also noted that DAIR AI emotions are a subset of SemEval emotions, which enabled the construction of a unified self-explaining model that predicts both the most prominent emotion and all secondary emotions from a single tweet, along with an explanation. This chapter established the foundational data necessary for training, evaluating, and preference-aligning models throughout the rest of the thesis.

In the next chapter, we will describe the methodology adopted in this work, including task definitions, evaluation metrics, modeling techniques, and the overall framework for classification and explanation generation.

Chapter 4

Methodology

4.1 Task Description

The multi-label emotion classification task assigns "neutral or no emotion" or one or more emotion labels shown in table 3.2 to each instance. It allows multiple emotions to be present, capturing the full range of emotional expression.

In contrast, the multi-class emotion classification task assigns each instance a single, dominant emotion label, from the emotions shown in table 3.6. This approach focuses on identifying the main emotion more simply.

Additionally, we explore explainability in emotion classification by generating explanations for model predictions using multiple methods, including SHAP, LIME, and Generative AI. Furthermore, we introduce our novel self-explaining model, which simultaneously predicts the multi-class dominant emotion and multi-label emotions present and provides detailed, contextually grounded explanations—all within a single unified model.

4.2 Evaluation Measures

4.2.1 Evaluation measures - Multi-class

The Dair AI dataset is a multi-class emotion classification dataset. We used common classification metrics: Accuracy, Precision, Recall, and F1 score to evaluate and compare our models with state-of-the-art results. These metrics are widely used in classification tasks to measure the effectiveness of models and are also the primary metrics used for leaderboard [1] rankings on this dataset. These metrics ensure a standardized evaluation of our model's performance against established benchmarks.

4.2.2 Evaluation measures - Multi-label

4.2.2.1 Jaccard score

In multi-label text classification, each instance has one or multiple gold emotion labels and one or multiple predicted emotion labels. Multi-label accuracy, also known as the Jaccard index, is characterized by the intersection of the predicted and gold tags divided by their union. This evaluation is performed for each sentence in the test dataset and then averaged over all instances in dataset D.

$$\text{Accuracy} = \frac{1}{D} \sum_{s \in D} \frac{G_s \cap P_s}{G_s \cup P_s} \tag{4.1}$$

where Gs is the set of gold tags for tweets, Ps is the set of predicted tags for tweets, and D represents the total number of tweets in the test set. In addition to the Jaccard index

[1] https://paperswithcode.com/sota/text-classification-on-emotion

(multi-label accuracy), we evaluated our models using the macro-averaged F-score and the micro-averaged F-score (Ameer et al., 2020) [6]. Equations to compute both scores can be found int eh following sections and, on the SemEval-2018 [74] competition's web page.

4.2.2.2 Micro F1

The micro-averaged F1 score aggregates the contributions of all emotion classes to compute the overall precision and recall before calculating the F1 score. It treats each prediction equally, regardless of its class, which is particularly useful when class distributions are imbalanced.

$$\text{Micro-avg Precision (micro-P)} = \frac{\sum_{e \in E} \text{TP}_e}{\sum_{e \in E} (\text{TP}_e + \text{FP}_e)} \tag{4.2}$$

$$\text{Micro-avg Recall (micro-R)} = \frac{\sum_{e \in E} \text{TP}_e}{\sum_{e \in E} (\text{TP}_e + \text{FN}_e)} \tag{4.3}$$

$$\text{Micro-avg F1} = \frac{2 \times \text{micro-P} \times \text{micro-R}}{\text{micro-P} + \text{micro-R}} \tag{4.4}$$

where TP_e is the number of tweets correctly assigned to emotion class e, FP_e is the number of tweets incorrectly assigned to emotion class e, and FN_e is the number of tweets from class e that were missed. E represents the set of all emotion labels.

4.2.2.3 Macro F1

The macro-averaged F1 score computes precision and recall independently for each class and then averages them, treating all classes equally regardless of their frequency. This metric is more sensitive to how the model performs on less frequent emotion classes.

$$P_e = \frac{\mathrm{TP}_e}{\mathrm{TP}_e + \mathrm{FP}_e} \tag{4.5}$$

$$R_e = \frac{\mathrm{TP}_e}{\mathrm{TP}_e + \mathrm{FN}_e} \tag{4.6}$$

$$F_e = \frac{2 \times P_e \times R_e}{P_e + R_e} \tag{4.7}$$

$$\text{Macro-avg F1} = \frac{1}{|E|} \sum_{e \in E} F_e \tag{4.8}$$

where P_e, R_e, and F_e represent the precision, recall, and F1 score for each emotion class e, respectively, and $|E|$ is the total number of emotion classes.

4.2.2.4 Exact match

In addition to the Jaccard score, Micro and Macro F1, we evaluate our multi-label emotion classification using the Exact Match measure, a more stringent evaluation metric. Unlike the Jaccard score, which considers the intersection of predicted and actual labels, the Exact Match measure only regards a prediction as correct if, and only if, the predicted set of emotion labels exactly matches the gold labels for a given tweet. Formally, the Exact Match accuracy can be expressed as:

$$\text{Exact Match Accuracy} = \frac{1}{|D|} \sum_{s \in D} \mathbb{I}(G_s = P_s) \tag{4.9}$$

where G_s is the set of gold labels for tweet s, P_s is the set of predicted labels, D is the test dataset, and $\mathbb{I}(\cdot)$ denotes the indicator function, returning 1 when the predicted labels exactly match the gold labels, and 0 otherwise. This metric heavily penalizes partially correct predictions, providing a rigorous evaluation of model performance on multi-label emotion classification tasks.

4.2.3 Evaluation measure - Explanations

4.2.3.1 Automatic measure

To automatically evaluate the quality of generated explanations, we adopt the Sufficiency metric from the Faithfulness-by-Construction (FRESH) framework [53]. Sufficiency aims to determine whether an explanation alone—without access to the original input—indicates the model's predicted label [51] [122]. The core idea is that a faithful and informative explanation should contain enough signal for a separate classifier to recover the model's decision.

In our implementation of FRESH, a BERT-based classifier [27] is trained on the explanation texts alone to perform the emotion classification task. If the classifier achieves high predictive accuracy using only these explanations, it is taken as evidence that they are sufficient, i.e., they encapsulate the critical reasoning behind the original model's prediction.

We apply this framework to evaluate explanations generated by our self-explaining model and compare them against baseline explanation methods such as LIME, SHAP, and zero-shot generative outputs from GPT-4o. A higher sufficiency score indicates better alignment between the explanation and the model's decision, making this an adequate proxy for explanation quality in an automated setting. This measure has been used for evaluating explanations in existing research [90] [34] [104].

4.2.3.2 Human evaluation

In addition to automatic evaluation using the sufficiency metric, we conducted a comprehensive human evaluation of the generated explanations to assess their interpretability and usefulness from a human-centered perspective. Humans evaluated random 300 explanations for each of the three models (GPT-4o, LLAMA, DeepSeek) for both pre-trained and preference-aligned models (6 models in total: 1x3 pre-trained, 1x3 preference aligned), scoring them based on four key qualitative dimensions:

- Correctness: Evaluates whether the response's predicted emotions align with the gold labels and whether the explanation logically supports those labels.
- Clarity: How clear and unambiguous the explanation is. Does the explanation use concise language and avoid vague or contradictory statements?
- Helpfulness: How effectively the explanation explains the reasoning behind the predicted emotions, highlighting salient parts of the tweet.
- Verbosity: How the explanation balances detail with conciseness—avoiding being too short (missing information) or too long (unnecessarily repetitive).

Each explanation is rated on a scale of 1–5 for each dimension.

This human evaluation process provides a robust assessment of explanation quality beyond what automatic metrics can capture, ensuring that the models produce faithful, understandable, and actionable explanations for end users.

4.2.3.3 Human evaluators

Three individuals conducted the human evaluation of the generated explanations. Two of the evaluators were volunteers who are professional data engineers and computer science graduates. Importantly, these two volunteers also served as annotators for the original preference dataset, discussed later in the thesis, giving them direct familiarity with the emotion classification and explanation task and its requirements. The third evaluator was the author of this thesis. This combination ensured that both prior annotation experience and subject-matter expertise informed the evaluation.

4.2.3.4 Why not other automatic measures?

There exist numerous evaluation techniques for explainability in XAI, as described by Pawlicki et al. [81], including faithfulness [4], sensitivity [108], and trustworthiness [69], among others. While each method offers valuable perspectives on different aspects of model interpretability, there is currently no universally accepted standard for evaluating explanations [81]. This diversity reflects the interdisciplinary and evolving nature of XAI evaluation, where different approaches emphasize distinct priorities.

Most established frameworks tend to focus on uncovering the inner workings of models, providing insights that are particularly useful for engineers and machine learning researchers. These lower-level explainability techniques are well-suited for tasks such as model diagnosis and debugging. In contrast, socio-behavioral evaluation methods prioritize explanation intelligibility and usefulness for non-technical end-users [83].

In our research, we generate text-based explanations that are not post-hoc, but are generated by the model at the time of predicting labels. These explanations are designed to be consumed by both end-users and technical experts. Such explanations can help engineers diagnose models as they can tell if there is a bias in the model's output. For this purpose, we employ the sufficiency metric [23] as an automatic, model-agnostic measure of explanation quality. In addition, we conduct a detailed human evaluation of the generated explanations, assessing them on correctness, clarity, helpfulness, and verbosity. This approach allows us to capture both the objective and subjective qualities of explanations, aligning evaluation with the needs and expectations of real-world users.

4.3 Transformer Models for Emotion Classification

Due to the success of the context-aware transfer learning models [112], such as BERT, RoBERTa, XLNet, etc., have been very effective in setting new state-of-the-art in numerous supervised NLP problems, such as language inference [20], language understanding [27], and machine translation [112]. We used the Transformers library by HuggingFace to use an improved version of BERT, which is RoBERTa (Robustly optimized BERT approach) [68]. Using the pre-trained RoBERTa model, we fine-tuned the final layer on our emotion classification corpus.

4.3.1 Hyperparameter settings

We used the hyperparameters defined in the original transformer papers for the transformer models. For the models without an attention mechanism (XLNet, DistilBERT, and RoBERTa), we used a learning rate of 0.01; the number of epochs is 20. For the models with multiple-attentions (XLNet-MA, DistilBERT-MA, RoBERTa-MA), the learning rate is 0.001, and the number of epochs is 25.

4.4 Instruction prompts

This part of the research aims to develop a model for emotion classification in tweets using fine-tuning techniques on the GPT architecture. To achieve this, a specialized instruction-tuning approach is used, where instruction prompts are crafted to guide the model in generating responses that indicate the presence or absence of a specific emotion in the tweet.

4.4.1 Prompt structure

The instruction prompts utilized in the fine-tuning process are structured using specific tags to outline different input components. Three key pairs of tags are employed:

1. **<TWEET_STARTS>** and **<TWEET_ENDS>:** These tags encapsulate the tweet text, providing the model with contextual information.
2. **<QUESTION_STARTS>** and **<QUESTION_ENDS>:** Encompassing the question regarding the emotion expressed in the tweet, these tags guide the model towards understanding the specific emotion query.
3. **<ANSWER_STARTS>** and **<ANSWER_ENDS>:** This pair of tags define the boundaries within which the model generates its response, representing either a 'yes' or 'no' depending on the presence or absence of the specified emotion in the tweet.

   In the results table **??**, we show results for three different proposed models, i.e., $GPT2-IT_A$,$GPT2-IT_B$ and $GPT2-IT_C$ (IT stands for Instruction Tuning). We experimented with a number of instruction formats to fine-tune the GPT model and report the top three best-performing ones in this paper. All these models use the same GPT2-medium pre-trained model from the huggingface library and fine-tuned it on different prompt formats. Model A, the best performing one, has its structure elaborated in the section below. Following are the formats for the three models:

- $GPT2{-}IT_A$: <TWEET_STARTS> **tweet** <TWEET_ENDS> <QUESTION_STARTS> **question** <QUESTION_ENDS> <ANSWER_STARTS> **Yes/No** <ANSWER_ENDS>
- $GPT2 - IT_B$: <SENTENCE> **tweet** <SENTENCE> <EMOTION> **emotion** <EMOTION> <PRESENT> **Yes/No** <PRESENT>
- $GPT2 - IT_C$: <TQA_START> <TWEET> **tweet** <QUESTION> **question** <ANSWER> **Yes/No** <TQA_END>

We opted to use the GPT2-medium model which is a considerably sized version compared to the bigger GPT2-large and GPT2-XL.

4.4.2 Fine-tuning process

The GPT-2 model is fine-tuned using the entire instruction prompt in a generative setting. The training dataset comprises tweets annotated with multiple emotions, providing a diverse range of examples for the model to learn from. During the fine-tuning process, the model adjusts its parameters to predict the next word in the sequence better, handling the relationships between tweets, emotion-related questions, and corresponding answers.

4.4.3 Inference mechanism

For inference, a specialized generation approach is employed. The model is prompted to generate the next word after encountering the <ANSWER_STARTS> tag. The generated output corresponds to either 'yes' or 'no,' indicating the model's prediction regarding the

Setup	Emotion	Emotion present	Prompt
Fine-tuning	Anger	Yes	<TWEET_STARTS> Unruly kids at am in the morning nothanks ripping the flower beds up by the roots while their parents watch shocking <TWEET_ENDS> <QUESTION_STARTS> Is the person in the tweet angry? <QUESTION_ENDS> <ANSWER_STARTS> Yes <ANSWER_ENDS>
Testing	Anger	No	<TWEET_STARTS> Going to get myself copy of StephenKing CUJO for an upcoming project that can talk about just yet am writing <TWEET_ENDS> <QUESTION_STARTS> Is the person in the tweet angry? <QUESTION_ENDS> <ANSWER_STARTS> **model generates the answer here**

Table 4.1: Example prompts for fine-tuning and testing setups. Similar prompts are generated for each emotion.

presence of the queried emotion in the given tweet. Table 4.1 provides examples of fine-tuning and inference prompts.

4.4.4 Experimental settings

We create instruction-tuning prompts based on the emotions in our dataset. We use a batch size of 16 during training and train the model over three epochs. The AdamW optimizer with a learning rate of 3e-5 is used to update the model's parameters, along with 5000 warm-up steps for stability. After training, the model is saved and later used for evaluation on the test set, where it predicts the emotions in unseen tweets. Table 4.1 illustrates example prompts for fine-tuning and testing.

4.5 Zero-shot Classification

For the zero-shot experiments, a prompt was crafted to guide the GPT-4o model in classifying tweets into one or more predefined emotions. The prompt structure is designed to provide the necessary context and instructions for the model to understand the task requirements. The prompt used is shown in table 4.2.

This prompt template includes placeholders for the tweet text and the list of emotions to classify. It instructs the model to analyze the tweet text and determine the presence or absence of each emotion listed, assigning the corresponding emotion labels accordingly. Each tweet is formatted as per the prompt structure, with the tweet text inserted into the designated placeholder. The formatted prompts, along with the tweet texts, are submitted to the models for inference.

4.6 Few-shot Classification

Few-shot learning enables models to efficiently learn from a minimal number of labeled examples, often as few as one to five instances per class. This approach is significant for text classification tasks using LLMs to provide examples from the training set to the model inside the prompt so that the model can see examples of annotated data during inference time.

4.6.1 Vanilla few-shot

In the vanilla few-shot setting, we provide the model with k annotated examples from the training set (Multi-label) within the prompt, ensuring that these examples collectively cover

Setup	Prompt
Zero-shot multi-class	Here is a tweet: [Insert Tweet Text Here]. Classify the following tweet into one of the following emotions depending on the presence of that emotion in tweet text. Emotions: ’anger’, ’fear’, ’joy’, ’love’, ’sadness’, ’surprise’. Limit the response to only the emotions.
Zero-shot multi-label	Here is a tweet: [Insert Tweet Text Here]. Classify the following tweet into one or more of the following emotions depending on the presence of that emotion in tweet text. Emotions: ’anger’, ’anticipation’, ’disgust’, ’fear’, ’joy’, ’love’, ’optimism’, ’pessimism’, ’sadness’, ’surprise’, ’trust’. Limit the response to only the emotions.
Few-shot	Your task is to analyze the content of the given tweet and classify it based on the following emotion labels: Emotion Labels: [anger, fear, love, surprise, sadness, joy, anticipation, disgust, optimism, pessimism, trust, neutral] Tweet: "" Instructions: 1. Analyze the Content: Thoroughly analyze the tweet’s content, tone, and context. Look for linguistic cues, such as words, phrases, or implied sentiments that might indicate one or more emotions. 2. Chain of Thought Reasoning: Break down the tweet’s components step by step. Consider the following questions as part of your reasoning: 2a. What is the speaker expressing? 2b. Does the speaker show a positive, negative, or neutral sentiment? 2c. Are there hints of longing, hope, frustration, or any other specific emotion? 2d. Is there any combination of emotions suggested by the words or tone? 3. Classify Emotions: Use the reasoning process to identify: 3a. The Most Prominent Emotion: Select the single most dominant emotion that captures the essence of the tweet. 3b. Other Present Emotions: List other emotions that might be present in the tweet, even if they are secondary. 4. If no emotions are present, then put neutral in both the most prominent emotion and other present emotions. Here are some examples of similar multilabel classifications: Examples: Example 1: Tweet: "" - multiple_emotions_present: "" Example 2: Tweet: "" - multiple_emotions_present: "" Example 3: Tweet: "" - multiple_emotions_present: ""
System prompt	You are an expert emotion classifier that accurately identifies emotions in tweets.

Table 4.2: The used zero-shot, few-shot, and system prompts.

all possible emotion categories. Specifically, we perform experiments with small batches (to see what value of k is the best), classifying 100 tweets each for three different prompt structures (refer to Appendix) while varying the number of provided examples (k) as 3, 5, and 10. These experiments showed us that a value of $k = 3$ yielded the most effective performance, balancing prompt complexity with predictive accuracy. The best-performing few-shot prompt is shown in table 4.2.

4.6.2 Dynamic few-shot

Dynamic few-shot learning (DFSL) represents an innovative adaptation of traditional few-shot learning techniques, emphasizing the dynamic selection of relevant examples during inference rather than relying on fixed, pre-selected instances. This evolving approach uses similarity metrics, often derived from embeddings or semantic analysis, to identify the most pertinent examples relative to each query, thus enhancing the classifier's performance by providing contextually relevant information tailored to the specific instance. Recent research has illustrated the efficacy of DFSL in various text classification tasks. For instance, Geng et al. proposed Dynamic Memory Induction Networks, which enable the dynamic selection of relevant memory instances based on contextual similarity, successfully improving classification accuracy in few-shot scenarios [37]. Moreover, Chen et al. introduced ContrastNet, which employs a contrastive learning framework to determine the similarity among examples dynamically, achieving significant performance gains in few-shot text classification tasks [15]. By adapting to the context and dynamically selecting instances, these models often outperform traditional few-shot approaches, highlighting the importance of context-aware selections in enhancing model robustness and accuracy across diverse natural language processing applications, as further supported by Liu and Yang, who discuss advancements in few-shot learning for various NLP tasks [67]. Algorithm 4.1 shows the pseudocode of the Dynamic few-shot approach

Algorithm 4.1 Dynamic Few-shot Prompt Generation using Cosine Similarity

Input: Input tweet T_{test}, training dataset D_{train}, embedding model, k (number of examples)

Output: Few-shot prompt P containing k most similar examples

Step 1: Embed Training Data
for all tweet $t_i \in D_{\text{train}}$ **do**
    Compute embedding $v_i = \text{Embed}(t_i)$
    Store $(v_i, t_i, \text{label}_i)$ in vector database
end for
Step 2: Embed Test Tweet
$v_{\text{test}} \leftarrow \text{Embed}(T_{\text{test}})$
Step 3: Compute Cosine Similarity
for all $(v_i, t_i, \text{label}_i)$ in vector database **do**
    Compute $\text{sim}_i = \text{cosine}(v_{\text{test}}, v_i)$
end for
Retrieve top-k examples with highest similarity scores
Step 4: Construct Prompt
Initialize prompt $P \leftarrow$ empty string
for all retrieved pair (t_j, label_j) **do**
    Append example to prompt:
        `"Tweet:` t_j `\n Labels:` $\texttt{label}_j$`"`
end for
Append test tweet instruction to prompt:
    `"Tweet:` $T_{\texttt{test}}$ `\n Labels:"`
return Few-shot prompt P

4.6.2.1 Our approach

Relevant examples are dynamically selected for each inference tweet based on similarity in embeddings. Given a new tweet to classify, we generate its embedding vector. Using cosine similarity, we retrieve the most similar tweets and their emotion labels from the training set, and append them to the model's prompt. By leveraging semantic similarity and testing on various embedding models, our dynamic few-shot method consistently outperformed traditional zero-shot and vanilla few-shot approaches, demonstrating its effectiveness in providing richer context and enhancing classification accuracy. We used the same prompt as vanilla few-shot learning 4.2 for dynamic few-shot, formatting the prompt with the top 3 most relevant tweets.

4.6.2.2 Embedding models used

To facilitate dynamic few-shot learning, we experimented with the following three embedding models to capture semantic similarities effectively:

- Used BerTweet [77]; a natural choice as it is a transformer-based BERT model trained explicitly on Twitter data, making it well-suited for capturing tweet-specific semantics. The embedding size for BerTweet is 768, extracted from the last hidden state.
- Used embeddings generated by SentenceTransformers [92] model "all-mpnet-base-v2", a widely adopted model known for semantic similarity tasks, especially prominent in Retrieval-Augmented Generation (RAG) systems due to its ability to encode contextual meaning. The embedding size for sentence transformers is 768.
- Used the OpenAI embeddings [2] model (Large), recognized for its generalizable representations, providing comprehensive semantic understanding across textual contexts. The embedding size is 3072.

[2] https://platform.openai.com/docs/guides/embeddings

4.7 Explainable AI

To enhance the interpretability of emotion-classification models, we explore a range of Explainable AI (XAI) techniques that provide insight into model predictions. This section presents three categories of explanation methods employed in this thesis: (i) post-hoc explanation methods, which interpret predictions of pre-trained models by analyzing input-output perturbations; (ii) generative explanations produced by prompting large language models like GPT-4o in a zero-shot setting; and (iii) our proposed self-explaining model, which is designed to generate emotion predictions along with natural language explanations jointly. Together, these approaches allow us to compare traditional interpretability techniques with the emerging capabilities of generative LLMs, both in standalone and preference-aligned settings.

4.8 Post-hoc Explainers

To generate post-hoc explanations for model predictions, we used two widely adopted explainability techniques: LIME and SHAP, employing their respective Python libraries.

First, we fine-tuned a RoBERTa model on the training set of the SemEval-2018 Task E-c multilabel emotion classification dataset. This fine-tuned model served as the base classifier whose predictions were subsequently explained using LIME and SHAP.

LIME operates by constructing a locally interpretable model around each prediction. Specifically, we used the LimeTextExplainer object's explain_instance function to perturb the input text and observe how small changes influence the model's output. The features identified as important are enumerated to create human-readable explanations, highlighting whether each feature steers the model toward a positive or negative prediction.

SHAP uses a game-theoretic framework based on Shapley values to attribute contributions to each feature. We utilized the KernelExplainer method to compute Shapley values for individual predictions made by the fine-tuned RoBERTa model. These values enabled us to construct explanations that rank feature importance and quantify each feature's impact on the prediction, offering deeper insight into the model's reasoning.

For both LIME and SHAP, we extracted features (which correspond to words in the tweet) whose importance scores exceeded a pre-defined threshold of 0.4. These selected features were concatenated into a single string, which served as the explanation text generated by each method. This approach allowed us to standardize the explanations for subsequent evaluation.

It is important to note that the way we constructed these explanation strings—by thresholding and concatenation—could influence evaluation outcomes. For example, setting the threshold at 0.4 balances between including relevant features and avoiding noise, but alternative thresholds or aggregation methods might yield explanations with different granularity or interpretability.

4.9 Generative AI Explanations Using Pre-trained Models

For generative AI explanations, we used the GPT-4o chat completions API. In the same request used for classification, we added a prompt asking the model to provide a brief explanation in plain English for its choice of emotions. Combining classification and explanation in a single step enables the model to generate natural language explanations, making its decisions more interpretable and easier to understand. Without the few-shot examples, we used the few-shot prompt in table 4.2. Table 4.3 shows examples of explanations generated through different explainable AI models. The outputs of SHAP and LIME are numerical weights for each word given an emotion. Just for this table, we made the explanations for these models human-readable for better interpretation.

XAI model	Explanation
SHAP	For the emotion labels 'joy', 'optimism' and 'love', the word 'smiling' has the highest score. The phrases 'I'm doing' and 'make sure' also contribute slightly to these emotions
LIME	For the emotion labels 'joy' and 'optimism', the word 'smiling' and the phrase 'make sure' have the highest score.
GPT-4o	I'm doing all of this to ensure that you're smiling down on me, bro. This expression conveys a mix of emotions: anticipation, joy, optimism, sadness, and trust. Anticipation is evident as I work toward a future outcome. Joy and optimism are reflected in my desire to bring happiness to you. Sadness is implied by the fact that you are not present. Lastly, trust is shown in my belief that you are watching over me.
Tweet	I'm doing all this to make sure you smiling down on me bro

Table 4.3: Examples of explanations generated through different explainable AI models

4.10 Self-explaining Model

4.10.1 What is a self-explaining model, and why do we need one?

A self-explaining model is an inherently interpretable model [90] designed to simultaneously produce predictions and clear, human-readable explanations that articulate the reasoning behind each decision. In emotion classification, where accurately interpreting nuanced human sentiments is crucial, a self-explaining model explicitly clarifies which parts of the input text contribute to recognizing specific emotions, greatly enhancing transparency. Such interpretability is vital because traditional neural models, particularly large language models, tend to be opaque and susceptible to biases or inaccuracies like hallucinations, thereby reducing user trust. By clearly communicating how and why a particular emotion was identified, self-explaining models enable stakeholders—such as researchers, organizations, and policymakers—to trust model outputs, verify their reliability, and understand complex emotional dynamics in social media content, ultimately leading to more effective decision-making.

4.10.2 What should explanations/output look like?

4.10.2.1 Explanation quality

To ensure transparency and interpretability, the explanations generated by our models are expected to be structured in a way that closely mimics the classification decisions. Specifically, a high-quality explanation should explicitly reference key phrases or linguistic cues from the input tweet and logically connect them to the corresponding predicted emotions. For example, if a tweet includes phrases such as "I'm scared" or "so anxious," the explanation should identify these as indicators of fear and articulate that reasoning clearly. This connection between the input text and the predicted emotion enables human evaluators to understand not just what the model predicted, but why.

All generated explanations are evaluated based on four key qualities:

- Correctness: Evaluates whether the response's predicted emotions align with the gold labels, and whether the explanation logically supports those labels.

- Clarity: How clear and unambiguous the explanation is. Does the explanation use concise language and avoid vague or contradictory statements?
- Helpfulness: How effectively the explanation explains the reasoning behind the predicted emotions, highlighting salient parts of the tweet.
- Verbosity: How the explanation balances detail with conciseness—avoiding being too short (missing information) or too long (unnecessarily repetitive).

4.10.2.2 Output

We define a structured output response for our self-explaining model, so it can be consistent and parsed using JSON. Here is a sample response. Table 4.4 shows an example output for one of the tweets from the SemEval 2018 dataset.

Field	**Value**
ID	2018-En-12345
Tweet	There's no one right way to live life but: be kind, be genuine, be honest, be open. #Love yourself & #laugh often.
Gold multilabel	['joy', 'love', 'optimism']
Output	**"explanation"**: "The tweet emphasizes positivity and self-care, which are indicative of emotions like love and joy. The use of words such as 'be kind,' 'be genuine,' and 'love yourself' strongly suggests love as the primary emotion. The encouragement to 'laugh often' and the hashtag #Mindfulness contribute to a sense of joy and optimism. The overall tone is uplifting and encouraging, promoting a positive outlook on life.", **"most_prominent_emotion"**: "love", **"multiple_emotions_present"**: ["love", "joy", "optimism"]

Table 4.4: Output example from the self-explaining model

We put the explanation section first because it gives the model a chain-of-thought reasoning approach to describing the emotions and then generating the prediction fields. The response works for multi-class (most_prominent_emotion) and multi-label (multiple_emotions_present).

4.10.3 How? Pre-trained LLMs

As a first step toward building a self-explaining emotion classification system, we explore the capabilities of pre-trained large language models (LLMs) to generate emotion predictions and accompanying natural language explanations in a zero-shot setting. For this purpose, we selected three state-of-the-art pre-trained models.

We use GPT-4o, OpenAI's most advanced publicly accessible model, through its Chat Completions API. GPT-4o can generate structured outputs and demonstrates strong performance in reasoning and natural language generation. To benchmark its performance against open-source alternatives, we also include two high-performing instruction-tuned models: Meta-Llama-3.1-8B-Instruct-bnb-4bit [3] and DeepSeek-R1-Distill-Qwen-1.5B-unsloth-bnb-4bit [4]. These models were selected based on their ranking on the Hugging Face Open LLM Leaderboard [5] within their respective parameter size categories. Furthermore, both models are supported by the Unsloth [6] library, enabling faster inference and fine-tuning with reduced memory overhead through 4-bit quantization.

All three models are prompted in a structured format to behave as self-explaining models, generating outputs that include: (i) an explanation of their decision, (ii) the most prominent emotion (for multi-class classification), and (iii) the set of multiple emotions

[3] https://huggingface.co/unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit
[4] https://huggingface.co/unsloth/DeepSeek-R1-Distill-Qwen-1.5B-unsloth-bnb-4bit
[5] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/
[6] https://docs.unsloth.ai/get-started/all-our-models

present (for multi-label classification), as described in table 4.4.

These outputs adhere to a predefined schema that ensures consistency and interpretability. This zero-shot evaluation is a baseline for comparison with models fine-tuned via supervised and preference-aligned approaches. We use our best-performing prompt, i.e., the dynamic few-shot prompt from table 4.2.

4.10.4 Making LLAMA and DeepSeek as good as GPT-4o

As shown in Table 5.1, GPT-4o outperforms all other models in our self-explaining emotion classification task. This is expected, as GPT-4o is not only the largest model among those evaluated, but also one of the most powerful and instruction-aligned models currently available—often serving as the benchmark for comparing open-source alternatives. However, due to its closed-source nature and high usage costs, it is not always practical for large-scale or production-level deployment. To address this, we aim to tune smaller, open-source models like LLAMA and DeepSeek to mimic GPT-4o's behavior and performance. Specifically, we explore two strategies to achieve this: (i) Supervised Fine-tuning, and (ii) Preference Alignment. In the following sections, we describe these techniques in detail and explain how we implemented them to improve the performance and interpretability of smaller models using GPT-4o as a reference.

4.10.5 How? Supervised fine-tuning

To enable smaller open-source models to mimic the high-quality responses of GPT-4o, we employ a supervised fine-tuning (SFT) approach. The goal is to train models like LLaMA 3 (8B) and DeepSeek-R1 (Distilled Qwen 2.5, 32B) to replicate the explanation and classification outputs generated by GPT-4o. For this, we utilize the SFTTrainer module from the TRL (Transformers Reinforcement Learning) library, which provides an efficient interface for fine-tuning large language models on input-output pairs.

Rather than updating the full set of model parameters—which would be computationally expensive and potentially lead to overfitting or catastrophic forgetting—we apply Low-Rank Adaptation (LoRA) techniques. LoRA allows us to fine-tune only a small subset of trainable weights by injecting low-rank matrices into the transformer layers, significantly reducing memory usage and improving training stability.

We construct the fine-tuning dataset by collecting the prompts and corresponding structured responses generated by GPT-4o on the training split. These input-output pairs are then used to train LLaMA and DeepSeek models in a supervised fashion, enabling them to learn to generate explanations, the most prominent emotion, and multiple emotions present—mimicking the behavior of GPT-4o as a self-explaining model.

4.10.6 How? Alignment

Developing self-explaining architectures aims to incorporate interpretability into neural text classifiers inherently, [89]. The advancements in AI, particularly with the emergence of LLMs, have transformed various applications, but making these models behave according to task-specific needs remains challenging. Preference alignment algorithms have gained attention as a means to align AI models with human preferences, ensuring that AI outputs are more relevant and acceptable to users [72]. Various preference alignment techniques, such as reinforcement learning from human feedback (RLHF), have been instrumental in fine-tuning models for applications ranging from content moderation to personalized recommendations [72]. These algorithms enable models to learn from user interactions, thereby enhancing their performance in real-world scenarios [72]. Despite advancements in LLMs, uncontrolled generation remains a significant challenge, often leading to hallucinations—instances where the model generates false or nonsensical information. This phenomenon can undermine the reliability of AI systems, particularly in applications requiring factual accuracy, such as news generation or medical advice [14] [63]. The literature indicates that hallucinations can arise from various factors, including insufficient training data and the inherent complexity of language generation tasks [14] [63]. Addressing this issue is critical to ensuring the safe deployment of LLMs in sensitive applications. Preference alignment enhances the relevance of AI outputs and can be leveraged to create self-explaining models with controlled quality explanations. By aligning model behavior with user preferences, developers can ensure that the explanations generated by AI systems are not only accurate but also tailored to the user's context and needs [72]. This approach can significantly improve user trust and satisfaction, as users are more likely to engage with systems that provide clear and relevant explanations for their outputs [72]. Integrating preference alignment in self-explaining models represents a promising direction for research, aiming to bridge the gap between complex AI systems and user understanding.

4.10.7 Preference dataset for alignment

A core contribution of this thesis is the creation of a preference dataset [7] tailored to emotion classification and explanation. In the context of LLMs, a preference dataset consists of multiple candidate responses for the same input, accompanied by explicit human judgments of which response is preferred. These human judgments provide the model with valuable signals on what constitutes a "better" output, enabling fine-tuning toward outputs more closely aligned with human expectations.

[7] https://huggingface.co/datasets/imhmdf/ExplainableAI-emotions-DPO-ORPO-RLHF

4.10.7.1 Base dataset and prompting strategy

We utilized the GPT-4o model to create a synthetic dataset, following a methodology similar to the Self-Instruct [115] approach. We start with the SemEval 2018-Ec multilabel emotion classification dataset. Specifically, we use the training and validation splits (7,724 tweets in total). For each tweet, we prompt GPT-4o (detailed system and user prompts are included in the Appendix) to generate two candidate responses. Each response is structured in a JSON-like format containing the fields shown in table 4.5. The explanation field contains a concise but informative rationale linking textual elements in the tweet to the predicted emotions. The most_prominent_emotion field indicates the single dominant emotion in the tweet, while multiple_emotions_present enumerates all predicted emotions.

4.10.7.2 Correctness criterion and dataset filtering

To ensure high-quality preference samples, we include only those pairs of responses where at least one of the two responses meets our correctness threshold. A response is deemed correct if (1) the multiple_emotions_present field exactly matches the gold set of emotions for that tweet, and (2) the most_prominent_emotion is one of the correctly identified emotions. Out of the 7,724 tweets, 1,063 met these criteria-constituting our final preference dataset.

4.10.7.3 Human annotation and scoring

Each of the 1,063 pairs of GPT-4o responses was then manually labeled by two domain experts and one adjudicator to decide on conflicts (details on annotation guidelines are in the Appendix). The annotators scored both responses along the following four dimensions:

1. Correctness: Evaluates whether the response's predicted emotions align with the

gold labels, and whether the explanation logically supports those labels. This is the highest-priority dimension, particularly when deciding which response is ultimately "preferred."

2. Clarity: Measures how unambiguous and straightforward the explanation is. Vague or contradictory responses receive lower scores.

3. Helpfulness: Assesses how effectively the explanation explains the reasoning behind the predicted emotions, highlighting salient parts of the tweet.

4. Verbosity: Gauges whether the explanation is sufficiently detailed without being excessively long or repetitive. Overly concise or unnecessarily verbose explanations receive lower scores.

5. Overall: An overall assessment of each response, considering the four specific dimensions (Clarity, Correctness, Helpfulness, Verbosity).

Each dimension is rated on a scale of 1 to 5, and an overall rating (1–5) summarizes the annotators' judgment across all four dimensions for each of two responses for each tweet. While these scores were not strictly required by the specific preference alignment algorithm that we employed [45] [88], they provide rich metadata for future work in reinforcement learning from human feedback (RLHF), Group Relative Policy Optimization (GRPO), and other reward-based preference-optimization approaches. During the annotation process, we also fixed minor errors in the chosen response to perfect its correctness, clarity, helpfulness, and verbosity. Table 4.5 shows an annotated example from our preference dataset.

Field	Value
ID	2018-En-12345
Tweet	There's no one right way to live life but: be kind, be genuine, be honest, be open. #Love yourself & #laugh often.
Gold multilabel	['joy', 'love', 'optimism']
Response 1	**"explanation"**: "The tweet promotes positivity and well-being, focusing on being kind, genuine, and open. The use of hashtags like #Love and #laugh often, along with the mention of mindfulness, suggests a strong emphasis on love and joy. **"most_prominent_emotion"**: "love", **"multiple_emotions_present"**: ["love", "joy", "trust", "optimism"]
Response 2	**"explanation"**: "The tweet emphasizes positivity and self-care, which are indicative of emotions like love and joy. The use of words such as 'be kind,' 'be genuine,' and 'love yourself' strongly suggests love as the primary emotion. The encouragement to 'laugh often' and the hashtag #Mindfulness contribute to a sense of joy and optimism. The overall tone is uplifting and encouraging, promoting a positive outlook on life.", **"most_prominent_emotion"**: "love", **"multiple_emotions_present"**: ["love", "joy", "optimism"]
Chosen response	2 (according to correctness)
Response 1 ratings	Correctness[1-5], Clarity[1-5], Helpfulness[1-5], Verbosity[1-5], Overall[1-5]
Response 2 ratings	Correctness[1-5], Clarity[1-5], Helpfulness[1-5], Verbosity[1-5], Overall[1-5]
Chosen response fixed	Yes or No (if the chosen response was manually fixed for errors)

Table 4.5: Example of annotated response pairs from our preference dataset

4.10.7.4 Annotators

The two primary annotators involved in the scoring process are graduate-level computer scientists with domain knowledge in natural language processing and real-world experience as professional data engineers in the industry. Their technical background and familiarity with language model outputs enabled them to evaluate explanation quality across multiple dimensions critically. Their involvement in this annotation exercise is as volunteers. The third member of the annotation team, serving as the adjudicator in cases of disagreement, is the author of this thesis. The adjudicator reviewed all flagged instances and resolved any annotation conflicts to ensure consistency and reliability in the final preference dataset.

4.10.7.5 Annotator agreement - Krippendorff's Alpha (α)

To ensure the reliability and consistency of the human-annotated preference dataset, we computed inter-annotator agreement between the two primary annotators across all four evaluation dimensions—Correctness, Clarity, Helpfulness, and Verbosity—for both response 1 and response 2 in each annotation pair.

We used Krippendorff's Alpha (α) [62] as the statistical measure of agreement. This metric is well-suited for our setting because it supports ordinal data, such as the 1–5 Likert-scale ratings used in this study. Krippendorff's Alpha also accounts for the magnitude of disagreement, making it more informative than nominal agreement measures for subjective, graded annotations.

Other standard agreement metrics, such as Cohen's Kappa, are not suitable in this context [9], as they are designed for categorical (nominal) data and assume a binary or unordered label space. They do not capture the ordinal relationships between scores (e.g., the fact that a disagreement between 4 and 5 is less severe than one between 1 and 2), which is critical for interpreting explanation quality scores.

Table 4.6 displays the Krippendorff's Alpha reliability scores for each annotation dimension—Clarity, Correctness, Helpfulness, and Verbosity—for both response 1 and re-

Response	Clarity	Correctness	Helpfulness	Verbosity	Overall
Response 1	0.65	0.97	0.58	0.59	0.48
Response 2	0.43	0.90	0.44	0.52	0.46

Table 4.6: Krippendorff's Alpha values indicating reliability score across all explanation quality dimensions for both responses

sponse 2. Krippendorff's Alpha ranges from -1 (systematic disagreement) to 1 (perfect agreement), with 0 indicating chance-level agreement. In practice, scores above 0.8 are considered strong, values between 0.67 and 0.8 are acceptable, and values between 0.4 and 0.67 reflect moderate, but still usable, reliability [79].

Correctness scores were extremely high—0.97 for Response 1 and 0.90 for Response 2—indicating near-perfect agreement among annotators. This is expected, as correctness was based on a relatively objective comparison: annotators judged whether the model's multilabel classification results matched the gold-standard labels in the dataset. The clear, rule-based nature of this dimension naturally led to stronger consensus.

For the more subjective dimensions—Clarity (0.65/0.43), Helpfulness (0.58/0.44), and Verbosity (0.59/0.52)—agreement was moderate, which is common for qualitative assessments that involve individual interpretation. Nonetheless, these values indicate a reasonable level of annotator consistency, supporting the usability of the dataset for preference alignment and evaluation.

The Overall dimension, which represents the annotators' holistic assessment of each response, yielded reliability scores of 0.48 (Response 1) and 0.46 (Response 2), indicating moderate agreement. This score captures the overall quality of each explanation as perceived by the annotators, integrating their impressions across all specific dimensions.

Taken together, these Krippendorff's Alpha results indicate that the annotation process provided very high reliability for correctness, moderate consistency for subjective dimensions, and acceptable agreement for overall response quality. This supports the robustness and trustworthiness of the preference dataset, ensuring its suitability for training and evaluating preference-aligned models in this thesis.

4.10.7.6 Data generation settings

We configured GPT-4o to produce two responses per tweet with moderate stochasticity to encourage diverse outputs. The prompt used to generate these responses is available in Section 1.1.2.4 of the Appendix. We set the model's temperature to 0.4 and top-p to 0.8, aiming to strike a balance between creativity and focus. Higher temperature or top-p values could introduce more varied yet potentially off-topic content; lower values might yield overly deterministic responses that lack richness. In practice, this configuration allowed us to gather responses that varied enough in style and detail to be meaningfully compared and scored by human annotators. To further ensure the generated responses conform to a consistent structure, we employed the OpenAI structured outputs utility. We defined the required JSON schema using Python's Pydantic library, and then passed that schema to the chat_completions function call. This setup enforced that GPT-4o produced outputs in our desired format and emotion values comprising explanation, most_prominent_emotion, and multiple_emotions_present—thereby streamlining both automatic parsing and human annotation.

We evaluated our model on our chosen datasets for the thesis. The DAIR AI dataset [97] contains six possible emotions—anger, fear, sadness, joy, disgust, fear—where each tweet is assigned a single most prominent emotion. In contrast, the SemEval 2018 E-c dataset [74] contains eleven possible emotions—anger, fear, sadness, joy, disgust, fear, optimism, pessimism, sadness, surprise, trust, neutral—and permits multiple emotions to co-occur within a single tweet. Accordingly, our model uses the most_prominent_emotion field to handle DAIR AI and the multiple_emotions_present field for SemEval 2018 E-c.

To ensure that the model's outputs conform to these respective label sets, we defined two Pydantic schema objects reflecting the allowable outputs for each dataset. These schemas were then passed as constraints to the Chat Completions API (via OpenAI's structured outputs utility). We could have included these restrictions solely within the model's prompt, but large language models often struggle to consistently adhere to textual format directives alone [42], [17], [3]. By providing an explicit schema, the model is programmatically constrained to produce outputs that align with the specified fields and label sets for each dataset, leading to more reliable and parsable results. The output schema, and algorithm used to generate the dataset is given in the algorithm 4.2 The resulting preference dataset comprises 1,063 tweet-responses pairs with comprehensive human evaluations. This dataset constitutes the foundation for the subsequent preference alignment of GPT-4o and other LLMs, guiding them to produce explanations and predictions that match ground-truth labels more accurately and better align with human notions of quality, clarity, and helpfulness.

4.10.8 Methodology used for alignment

In this subsection, we explore the methodologies and processes involved in training our models to align with human preferences. We begin by describing both Direct Preference Optimization (DPO) and Odds Ratio Preference Optimization (ORPO), highlighting their role in leveraging the preference dataset. We then detail our model training setup, hyperparameter choices, and overall training workflow.

Algorithm 4.2 Structured Prediction Using GPT-4o

Input: Tweet input T, system prompt, user prompt
Output: Explanation E, most prominent emotion e_{main}, emotion set E_{multi}

Define Enum Classes:
`SingleEmotion = {anger, fear, sadness, joy, love, surprise, neutral}`
`MultipleEmotion = {anger, fear, sadness, joy, love, surprise, anticipation, disgust, optimism, pessimism, trust, neutral}`
Define Output Schema:

```
  EmotionOutput:
    explanation: str
    most_prominent_emotion: SingleEmotion
    multiple_emotions_present: List[MultipleEmotion]
```

Compose system and user prompts with input tweet T
Call `chat.completions.parse()` with:

```
  model = "gpt-4o"
  messages = [system_prompt, user_prompt]
  response_format = EmotionOutput
  temperature = 0.4
```

Parse output to extract E, e_{main}, and E_{multi}
return E, e_{main}, E_{multi}

4.10.8.1 Direct Preference Optimization

DPO [88] is a pairwise preference alignment technique that leverages human feedback to guide model responses. Given two candidate responses A and B for the same input, along with a human-annotated preference, DPO adjusts the model parameters so that the reward for the preferred response is higher. Formally, the DPO loss can be written as:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(A,B)\sim D}\big[\log \sigma\big(r_\theta(A) - r_\theta(B)\big)\big], \tag{4.10}$$

$r_\theta(A)$ and $r_\theta(B)$ are the learned reward scores assigned by the model to responses A and B, respectively. The expression $\sigma(r_\theta(A) - r_\theta(B))$ gives the probability that the model prefers A over B under the current parameters, where $\sigma(\cdot)$ is the logistic (sigmoid) function. By taking the negative log of this probability and averaging over all pairs in the dataset D, the model is penalized when it fails to assign a higher reward to the human-preferred response. Minimizing $\mathcal{L}_{\text{DPO}}(\theta)$ thus encourages the model to consistently rank the chosen response more favorably than the non-chosen one, aligning outputs with human preferences.

We adopted DPO as the sole preference alignment strategy for tuning GPT-4o, as it is the only such algorithm currently offered by OpenAI for their models. This approach allowed us to leverage pairwise preferences collected in our dataset to optimize GPT-4o's responses toward higher human satisfaction. Consequently, GPT-4o achieved stronger alignment with human judgments regarding clarity, correctness, and helpfulness of its outputs.

4.10.8.2 Odds Ratio Preference Optimization

ORPO [45] is a pairwise preference alignment technique that optimizes the ratio of predicted probabilities assigned to the preferred versus the non-preferred response. Formally, its loss function can be expressed as:

$$\mathcal{L}_{\mathrm{ORPO}}(\theta) = -\mathbb{E}_{(A,B)\sim D}\left[\log\left(\frac{p_\theta(A)}{p_\theta(B)}\right)\right], \tag{4.11}$$

where $p_\theta(\cdot)$ indicates the model's predicted probability for each candidate response A or B. Minimizing this term encourages the model to assign higher likelihood to the chosen (human-preferred) response, thus aligning outputs with annotator judgments. Compared to DPO—which uses a sigmoid-based difference in reward scores—ORPO operates directly on the odds ratio, potentially offering more stable updates for smaller models.

We used ORPO for its superior stability over DPO [8] [60] [107], especially when training open-source models. We applied ORPO to preference-align two open-source, instruction-tuned language models, achieving improved alignment to human-labeled preferences.

4.10.8.3 Model training

We aligned three models using our preference dataset. Specifically, we applied DPO to train GPT-4o, while for the two open-source models—LLaMA 3 (8B) and DeepSeek R1 (Distilled Qwen 2.5, 32B)—we employed ORPO. This setup allowed each model to leverage human preference annotations in a manner best suited to its respective infrastructure.

4.10.8.4 Hyperparameter setup

For DPO training on GPT-4o, we followed the default preference alignment workflow provided by OpenAI. We ran the fine-tuning for 2 epochs with a batch size of 32, a learning rate multiplier of 1. Additionally, we used $\beta = 0.1$ in the loss computation, which controls the gradient update weight for the preference signal. For ORPO training on the open-source models (LLaMA 3 and DeepSeek R1), we employed a maximum input and prompt length of 1,024 tokens. The per-device train and per-device eval batch sizes were both set to 16, with 2 gradient accumulation steps to effectively reach an overall batch size of 32. We used a learning rate of 2×10^{-4}, adamw8bit optimization, and weight decay of 0.01. The training proceeded for 100 steps, with an evaluation step every 10 steps. This configuration was managed using a HuggingFace-compatible ORPOTrainer module, ensuring consistent training parameters across our preference alignment experiments.

4.10.8.5 Why not use Reinforcement Learning with the rewards (scores) in the dataset?

Although Reinforcement Learning from Human Feedback (RLHF) [80] has emerged as a robust framework for aligning large language models with human preferences, we did not adopt it in this work due to its considerable implementation complexity and high computational demands. RLHF relies on a multi-stage pipeline involving a reward model, a policy model, and optimization algorithms like Proximal Policy Optimization (PPO). These components collectively introduce substantial memory and scaling challenges.

As highlighted in prior work [43]; [48], the need to manage multiple models during training, including a reward model, significantly increases the overhead associated with memory consumption and training complexity. RLHF also tends to produce diminishing returns in some metrics while potentially degrading performance in others [44], further complicating its deployment in resource-constrained research environments.

In contrast, we opted to use DPO and ORPO—two recently introduced methods that offer more straightforward, more scalable alternatives to RLHF. These techniques enable the direct optimization of human preferences without requiring a separate reward model. They are significantly more straightforward to implement, computationally efficient, and better suited for fine-tuning instruction-tuned models on limited hardware [124]. In our experiments, DPO was used to align GPT-4o. ORPO was used to fine-tune the smaller open-source models, achieving substantial performance improvements without requiring multi-model training infrastructure.

That said, our human-annotated preference dataset—which includes detailed quality scores for correctness, clarity, helpfulness, and verbosity—is fully compatible with RLHF-style optimization and can be utilized in future research where larger GPU compute resources are available. With sufficient hardware, the preference scores can be leveraged to train a reward model and perform full RLHF or GRPO (Group Relative Preference Optimization) to enhance model alignment further. Therefore, while we did not adopt RLHF due to current resource limitations, our dataset lays the groundwork for such experimentation in future iterations of this research.

4.11 Architecture Diagram

The architecture of the proposed self-explaining model for emotion classification in tweets integrates multiple components that work together to enable dynamic few-shot prompting, explanation generation, and preference alignment, as depicted in the diagram 4.1

1. **Embedding-based Retrieval:** The process begins with the SemEval train+dev dataset, where each tweet is encoded into a dense vector using OpenAI's embedding engine. These vectorized tweets are stored in a vector database to facilitate semantic retrieval.
2. **Dynamic Few-shot Prompt Construction:** For each SemEval and DAIR AI test set tweet, the system queries the vector database to retrieve the top 3 most semantically similar tweets and their emotion labels using cosine-similarity. These retrieved examples are used to dynamically construct few-shot prompts tailored to each input tweet dynamically, enhancing contextual relevance.
3. **Initial Response Generation (GPT-4o):** The constructed prompts are sent to GPT-4o, which outputs three elements simultaneously: (i) a natural language explanation, (ii) the most prominent emotion (multi-class), and (iii) the set of emotions present (multi-label). This marks the self-explaining model's core, where prediction and explanation are generated together.
4. **Preference Dataset Construction:** The generated outputs are collected as response pairs and subjected to human annotation. Annotators evaluate each response along four key dimensions—correctness, clarity, helpfulness, and verbosity. These labeled pairs form the preference dataset, crucial for aligning model outputs with human expectations.
5. **Model Fine-tuning via Preference Optimization:** The preference dataset is then used to fine-tune three models: GPT-4o using DPO, and DeepSeek-Distilled Qwen2.5 32B and Meta LLaMA 3 8B using ORPO. This step improves both the classification accuracy and the quality of explanations.
6. **Final Inference and Evaluation:** The aligned models are then used for emotion classification and explanation generation on test data using dynamic few-shot prompts. The outputs are evaluated using classification metrics (accuracy, F1, exact match) and explanation metrics such as sufficiency and human-annotated dimensions, completing

the self-explanation pipeline.

This architecture integrates dynamic few-shot prompting, structured output generation, human-in-the-loop preference annotation, and preference-aligned fine-tuning, providing a unified and interpretable framework for emotion classification in tweets.

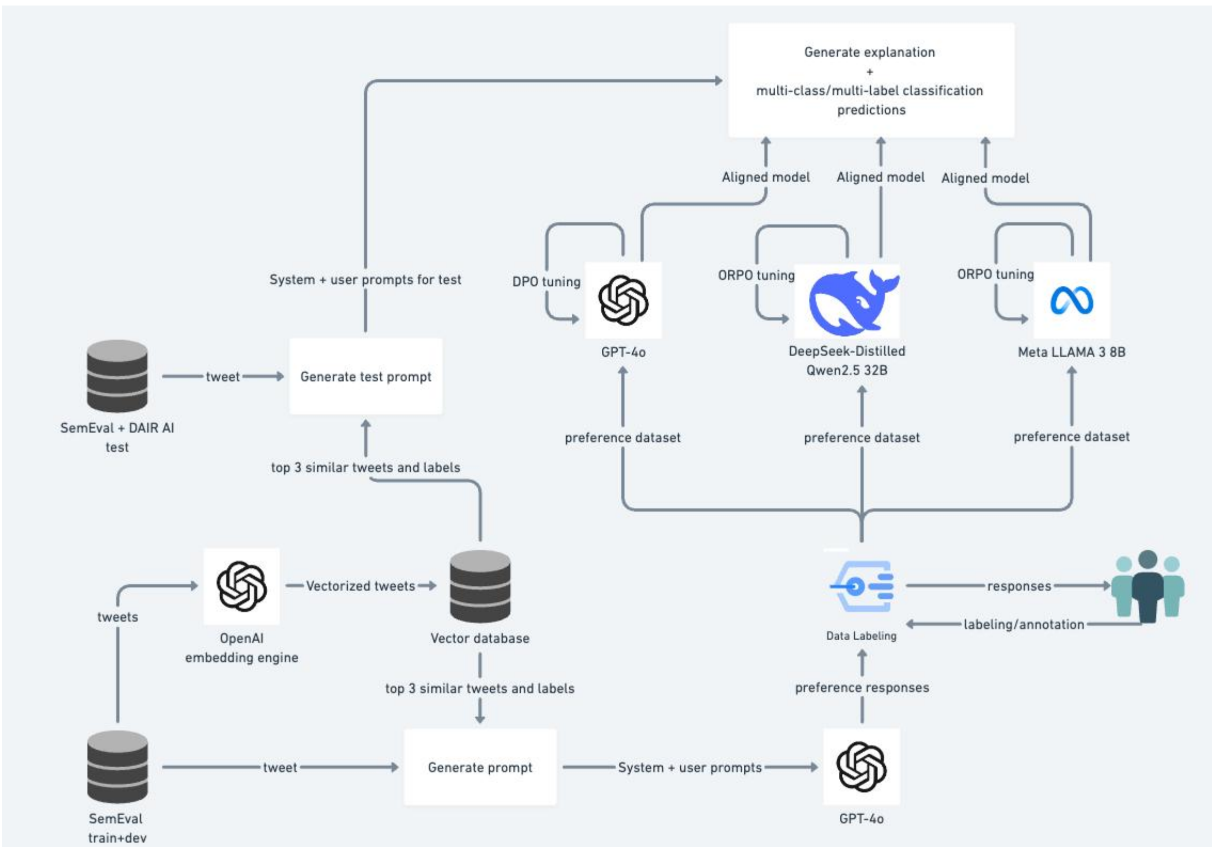

Figure 4.1: System architecture diagram for self-explaining model

4.12 Summary

In this chapter, we detailed the methodology used to address the emotion classification and explanation generation tasks in this thesis. We began by formally defining the multi-label and multi-class classification tasks and the corresponding evaluation metrics, including accuracy, Micro-F1, Macro-F1, and Exact Match for classification, and sufficiency and human scores for explanations. We then described the range of models and techniques employed, including transformer-based classifiers, instruction-tuned GPT2 models, zero-shot and few-shot prompting strategies, and dynamic few-shot prompting using embedding-based retrieval.

The chapter further introduced various explainability techniques explored in this work, such as post-hoc methods (LIME and SHAP), generative AI explanations from large language models, and the development of a novel self-explaining model that produces structured outputs including explanations and emotion predictions. We concluded by presenting the architecture of our overall system and the components involved in structured generation, supervised fine-tuning, and preference alignment.

In the next chapter, we will present the experimental results, comparing model performance across classification tasks, and evaluating the quality and interpretability of generated explanations through both automatic and human-centered metrics.

Chapter 5

Evaluation

This chapter presents a comprehensive evaluation of the various methods explored in this thesis for emotion classification and explanation generation. We begin by analyzing the classification performance of baseline transformer models, followed by results from instruction-tuned models, zero-shot methods, and few-shot learning approaches, including both vanilla and dynamic variants. We then evaluate the performance of different explainability techniques using the sufficiency metric to quantify explanation quality. Finally, we present and compare the results of our proposed self-explaining models under three configurations: pre-trained generative models, supervised fine-tuning, and preference-aligned models. Each section includes detailed performance metrics for both multi-label (SemEval 2018 E-c) and multi-class (DAIR AI) emotion classification tasks, accompanied by discussion subsections that provide analysis and insights into the observed trends. This evaluation establishes a robust foundation for understanding the trade-offs between interpretability, accuracy, and scalability across different modeling strategies.

5.1 Classification results

5.1.1 Transformer-based models

The first section presents the classification results obtained from transformer-based models applied to multi-label (SemEval 2018 E-c) and multi-class (DAIR AI) emotion classification tasks. The multi-label results are drawn from our prior research conducted before the start of this thesis. However, since the same author carried out that work and utilizes the same dataset explored in this thesis, we include those results here for completeness. Moreover, they serve as a valuable baseline against which the improvements introduced in this thesis can be measured and evaluated through instruction tuning, few-shot learning, and self-explaining models.

5.1.2 Instruction tuning prompts

The second section presents the results of our instruction-tuned models based on the GPT-2 architecture. These models are fine-tuned using a next-word generation objective, with emotion classification framed as an instruction-following task. The model $GPT2 - IT_A$ refers to the GPT-2 model that has been Instruction Tuned (IT) using Prompt Setting A, one of the structured prompt formats 4.4.1 defined and evaluated in this thesis. The results reported here reflect how well these models perform on both multi-label and multi-class emotion classification tasks under instruction tuning and serve as a strong benchmark for comparison with more advanced generative and self-explaining models introduced later.

5.1.3 Zero-shot

The third section reports the results for zero-shot emotion classification using GPT-4o. In this setting, the model is prompted with an instruction and a single tweet, without being provided any in-context examples. Despite the absence of fine-tuning or task-specific training, GPT-4o demonstrates strong zero-shot capabilities on both multi-label and multi-class classification tasks. These results highlight the effectiveness of large, instruction-aligned language models in performing complex classification tasks with minimal setup.

5.1.4 Few-shot, dynamic few-shot and self-explaining models

The fourth section presents the results for few-shot and dynamic few-shot emotion classification. In both approaches, we use a structured generation format with a unified prompt template, as detailed in table 4.2, excluding the instruction for generating explanations. The vanilla few-shot setting involves fixed, manually selected examples, while the dynamic few-shot method retrieves the most relevant examples based on embedding similarity. These experiments evaluate how exposure to a few labeled instances influences model performance in multi-label and multi-class classification tasks. Alongside these, we report classification performance from the self-explaining models in their various configurations—pretrained, supervised, fine-tuned, and preference-aligned—demonstrating how explanation-guided outputs can also enhance classification accuracy.

The table 5.1 shows the results obtained on the SemEval multi-label dataset, and 5.2 shows the results obtained on the Dair AI multi-class dataset..

Models	**Accuracy %**	**Micro F1**	**Macro F1**	**EM %**
Self-explaining preference-aligned models				
Self-Explaining GPT-4o (DPO)	**68.61**	**78.38**	68.44	**47.47**
Self-Explaining DeepSeek 32B (ORPO)	65.66	76.14	65.65	41.98
Self-Explaining LLAMA 8B (ORPO)	64.12	74.88	63.19	39.49
Self-explaining pre-trained models				
Self-Explaining GPT-4o	67.94	78.26	**68.97**	46.79
Self-Explaining DeepSeek 32B	64.01	74.80	62.95	39.64
Self-Explaining LLAMA 8B	63.18	74.35	63.18	38.20
Zero-shot and Few-shot models				
Dynamic Few-shot GPT-4o	67.61	77.66	67.92	45.78
Few-shot GPT-4o	63.92	74.74	63.04	39.25
Zero-shot GPT-4o	66.73	77.12	66.52	43.97
Instruction-tuned models				
$GPT2-IT_A$	67.56	77.63	67.67	45.81
$GPT2-IT_B$	64.69	75.32	63.59	40.75
$GPT2-IT_C$	57.57	69.85	53.93	28.57
Transformer-based models				
RoBERTa MA	62.4	74.2	60.3	
Chochlakis et. al [18]	61.1	72.3	56.1	
Baziotis et. al [11]	58.8	70.1	52.8	

Table 5.1: Emotion classification results on different models for SemEval-2018 dataset. The best values are in bold.

Models	**Accuracy**	**Precision**	**Recall**	**F1**
Self-explaining preference-aligned models				
Self-Explaining GPT-4o	93.20	89.67	89.16	89.29
Self-Explaining DeepSeek 32B	90.6	85.37	86.59	85.76
Self-Explaining LLAMA 8B	89.10	82.90	85.03	83.68
Self-explaining pre-trained models				
Self-Explaining GPT-4o	93.10	89.49	89.11	89.19
Self-Explaining DeepSeek 32B	90.75	85.55	86.67	85.88
Self-Explaining LLAMA 8B	88.22	81.61	84.00	82.48
Zero-shot and Few-shot models				
Dynamic Few-shot GPT-4o	93.0	89.43	88.98	89.08
Few-shot GPT-4o	92.20	87.93	88.10	87.84
Zero-shot GPT-4o	92.40	88.39	88.29	88.18
Instruction-tuned models				
$GPT2 - IT_A$	93.3	89.74	89.34	89.43
$GPT2 - IT_B$	90.1	84.43	86.21	85.04
$GPT2 - IT_C$	88.8	82.51	84.76	83.32
Transformer-based models				
RoBERTa	**93.4**	89.97	89.39	**89.58**
BERT	93.3	89.4	87.6	88.4
BERTweet	93.3	89.56	88	88.63
Current State-of-the-art models				
sagemaker-roberta-base-emotion [1]	93.1	88.3	**90.9**	89.5
roberta-base-emotion [2]	93.1	**91.7**	87.4	88.2

Table 5.2: Emotion classification results on different models for Dair AI emotion dataset. The best values are in bold.

5.1.5 Multi-label results

The results in Table 5.1 highlight the performance of various modeling approaches on the SemEval-2018 multi-label emotion classification task, evaluated using Accuracy, Micro F1, Macro F1, and Exact Match (EM). Among all models, the Self-Explaining GPT-4o (DPO-aligned) model achieves the best overall performance, setting a new state-of-the-art on this dataset with an accuracy of 68.61%, Micro F1 of 78.38%, and Exact Match score of 47.47%. This demonstrates that aligning a powerful generative model like GPT-4o using a preference dataset significantly improves both prediction quality and explanation fidelity.

The preference-aligned open-source models—DeepSeek 32B (ORPO) and LLaMA 8B (ORPO)—also show notable performance gains compared to their pre-trained versions. For instance, DeepSeek improves from 64.01% to 65.66% accuracy and LLaMA from 63.18% to 64.12%, with corresponding improvements in Micro F1 and EM. Although these models do not outperform GPT-4o, the performance lift through preference alignment illustrates the effectiveness of our dataset in transferring explanatory reasoning capabilities to smaller, open-access models.

The pre-trained self-explaining models perform reasonably well without fine-tuning or alignment—especially GPT-4o, which achieves 67.94% accuracy and the highest Macro F1 score of 68.97% among all models. This suggests that GPT-4o demonstrates a good balance across emotion classes, including the underrepresented ones, even without alignment. However, compared to the DPO-aligned version, it lags in EM and consistency, emphasizing the role of alignment in enhancing structured output quality.

The few-shot and dynamic few-shot models reveal interesting trends. The Dynamic Few-shot GPT-4o model achieves 67.61% accuracy and 45.78% EM, closely matching the performance of the pre-trained GPT-4o. This confirms the effectiveness of retrieval-based contextualization using similar examples from the training set. The standard few-shot GPT-4o model, however, performs slightly worse, suggesting that static examples may not generalize as effectively as dynamically retrieved ones tailored to each input.

Among the instruction-tuned models, $GPT2 - IT_A$ performs the best, with an accuracy of 67.56% and Micro F1 of 77.63%, on par with the dynamic few-shot model and only slightly behind pre-trained GPT-4o. The drop in performance in $GPT2 - IT_B$ and $GPT2 - IT_C$ highlights the sensitivity of instruction-tuned models to prompt design. This emphasizes the need for well-engineered prompts in instruction-based fine-tuning and also reveals the limits of smaller-scale models like GPT-2 when compared to larger architectures.

The transformer-based baseline models perform considerably worse across all metrics. RoBERTa MA achieves the highest accuracy among them (62.4%), while previously reported models like Chochlakis et al. and Baziotis et al. score even lower. Their significantly reduced Macro F1 and Exact Match scores further indicate that these models struggle with label diversity and fail to capture the nuance of multiple overlapping emotional states in tweets.

Overall, these results demonstrate the effectiveness of self-explaining models, particularly when preference-aligned, in capturing multi-label emotional signals with improved interpretability and structural precision. Preference alignment boosts classification accuracy and F1 scores and significantly enhances Exact Match performance—a strict metric—suggesting the aligned models make fewer partial or structurally inconsistent predictions. These findings highlight the promise of our alignment strategy in creating accurate and reliable models in real-world emotion classification scenarios.

5.1.6 Multi-class results

The results in Table 5.2 summarize the performance of various models on the DAIR AI emotion classification dataset, a multi-class classification task involving six emotion categories. A key observation from these results is that, unlike the multi-label setting, traditional transformer-based models such as RoBERTa, BERT, and BERTweet outperform most of the large language model (LLM)-based approaches. Specifically, RoBERTa achieves the highest F1 score of 89.58% and an accuracy of 93.4%, closely followed by BERT and BERTweet, all of which perform on par with or slightly above the current published state-of-the-art models.

This outcome highlights an important insight: multi-class classification with six well-defined emotion labels is a relatively simpler task, where compact transformer models with fewer parameters are able to generalize effectively. In this context, the additional reasoning capabilities and generative flexibility of LLMs do not translate into a performance advantage. While LLMs such as $GPT2 - IT_A$ and GPT-4o (zero-shot, few-shot, and dynamic) deliver strong results—within 1–1.5 percentage points of the top-performing models—they do not surpass the RoBERTa baseline.

Despite this, the performance of the self-explaining models—particularly GPT-4o (DPO-aligned), which achieves 93.20% accuracy and an F1 score of 89.29%—remains impressive, especially considering that these models were not tuned on the DAIR AI dataset. Instead, the preference alignment was conducted using a different dataset (SemEval 2018 E-c), and yet the models were able to generalize reasonably well to this new task and domain. The open-source aligned models (DeepSeek 32B and LLaMA 8B) also achieved competitive scores, demonstrating the robustness of our preference dataset and alignment strategy, even in a cross-dataset generalization setting.

Additionally, the instruction-tuned $GPT2 - IT_A$ model performs strongly across all metrics, nearly matching RoBERTa with an F1 score of 89.43%, showing that carefully designed prompts and supervised instruction tuning can be highly effective for well-scoped classification problems. Similarly, zero-shot and few-shot variants of GPT-4o show competitive performance in the range of 88–89% F1, confirming that LLMs can generalize well with minimal data when the task is well-constrained.

In summary, although the transformer-based models achieve slightly better performance in this simpler classification task, the LLM-based and self-explaining models demonstrate acceptable and consistent performance, particularly considering they were not explicitly fine-tuned on the DAIR dataset. These findings reinforce that LLMs are highly capable, and with proper alignment, can approach or even match task-specific models in more controlled tasks like multi-class classification.

5.2 Explainable AI Evaluation

5.2.1 Evaluating LIME, SHAP and Generative AI Explanations using the automatic measure

We use a quantitative evaluation metric, the sufficiency metric, to provide an automated assessment of explanation quality. The sufficiency metric leverages a BERT model trained solely on the explanations rather than the entire input tweet. Subsequently, this trained BERT model is utilized to perform inference on the test set, with the accuracy of the BERT model serving as the sufficiency metric. A higher accuracy of the BERT model indicates better explanations, thereby setting a standard for the trustworthiness of explanations.

The BERT score was used for evaluating explanations for other classification tasks [90] (not for emotion classification). It calculates the ability of a simple BERT classifier to produce the same labels (when the input is the explanation text, as mentioned above) as the classifier that we are explaining (on the initial text). It is not a perfect measure, but it is automatic and easy to compute, while evaluating the explanations with human judges is time-consuming and not feasible for large test sets and multiple explanation methods.

5.2.2 Self-explaining model

This section evaluates the explanation generation capabilities of our proposed self-explaining models. We assess three different configurations: (i) structured generation using pre-trained models, (ii) supervised fine-tuning of smaller open-source models using prompt-response pairs generated by GPT-4o, and (iii) preference-aligned models, fine-tuned using our annotated preference dataset with DPO (for GPT-4o) and ORPO (for LLaMA and DeepSeek). We compare the generated explanations' quality across these configurations against the generative explanation baselines described in Section 5.2.1. Evaluation uses two complementary approaches: an automatic evaluation metric, specifically the BERT score, and a human review, where annotators rate explanations on four key dimensions—correctness, clarity, helpfulness, and verbosity.

5.2.2.1 Structured Generative AI pre-trained

Using a structured output format in this setting, we evaluate the explanation generation capabilities of three pre-trained models—GPT-4o, LLaMA 3 (8B), and DeepSeek-R1 (Distilled Qwen 2.5, 32B). During inference, each model is prompted to produce explanations alongside the most prominent emotion and the list of multiple emotions present, adhering to a predefined JSON schema outlined in Table 4.4.

Among the three, GPT-4o demonstrated the most robust structured generation, with fewer than 1% of responses failing to conform to the expected format. These malformed outputs typically lacked strict adherence to the schema, requiring minimal manual parsing to extract the relevant fields.

We employed the Outlines[3] library for the open-source models, a HuggingFace-compatible tool designed to guide language models toward structured text generation. However, despite this constraint, the outputs from LLaMA and DeepSeek were considerably less consistent than those of GPT-4o. Across multiple inference runs, we observed that approximately 30–35% of the responses from these models were malformed—either deviating from the schema or producing incomplete structures. These responses were post-processed and corrected using a custom parsing script to extract the expected fields.

Figure 5.4 results of this structured generation evaluation. While open-source models show promise, the high rate of malformed outputs highlights a key limitation in their current robustness for structured generation. Reducing such inconsistencies remains essential to making these models more reliable for real-world deployment in self-explaining systems.

5.2.2.2 Supervised fine tuning

Given the high-quality explanations, accurate multi-label and multi-class predictions generated by GPT-4o, and its strong consistency in adhering to the desired output structure, we aimed to transfer this capability to smaller, open-source models through supervised fine-tuning (SFT). The goal was to improve both the output quality and structural consis-

[3] https://github.com/dottxt-ai/outlines

tency of models like LLaMA 3 (8B) and DeepSeek-R1 (Distilled Qwen 2.5, 32B) by training them on prompt-response pairs collected from GPT-4o.

To implement SFT, we used the SFTTrainer module from the TRL[4] (Transformers Reinforcement Learning) library, fine-tuning the models on GPT-4o-generated outputs. While supervised fine-tuning is a promising approach for adapting models to specific tasks and improving their performance, our experiments revealed several unexpected issues. Despite successful training, the outputs from the fine-tuned models were significantly more malformed than those of their pre-trained counterparts. On the test set, nearly all responses failed to conform to the expected structure, with the models frequently ignoring the defined output format.

More critically, both LLaMA and DeepSeek models exhibited unusual behavior. They continued generating repeated content indefinitely, ignoring the End-of-Sequence (EOS) token and producing outputs until the maximum token limit was reached. Despite multiple debugging attempts, we could not resolve the issue, and no apparent flaw was found in the training or inference pipeline.

We suspect this behavior stems from two main limitations: (i) the relatively small size of the supervised dataset and (ii) the restricted number of trainable parameters due to LoRA-based fine-tuning, which we adopted to remain within available GPU constraints. These factors may have hindered the models' ability to learn the response structure adequately and stopping behavior, making SFT less effective in our setup than anticipated.

[4] https://huggingface.co/docs/trl/en/index

5.2.2.3 Preference-aligned models

To address the limitations observed with supervised fine-tuning, we explored preference alignment using ORPO for open-source models and DPO for GPT-4o. These methods leverage human preference data to directly optimize model behavior without requiring complex reward modeling, making them more efficient and stable alternatives to traditional RLHF. Using our manually annotated preference dataset, which includes pairs of GPT-4o-generated explanations rated for correctness, clarity, helpfulness, and verbosity, we aligned LLaMA and DeepSeek models with ORPO and fine-tuned GPT-4o using DPO via OpenAI's preference training interface. Unlike SFT, preference-aligned models showed substantial improvements in explanation quality and structural consistency. Notably, the open-source models no longer exhibited repetitive or malformed outputs, and their explanations were more aligned with human-preferred reasoning. This demonstrates that preference optimization can significantly enhance the reliability and interpretability of self-explaining models, even under resource constraints. Table 5.3 presents the Sufficiency metric results, which gauge whether a model's explanation text alone is predictive of its final classification.

5.2.3 Automatic evaluation discussion

The results in Table 5.3 offer a comprehensive comparison between our proposed self-explaining models and existing explainable AI (XAI) methods, as measured by the sufficiency metric. Notably, our preference-aligned generative models consistently outperform current approaches, with GPT-4o-DPO achieving the highest sufficiency score of 63.66

In contrast, established statistical explanation methods such as LIME and SHAP, both coupled with RoBERTa classifiers, yield substantially lower sufficiency scores (53.22% and

54.16%, respectively). While these techniques have traditionally been valued for their ability to provide local, model-agnostic insights, their explanations are limited in scope and depth. Specifically, LIME and SHAP explanations often depend heavily on the quality and representativeness of the data, and they primarily focus on feature attribution within a narrow context [104]. This can result in explanations that, while technically accurate, may lack narrative richness and fail to capture the broader reasoning process underlying a model's decision, particularly in complex, subjective tasks such as emotion classification.

By contrast, our generative models, especially when preference-aligned, can leverage large-scale pre-training and explicit human preference signals to produce explanations that are both human-readable and contextually grounded. These models not only provide more comprehensive rationales behind each classification decision but also demonstrate greater capacity to articulate underlying emotional concepts in everyday language. The higher sufficiency scores for the proposed models underscore their ability to generate explanations that stand on their own, making them more accessible and useful to end-users.

Furthermore, the strong performance of smaller models like DeepSeek R1 (61.12%) and LLAMA 3 8B (60.98%) after ORPO fine-tuning demonstrates that preference alignment is effective even in compute-constrained settings. This scalability is significant for real-world applications where model size and inference efficiency are practical concerns.

In summary, these results highlight the limitations of traditional statistical XAI methods such as LIME and SHAP for generating user-facing explanations and demonstrate the clear advantages of generative, preference-aligned approaches in both explanation quality and sufficiency.

Explainable AI models	Sufficiency
Our proposed models	
GPT-4o - DPO	**63.66**
DeepSeek R1 (Distilled Qwen 32B) - ORPO	61.12
LLAMA 3 8B - ORPO	60.98
Current State-of-the-art models [33]	
GPT-4o	59.66
SHAP-RoBERTa	54.16
LIME-RoBERTa	53.22

Table 5.3: Sufficiency metric for evaluating explanations for the models

5.2.4 Human evaluation

In addition to automatic evaluation using the sufficiency metric (e.g., BERTScore), we conducted a comprehensive human assessment of the generated explanations to better understand their interpretability and usefulness from a human-centered perspective.

For the human evaluation, we randomly sampled 300 tweets from the test set. For each tweet, we generated two responses: one using the pre-trained model and one using the corresponding preference-aligned model. These paired explanations were then assessed by a team of three human annotators, consisting of two experienced data engineers (the same used for preference-dataset annotations) and the author of this thesis. Each annotator evaluated 100 tweets per model, ensuring that at least one human evaluator independently assessed every model (pre-trained and aligned) for a diverse and generalized set of generations. Each explanation was scored across four key qualitative dimensions as shown in 4.2.3.2. All dimensions were rated on a 5-point Likert scale (1–5), allowing for fine-grained comparisons between model outputs. We use star graphs to show human evaluation scores for comparison between pre-trained and preference-aligned models.

The three radar plots 5.1, 5.2, and 5.3 provide a comparative visualization of the human

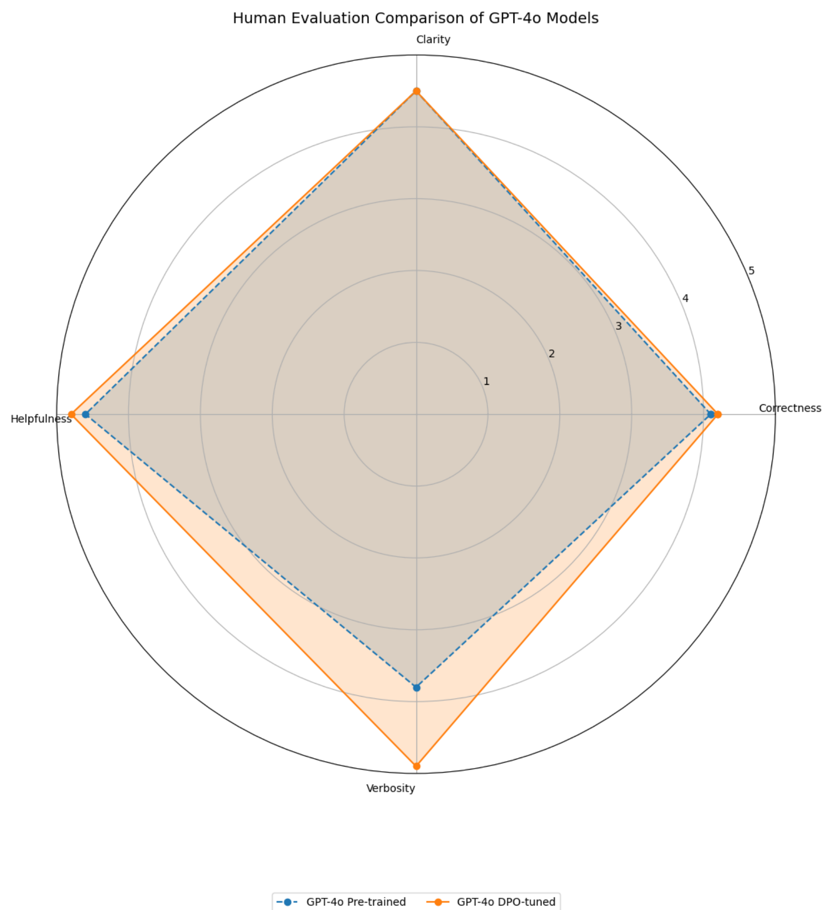

Figure 5.1: Human Evaluation Comparison of GPT Models

evaluation results across four key qualitative dimensions — Correctness, Clarity, Helpfulness, and Verbosity — for each of the three models (GPT-4o, LLaMA 8B, and DeepSeek 32B), both in their pre-trained and preference-aligned versions.

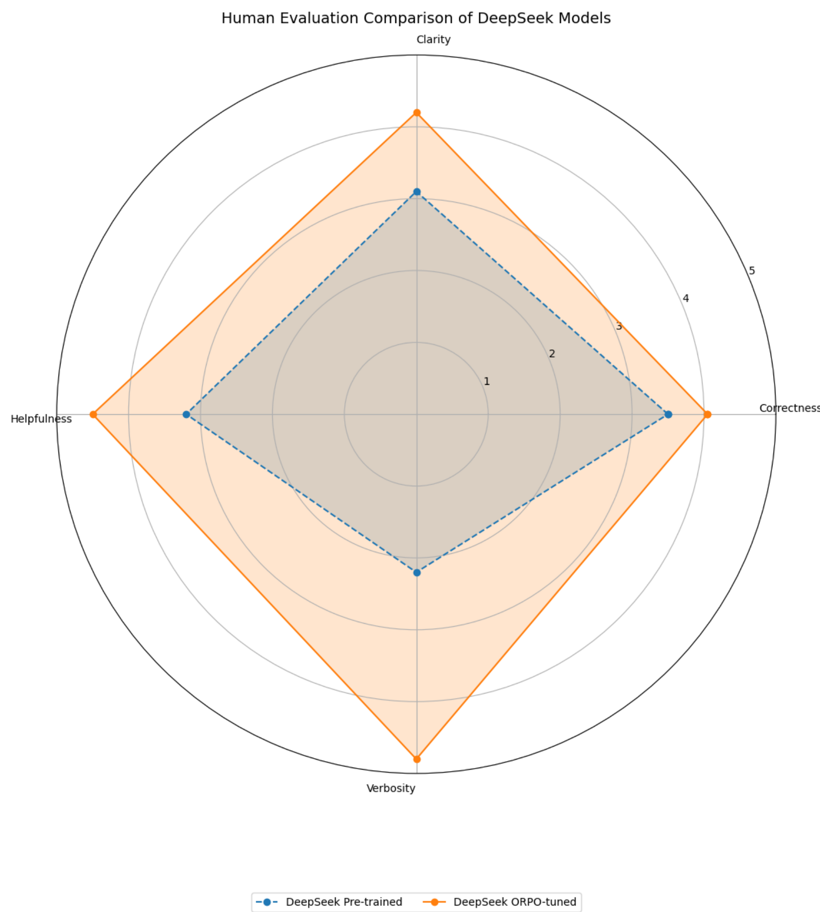

Figure 5.2: Human Evaluation Comparison of DeepSeek Models

5.2.4.1 GPT-4o Results

In the GPT-4o comparison, we observe that the preference-aligned model (DPO-tuned) consistently outperforms or matches the pre-trained version across all four dimensions. Notably, the largest improvement is seen in Verbosity, where the preference-aligned model reaches close to a perfect score. This indicates that DPO alignment not only maintains GPT-4o's high correctness and clarity but significantly enhances its ability to produce

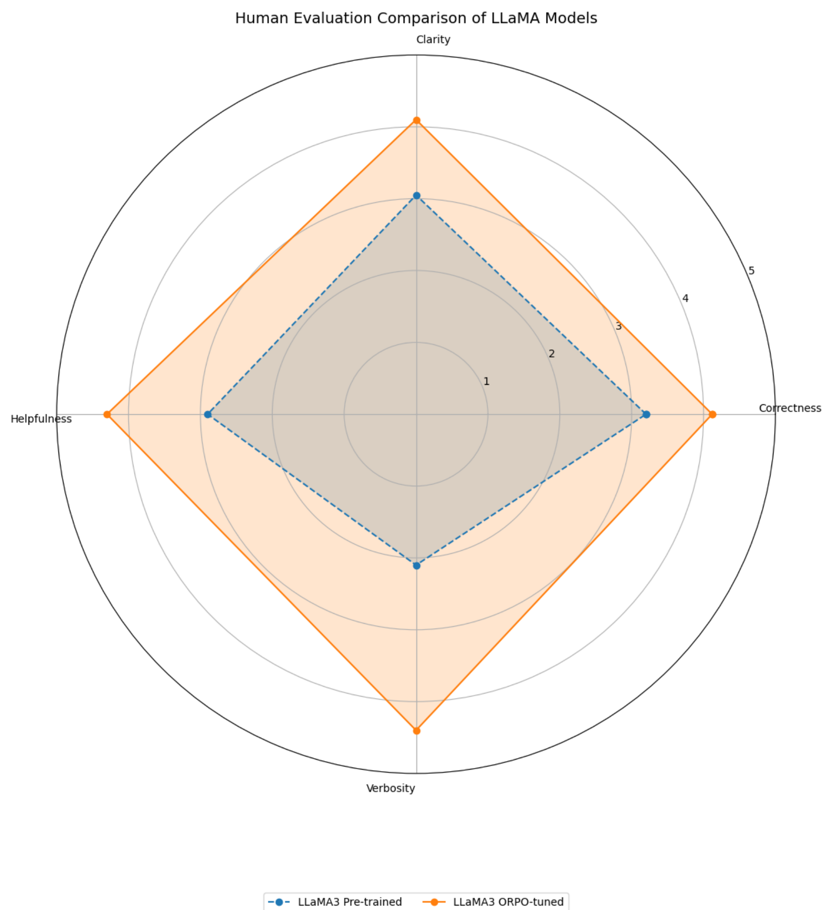

Figure 5.3: Human Evaluation Comparison of LLaMA Models

explanations that are detailed yet concise — an important quality for real-world interpretability.

5.2.4.2 LLaMA 8B results

The LLaMA radar plot shows even more dramatic improvements from preference alignment. All four dimensions — especially Clarity, Helpfulness, and Verbosity — show a

Model	Clarity	Correctness	Helpfulness	Verbosity
LLaMA 3 8B (pre-trained)	3.2	2.9	2.8	3.5
LLaMA 3 8B (ORPO-tuned)	4.2	4.0	4.1	3.8
DeepSeek R1 (pre-trained)	3.1	3.0	3.0	3.2
DeepSeek R1 (ORPO-tuned)	4.1	4.3	4.2	3.6
GPT-4o (pre-trained)	3.8	3.5	3.6	3.3
GPT-4o (DPO-tuned)	4.4	4.3	4.1	3.7

Table 5.4: Average human evaluation scores (on a 5-point Likert scale) for explanation quality across four dimensions. Evaluations were performed on 300 randomly sampled tweets from the test set.

substantial boost after ORPO tuning. This suggests that the smaller LLaMA 8B model benefits greatly from preference alignment, learning to generate explanations that are not only more faithful but also more intelligible and useful. The change in Verbosity is particularly striking, indicating the model's improved ability to balance informative content with conciseness.

5.2.4.3 DeepSeek 32B results

DeepSeek exhibits a similar trend as LLaMA, with the ORPO-aligned model outperforming the pre-trained model across all metrics. Gains in Helpfulness and Verbosity are again notable, which aligns with the goal of preference optimization — to train models that better match human expectations in format, structure, and interpretive utility.

5.2.4.4 Human evaluations discussion

The human evaluation results in Table 5.4 provide a detailed comparison of explanation quality across pre-trained and preference-aligned versions of three large language models—LLaMA 3 8B, DeepSeek R1, and GPT-4o—measured along four qualitative dimensions: clarity, correctness, helpfulness, and verbosity. These dimensions were selected to capture both the interpretability and usefulness of the generated explanations in the context of emotion classification.

Across all models, we observe a clear and consistent improvement in performance after preference alignment. The most significant gains were seen in the open-source models. For instance, LLaMA 3 8B's clarity increased from 3.2 to 4.2, and correctness rose from 2.9 to 4.0, indicating that the model became substantially more capable of expressing coherent and accurate reasoning after alignment. Similarly, DeepSeek R1 improved from 3.1 to 4.1 in clarity and from 3.0 to 4.3 in correctness. The improvement in helpfulness scores (from 2.8 to 4.1 for LLaMA and 3.0 to 4.2 for DeepSeek) reflects that explanations became more informative, offering clearer links between textual cues and predicted emotions.

Interestingly, even GPT-4o—already a strong baseline—benefited from preference alignment. Its clarity improved from 3.8 to 4.4, and correctness from 3.5 to 4.3. These improvements, though more modest than those of smaller models, suggest that preference alignment helped the model better internalize what constitutes a "good" explanation. Because GPT-4o was trained using a dataset that included both high- and low-quality examples (with one response preferred over the other), it learned to distinguish between more and less effective explanations. This exposure reinforced patterns associated with desired outputs and discouraged suboptimal generations, thereby enhancing its consistency and alignment with human expectations.

Verbosity scores remained relatively stable across models, indicating that the alignment process did not yield overly verbose or underinformative outputs. The aligned models maintained or slightly improved their balance of detail and conciseness, which is critical for practical usability.

Overall, these results strongly support the value of preference alignment. Not only does it substantially enhance explanation quality for open-source models that lack structured output utilities, but it also sharpens the output of even high-performing models like GPT-4o by guiding them toward more human-aligned behavior through exposure to contrastive examples. This demonstrates the broad utility and effectiveness of alignment methods for building trustworthy, explainable NLP systems.

5.2.5 Adherence to structure

One of the most practical advantages of preference alignment revealed in our experiments is its remarkable ability to enforce strict output formatting in open-source models 5.4. Unlike proprietary models like GPT-4o, which benefit from OpenAI's structured output utilities, open-source models typically lack such built-in schema enforcement and often produce inconsistent or malformed outputs. However, after applying preference alignment using our custom preference dataset, models like DeepSeek and LLaMA showed a dramatic improvement in structural adherence — jumping from 66% and 68% respectively, to a near-perfect 99% in both cases. This makes these models significantly more robust and production-ready, as the aligned outputs can be parsed directly without requiring any manual intervention or post-processing logic.

To emphasize the significance of this improvement, we also compare it with the $GPT2-IT_A$ model, which, although not a generative model designed for structured outputs, was used in our instruction-tuned classification experiments. We observed that approximately 38% of its outputs required cleaning, correction, or manual parsing to extract usable predictions. The dashed horizontal line in the graph highlights this 62% adherence benchmark from $GPT2 - IT_A$. The fact that preference-aligned open-source models outperform this and even their own pre-trained counterparts by such a wide margin underscores the efficacy of preference alignment in improving the quality of responses and their structural consistency — a crucial requirement for real-world deployment and interpretability.

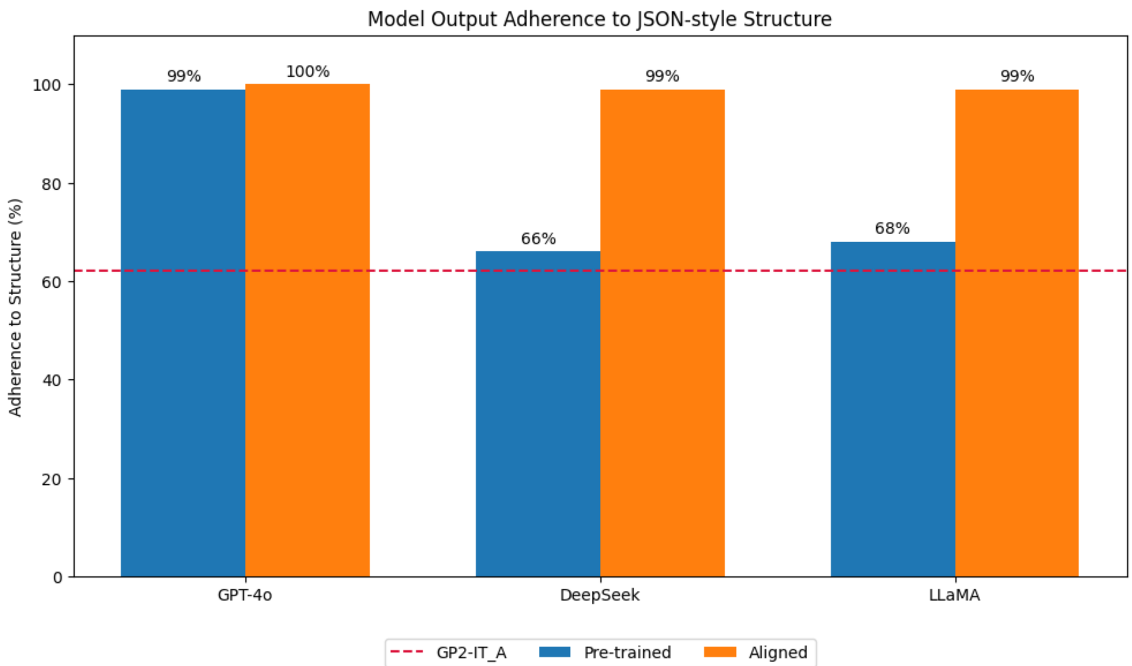

Figure 5.4: Model Output Adherence to JSON-style Structure

5.2.6 Cost and complexity vs. performance

When considering emotion classification tasks, the choice of model often balances between performance gains and computational cost. For multilabel classification, models such as GPT2-IT_A and RoBERTa demonstrate strong performance that is close to that of more complex models. The self-explaining GPT-4o model, while achieving marginally better results, does so at a substantially higher computational cost. This makes GPT2-IT_A a practical choice when the goal is solely multilabel classification, as the performance improvements from GPT-4o are relatively small and may not justify the increased resource requirements.

Similarly, for multiclass classification, RoBERTa offers state-of-the-art accuracy and F1 scores with lower computational overhead. The more complex self-explaining GPT-4o does not provide meaningful performance gains in this setting, reinforcing RoBERTa's suitability as a cost-effective solution for standard multiclass emotion classification.

However, the advantage of models like GPT-4o becomes clear in more advanced use cases. When the task requires not only identifying a prominent emotion but also capturing all other present emotions along with generating a narrated explanation, the enhanced interpretability and detailed outputs justify the higher computational expense. In these scenarios, the richer outputs from GPT-4o outweigh the increased complexity and resource needs.

Moreover, smaller models like GPT2-IT_A and RoBERTa offer practical benefits as they can perform inference without the need for specialized GPU hosting, unlike LLAMA, DeepSeek, and GPT-4o. This makes them more accessible and cost-effective for deployment in resource-constrained environments or where inference latency is a concern.

In summary, for pure classification tasks, smaller models provide a favorable balance of accuracy, efficiency, and cost. For applications demanding explainability and comprehensive emotion representation, the investment in GPT-4o is justified by its superior capabilities.

5.3 Summary

In this chapter, we presented a comprehensive evaluation of the models and methods proposed in this thesis, covering both emotion classification performance and explanation quality. We began by reporting results for traditional transformer-based classifiers and instruction-tuned models, followed by extensive evaluation of large language models (LLMs) in zero-shot, few-shot, and dynamic few-shot settings. The performance of our proposed self-explaining models was also analyzed in detail, including their pre-trained, supervised fine-tuned, and preference-aligned variants.

The evaluation extended beyond classification metrics to include structured assessments of explainability. We compared generative AI explanations from pre-trained LLMs with self-explaining models that are aligned, using both an automatic metric and human evaluations based on correctness, clarity, helpfulness, and verbosity. Our results demonstrated that preference alignment significantly improved explanation quality and structural consistency, particularly in open-source models like LLaMA and DeepSeek. Additionally, we analyzed the models' adherence to structured output formats and showed how alignment helped enforce predictable, machine-parsable outputs.

In the next chapter, we will synthesize the findings from this evaluation, discuss key takeaways, reflect on the limitations of this work, and outline future directions for extending self-explaining emotion classification models.

Chapter 6

Conclusion, Limitations and Future Work

6.1 Conclusion

This thesis presented a comprehensive exploration of emotion classification and explanation using large language models (LLMs), targeting both multi-label and multi-class emotion classification tasks on social media data. We approached the problem not only from a classification accuracy perspective but also through the lens of interpretability, aiming to build a robust self-explaining model capable of providing structured outputs that include both emotion labels and human-readable justifications.

In the classification experiments, we benchmarked a wide range of models—from traditional transformer-based baselines and instruction-tuned GPT2 variants to cutting-edge zero-shot, few-shot, and dynamically retrieved few-shot prompts with GPT-4o. Our results showed that while models like RoBERTa performed strongly on the simpler multi-class DAIR AI dataset, our preference-aligned self-explaining models achieved competitive performance across both multi-label and multi-class settings. Notably, GPT-4o aligned via

DPO attained state-of-the-art accuracy on the SemEval 2018 multi-label task, while the aligned DeepSeek and LLaMA models also demonstrated consistent improvements over their pre-trained counterparts.

We also highlighted the limitations of post-hoc explainability techniques, showing that while they offer localized insights, they fail to generate coherent or user-friendly explanations that explain why the model made a particular prediction. So, we proposed generative AI explanations using LLMs and introduced our novel self-explaining model, capable of jointly predicting emotions and generating contextual explanations. By integrating structured output formatting with multi-task generation, the self-explaining model allows us to enforce output schemas while providing greater interpretability than traditional post-hoc methods.

To evaluate the quality of explanations, we employed both automatic and human-centered metrics. The automatic evaluation leveraged the sufficiency metric from the FRESH pipeline, measuring whether explanations alone could recover the original prediction. Preference-aligned models significantly outperformed both pre-trained LLMs and traditional explainers under this metric, affirming the faithfulness of their explanations.

To complement this, we conducted an extensive human evaluation, where annotators rated 300 explanations from each model on four qualitative dimensions: correctness, clarity, helpfulness, and verbosity. Across the board, preference-aligned models consistently scored higher than their pre-trained counterparts, and in many cases, matched the output quality of GPT-4o. These findings validate the effectiveness of preference alignment as a powerful alternative to reinforcement learning with human feedback (RLHF), especially for resource-constrained settings where full-scale RLHF may be infeasible.

An important practical advantage uncovered through this research was the improvement in structured output adherence. Pre-trained open-source models often produced malformed outputs when asked to generate JSON-style responses. However, after preference alignment, both DeepSeek and LLaMA showed near-perfect conformance to the output structure (99%), making them viable for downstream applications without the need for additional output parsing—something not previously possible in such models. This structural consistency, combined with improved explanation quality and classification accuracy, demonstrates that preference alignment serves as a lightweight yet highly effective way to enhance model reliability and usability.

In conclusion, this thesis advances explainable emotion classification by developing a self-explaining model architecture, a novel preference-aligned dataset, and extensive experimental validation across multiple fronts. Through structured generation, preference alignment, and multi-faceted evaluation, we show that small, efficient, open-source models can be guided to behave comparably to larger proprietary systems. These findings pave the way for accessible, interpretable, and deployable LLMs for emotion understanding in real-world social media applications.

6.2 Summary of Contributions

6.2.1 Publications, datasets, and models

This thesis presents several significant contributions to the field of emotion classification and explainable AI. The publications from this thesis are listed in the section 1.5

Beyond published papers, the thesis introduces the following novel technical contributions:

- A self-explaining emotion classification model architecture, capable of simultaneously generating multi-label and multi-class emotion predictions along with structured, human-readable explanations. These models achieve state-of-the-art results on benchmark datasets such as SemEval 2018 E-c and DAIR AI, offering practical utility and interoperability.
- A preference-alignment dataset, constructed using chain-of-thought prompting and human-annotated scoring. This dataset is valuable for preference-aligned fine-tuning and as a reusable resource for future research in reward-based optimization of self-explaining models.

Together, these contributions address key gaps in interpretability, reliability, and accessibility of emotion classification systems.

6.2.2 Answering the research questions

This thesis effectively addresses all of the core research questions at the study's beginning.

6.2.2.1 Enhancing LLMs for emotion classification

RQ: How does the performance of the instruction-tuned GPT2 model compare to existing state-of-the-art models and the zero-shot and few-shot classification capabilities of a pre-trained LLM for the emotion classification task (on the selected datasets)? This thesis performed an extensive evaluation of transformer-based models, instruction-tuned GPT2 variants, and GPT-4o in both zero-shot and few-shot setups. The results showed that instruction tuning with prompt format A (GPT2–ITA) achieved strong results on the DAIR AI dataset, while GPT-4o outperformed all models on the multi-label SemEval 2018 E-c dataset, demonstrating the benefits of instruction tuning and few-shot prompting for LLM-based emotion classification.

RQ: How can LLMs and alignment be leveraged for data augmentation to address class imbalance and improve the robustness of emotion classification models? Although this thesis initially proposed data augmentation as a possible approach to address class imbalance, our empirical findings indicated that it was not necessary for the final models. Supervised fine-tuning experiments revealed that fine-tuning led to unstable model behavior, where models failed to learn effectively from the training data and often produced malformed outputs. In contrast, preference alignment techniques like ORPO and DPO yielded significantly better performance and structural consistency without requiring additional synthetic training data. Moreover, the dynamic few-shot approach, which selected relevant examples based on tweet embeddings, proved to be highly effective in boosting performance. Therefore, given the strong results from alignment and few-shot strategies, additional data generation or augmentation was deemed unnecessary within the scope of this thesis.

6.2.2.2 Explainable AI

RQ: How do generative AI explanations and traditional statistical methods (e.g., LIME, SHAP) help interpret the decisions of black-box models for emotion classification? The thesis conducted comparative analysis between post-hoc explainers and generative AI-based explanations produced by GPT-4o. These were evaluated using the sufficiency metric (an automatic measure) and a human evaluation. The results showed that generative explanations outperformed post-hoc techniques in producing coherent, interpretable, and task-relevant justifications for model predictions.

RQ: How can a self-explaining model for emotion classification be developed using LLMs to classify emotions and provide meaningful, concept-based explanations? The thesis introduced a novel self-explaining model that outputs both emotion predictions and explanations in a structured JSON format. This model was built using pre-trained LLMs (GPT-4o, LLaMA, DeepSeek) and improved through preference alignment. The resulting models demonstrated strong performance in both classification and explanation quality, satisfying the criteria for a robust, self-explaining system.

RQ: How can alignment techniques improve the quality of AI-generated explanations and ensure they align with human expectations, and how can we quantitatively and qualitatively evaluate them? Alignment techniques (DPO and ORPO) were applied using a custom preference dataset annotated for correctness, clarity, helpfulness, and verbosity. These preference-aligned models were evaluated using both automatic metrics (sufficiency) and human scoring across four explanation quality dimensions. The results confirmed that alignment significantly improved explanation faithfulness, fluency, and consistency— surpassing even pre-trained GPT-4o outputs in structure adherence and clarity.

6.3 Limitations

Despite the contributions of this thesis to emotion classification and explainable AI, some limitations exist that point to potential avenues for future research and refinement.

The alignment and fine-tuning experiments were restricted to relatively smaller open-source models (LLAMA 8B and DeepSeek-Qwen 32B) due to GPU and memory constraints. While these models demonstrated impressive performance after preference alignment, larger models could potentially achieve even greater accuracy and explanation quality, an avenue that remains unexplored in this work. Also, although supervised fine-tuning was attempted using high-quality GPT-4o responses, the models failed to generate consistently structured outputs. Most generations were malformed, despite using well-established training utilities like SFTTrainer from TRL. This suggests that either the dataset size was insufficient or supervised fine-tuning is ill-suited for this task under constrained resources. As a result, we relied more on preference alignment and dynamic few-shot prompting, which showed better performance and consistency.

For dynamic few-shot prompting and preference dataset generation, we only used examples from the SemEval 2018 E-c dataset. This was due to the nature of the DAIR AI dataset, which contains single-label annotations. Since our models are designed to output structured predictions, including multiple emotions, DAIR AI was not suitable for generating the structured preference data. Also, the preference dataset used for alignment consisted of only 1,063 examples. Although these examples were of high quality and manually verified, the limited dataset size may have restricted the extent to which models could fully learn nuanced behavior from preference signals. Additionally, the human evaluation of generated explanations was limited to a sample of 300 tweets, while this allowed for manageable annotation efforts, a larger sample size would provide more statistically robust insights.

Also, the study was confined to emotion classification in tweets. Tweets have unique linguistic characteristics such as abbreviations, emojis, and informal grammar, which may limit the generalizability of the models to other domains like product reviews, news articles, or multimodal settings. Lastly, while this thesis successfully utilized preference alignment through DPO and ORPO, we did not explore more advanced reinforcement learning-based alignment methods such as Reinforcement Learning from Human Feedback (RLHF) or Group Relative Preference Optimization (GRPO). This decision was primarily due to compute resource constraints and the complexity involved in implementing and scaling these methods, especially given the need for separate reward models and careful loss tuning. However, the preference dataset constructed in this work was annotated across four key dimensions—correctness, clarity, helpfulness, and verbosity, making it suitable for reward modeling. These scores provide a supervision signal that could be directly leveraged in RLHF or GRPO-based alignment pipelines to refine explanation quality. As such, while this thesis did not explore those avenues, it establishes the foundation for future research that seeks to combine the interpretability benefits of preference alignment with the precision of reward-optimized training.

6.4 Future Research Directions

This thesis has laid a strong foundation for self-explaning emotion classification using large language models and preference alignment. However, several promising directions remain for future exploration. One key avenue is the integration of reward modeling for preference-aligned fine-tuning using reinforcement learning techniques such as Reinforcement Learning from Human Feedback (RLHF) and Group Relative Preference Optimization (GRPO). Since the generated dataset already includes human-rated reward signals across dimensions such as correctness, clarity, helpfulness, and verbosity, it naturally supports more advanced alignment strategies based on reward learning. These approaches could help further refine explanation quality and overall model robustness.

Another direction involves scaling up model sizes. While this work primarily used instruction-tuned, mid-sized, and distilled models to remain compute-efficient, exploring larger open-source models or frontier foundation models could potentially lead to even more accurate and generalizable self-explaining emotion classifiers. Similarly, the performance gap between models like LLAMA or DeepSeek and GPT-4o could be further narrowed by increasing model capacity or combining alignment with instruction fine-tuning.

An especially intriguing research idea is to experiment with smaller prompts. Given that the same output (explanation + classification) in the dataset is generated from long, structured prompts, future work could train models using shorter or simplified prompts paired with these outputs. If successful, this would demonstrate that models can be taught to generate GPT-4o-level responses with significantly shorter input prompts—potentially reducing inference costs and latency, which is highly valuable for real-time or edge deployment.

Lastly, expanding this work into the multi-modal domain presents an exciting opportunity. Emotion classification is not limited to textual inputs—images, videos, audio, and physiological signals also play a significant role in emotional expression. Integrating modalities like images or voice (e.g., from video or audio tweets) into a unified, self-explaining framework could dramatically enhance emotion recognition capabilities and open new applications in fields such as mental health, customer support, and social media monitoring.

These directions pave the way for building more capable, aligned, and transparent AI systems that not only classify emotions with high accuracy but also explain their decisions in a trustworthy and human-understandable manner.

References

[1] Muhammad Abdul-Mageed and Lyle Ungar. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[2] Heyam H. Al-Baity, Hala J. Alshahrani, Mohamed K. Nour, Ayman Yafoz, Omar Alghushairy, Raed Alsini, and Mahmoud Othman. Computational linguistics based emotion detection and classification model on social networking data. *Applied Sciences*, 12(19), 2022.

[3] A. M. Alkalbani, A. S. Alrawahi, A. Salah, V. Haghighi, Y. Zhang, S. Alkindi, and Q. Z. Sheng. A systematic review of large language models in medical specialties: applications, challenges and future directions. 2024.

[4] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks, 2018.

[5] Iqra Ameer, Noman Ashraf, Grigori Sidorov, and Helena Adorno. Multi-label emotion classification using content-based features in twitter. *Computación y Sistemas*, 24, 09 2020.

[6] Iqra Ameer, Necva Bölücü, Muhammad Hammad Fahim Siddiqui, Burcu Can, Grigori Sidorov, and Alexander Gelbukh. Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*, 213:118534, 2023.

[7] Iqra Ameer, Grigori Sidorov, Helena Gómez-Adorno, and Rao Muhammad Adeel Nawab. Multi-label emotion classification on code-mixed text: Data and methods. *IEEE Access*, 10:8779–8789, 2022.

[8] S. H. Amini, C. R. Vass, M. Shahabi, and A. Noble. Optimization of coal blending operations under uncertainty – robust optimization approach. *International Journal of Coal Preparation and Utilization*, 42:30–50, 2019.

[9] Jean-Yves Antoine, Jeanne Villaneau, and Anaïs Lefeuvre. Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In Shuly Wintner, Sharon Goldwater, and Stefan Riezler, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 550–559, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

[10] Noman Ashraf, lal khan, Sabur Butt, Hsien-Tsung Chang, Grigori Sidorov, and Alexander Gelbukh. Multi-label emotion classification of urdu tweets. *PeerJ Computer Science*, 8:e896, 04 2022.

[11] Christos Baziotis, Athanasiou Nikolaos, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and

Alexandros Potamianos. NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning. In Marianna Apidianaki, Saif M. Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat, editors, *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 245–255, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[12] Andrew Beatty. Emotions in the field: What are we talking about? *The Journal of the Royal Anthropological Institute*, 11(1):17–37, 2005.

[13] Pavel Beránek and Vojtěch Merunka. Analyzing customer sentiments: A comparative evaluation of large language models for enhanced business intelligence. In João Paulo A. Almeida, Claudio Di Ciccio, and Christos Kalloniatis, editors, *Advanced Information Systems Engineering Workshops*, pages 229–240, Cham, 2024. Springer Nature Switzerland.

[14] Francisco Charte, Antonio J. Rivera, María José del Jesús, and Francisco Herrera. Dealing with difficult minority labels in imbalanced mutilabel data sets. *Neurocomputing*, 2019.

[15] J. Chen, R. Zhang, Y. Mao, and J. Xu. Contrastnet: a contrastive learning framework for few-shot text classification. *Proceedings of the Aaai Conference on Artificial Intelligence*, 36:10492–10500, 2022.

[16] Wenshi Chen, Bowen Zhang, and Mingyu Lu. Uncertainty quantification for multilabel text classification. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, 2020.

[17] Y. Chen, B. Xu, Q. Wang, Y. Liu, and Z. Mao. Benchmarking large language models on controllable generation under diversified instructions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:17808–17816, 2024.

[18] Georgios Chochlakis, Gireesh Mahajan, Sabyasachee Baruah, Keith A. Burghardt, Kristina Lerman, and Shrikanth S. Narayanan. Leveraging label correlations in a multi-label setting: a case study in emotion. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2022.

[19] Tulika Chutia and Nomi Baruah. A review on emotion detection by using deep learning techniques. *Artificial Intelligence Review*, 57, 07 2024.

[20] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[21] Antonio R. Damasio, B. J. Everitt, and D. Bishop. The somatic marker hypothesis and the possible functions of the prefrontal cortex [and discussion]. *Philosophical Transactions: Biological Sciences*, 351(1346):1413–1420, 1996.

[22] Yuhao Dan, Jie Zhou, Qin Chen, Qingchun Bai, and Liang He. Enhancing class understanding via prompt-tuning for zero-shot text classification. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4303–4307, 2022.

[23] Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. Framework for evaluating faithfulness of local explanations, 2022.

[24] Luna De Bruyne, Toni G.L.A van der Meer, Orphée De Clercq, and Véronique Hoste. Using state-of-the-art emotion detection models in a crisis communication context. *Computational Communication Research*, 6(1), 2024.

[25] De Langhe, Loic and Maladry, Aaron and Vanroy, Bram and De Bruyne, Luna and Singh, Pranaydeep and Lefever, Els and De Clercq, Orphée. Benchmarking zero-shot text classification for Dutch. *COMPUTATIONAL LINGUISTICS IN THE NETHERLANDS JOURNAL*, 13:63–90, 2024.

[26] Julien A. Deonna and Fabrice Teroni. *The Emotions: A Philosophical Introduction*. Routledge, New York, 2012.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[28] Arwa Diwali, Kawther Saeedi, Kia Dashtipour, Mandar Gogate, Erik Cambria, and Amir Hussain. Sentiment analysis meets explainable artificial intelligence: A survey on explainable sentiment analysis. *IEEE Transactions on Affective Computing*, pages 1–12, 2023.

[29] Sidney D'mello and Art Graesser. Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Trans. Interact. Intell. Syst.*, 2(4), jan 2013.

[30] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

[31] Paul Ekman. Basic emotions. In Tim Dalgleish and Mick Power, editors, *Handbook of Cognition and Emotion*, pages 4–5. Wiley, 1999.

[32] Paul Ekman and Wallace Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17:124–9, 02 1971.

[33] Muhammad Hammad Fahim Siddiqui, Diana Inkpen, and Alexander Gelbukh. Towards interpretable emotion classification: Evaluating lime, shap, and generative ai for decision explanations. 2024.

[34] Muhammad Hammad Fahim Siddiqui, Diana Inkpen, and Alexander Gelbukh. Towards interpretable emotion classification: Evaluating lime, shap, and generative ai for decision explanations. In *2024 28th International Conference Information Visualisation (IV)*, pages 1–6, 2024.

[35] Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:7692–7699, 04 2020.

[36] Senait Gebremichael Tesfagergish, Jurgita Kapočiūtė-Dzikienė, and Robertas Damaševičius. Zero-shot emotion detection for semi-supervised sentiment analysis using sentence transformers and ensemble learning. *Applied Sciences*, 12:8662, 08 2022.

[37] R. Geng, B. Li, Y. Li, J. Sun, and X. Zhu. Dynamic memory induction networks for few-shot text classification. 2020.

[38] Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. Affect-LM: A neural language model for customizable affective text generation. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[39] Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Cognitive Computation*, 14, 02 2021.

[40] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *Processing*, 150, 01 2009.

[41] James Jonathan Gross. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2:271 – 299, 1998.

[42] Z. Gu, X. Sun, F. Lian, Z. Kang, C. Xu, and J. Fan. Dingo: towards diverse and fine-grained instruction-following evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:18108–18116, 2024.

[43] A. Havrilla, M. Zhuravinskyi, D. Phung, A. Tiwari, J. Tow, S. Biderman, Q. Anthony, and L. Castricato. Trlx: a framework for large scale reinforcement learning from human feedback. 2023.

[44] N. Henry, M. Pedersen, M. Williams, J. Martin, and L. Donkin. Reducing echo chamber effects: an allostatic regulator for recommendation algorithms. 2023.

[45] Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model, 2024.

[46] Bairu Hou, Joe O'Connor, Jacob Andreas, Shiyu Chang, and Yang Zhang. Prompt-Boosting: Black-box text classification with ten forward passes. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 13309–13324. PMLR, 23–29 Jul 2023.

[47] D. Hu, E. Kalokerinos, and M. Tamir. Flexibility or instability? emotion goal dynamics and mental health. *Emotion*, 24:1078–1091, 2024.

[48] G. Hu, S. Wang, J. Yu, J. Zhang, Y. Sun, and D. Kong. A facile and scalable patterning approach for ultrastretchable liquid metal features. *Lab on a Chip*, 22:4933–4940, 2022.

[49] Doaa Mohey El-Din Mohamed Hussein. A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*, 30(4):330–338, 2018.

[50] Mohammed Jabreel and Antonio Moreno. A deep learning-based approach for multi-label emotion classification in tweets. *Applied Sciences*, 9:1123, 03 2019.

[51] Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. Understanding convolutional neural networks for text classification. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–65, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[52] Bhawna Jain, Gunika Goyal, and Mehak Sharma. Evaluating emotional detection classification capabilities of gpt-2 gpt-neo using textual data. In *2024 14th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, pages 12–18, 2024.

[53] Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. Learning to faithfully rationalize by construction. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online, July 2020. Association for Computational Linguistics.

[54] Cuiqing Jiang, Rao Muhammad Rashid, and Jianfei Wang. Investigating the role of social presence dimensions and information support on consumers' trust and shopping intentions. *Journal of Retailing and Consumer Services*, 51:263–270, 2019.

[55] Yan Jin, Augustine Pang, and Glen Cameron. The role of emotions in crisis responses: Inaugural test of the integrated crisis mapping (icm) model. *Corporate Communications: An International Journal*, 15:428–452, 10 2010.

[56] Brandon Joyce and Jing Deng. Sentiment analysis of tweets for the 2016 us presidential election. pages 1–4, 11 2017.

[57] Justine T. Kao and Dan Jurafsky. A computational analysis of poetic style: Imagism and its influence on modern professional and amateur poetry. *Linguistic Issues in Language Technology*, 12, October 2015.

[58] D. Kim, W. Son, S. Kwak, T. Yun, J. Park, and J. Lee. A hybrid deep learning emotion classification system using multimodal data. *Sensors*, 23:9333, 2023.

[59] Harald Kindermann and Melanie Schreiner. Iat measurement method to evaluate emotional aspects of brand perception – a pilot study. *Lecture Notes in Computer Science*, 06 2017.

[60] H. R. Kirk, B. Vidgen, P. Röttger, and S. A. Hale. Personalisation within bounds: a risk taxonomy and policy framework for the alignment of large language models with personalised feedback. 2023.

[61] A. Kolliakou, I. Bakolis, D. Chandran, L. Derczynski, N. Werbeloff, D. Osborn, K. Bontcheva, and R. Stewart. Mental health-related conversations on social media and crisis episodes: a time-series regression analysis. *Scientific Reports*, 10, 2020.

[62] klaus krippendorff. Computing krippendorff's alpha-reliability. 01 2011.

[63] Jaesung Lee, Wangduk Seo, and Dae-Won Kim. Effective evolutionary multilabel feature selection under a budget constraint. *Complexity*, 2018.

[64] Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. A text-driven rule-based system for emotion cause detection. In Diana Inkpen and Carlo Strapparava, editors, *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53, Los Angeles, CA, June 2010. Association for Computational Linguistics.

[65] Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Bin Liu, and Jianhua Tao. Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition. *Information Fusion*, 108:102367, 2024.

[66] Nankai Lin, Sihui Fu, Xiaotian Lin, and Lianxi Wang. Multi-label emotion classification based on adversarial multi-task learning. *Information Processing Management*, 59(6):103097, 2022.

[67] J. Liu and L. Yang. Knowledge-enhanced prompt learning for few-shot text classification. *Big Data and Cognitive Computing*, 8:43, 2024.

[68] Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[69] Pedro Lopes, Eduardo Silva, Cristiana Braga, Tiago Oliveira, and Luís Rosado. Xai systems evaluation: A review of human and computer-centred methods. *Applied Sciences*, 12(19), 2022.

[70] R. Mastoras, D. Iakovakis, S. Hadjidimitriou, V. Charisis, S. Kassie, T. Alsaadi, A. Khandoker, and L. Hadjileontiadis. Touchscreen typing pattern analysis for remote detection of the depressive tendency. *Scientific Reports*, 9, 2019.

[71] Michael Mctear, Zoraida Callejas, and David Griol. *The Conversational Interface*. 01 2016.

[72] Jérôme Michaud. Dynamic preferences and self-actuation of changes in language dynamics. *Language Dynamics and Change*, 2019.

[73] Saif Mohammad. emotional tweets. *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, 06 2012.

[74] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. SemEval-2018 task 1: Affect in tweets. In Marianna Apidianaki,

Saif M. Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat, editors, *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[75] Cezary Mordka. What are emotions? structure and function of emotions. *Studia Humana*, 5, 11 2016.

[76] ALENA NEVIAROUSKAYA, HELMUT PRENDINGER, and MITSURU ISHIZUKA. Affect analysis model: novel rule-based approach to affect sensing from text. *Natural Language Engineering*, 17(1):95–135, 2011.

[77] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English tweets. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online, October 2020. Association for Computational Linguistics.

[78] Shiwen Ni and Hung-Yu Kao. Kpt++: Refined knowledgeable prompt tuning for few-shot text classification. *Know.-Based Syst.*, 274(C), aug 2023.

[79] Stephen Oladele. BWorld Robot Control Software. https://encord.com/blog/interrater-reliability-krippendorffs-alpha/#:~:text=Krippendorff's%20Alpha%20values%20range%20from,implies%20systematic%20disagreements%20among%20raters.

[80] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell,

Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[81] Marek Pawlicki, Aleksandra Pawlicka, Federica Uccello, Sebastian Szelest, Salvatore D'Antonio, Rafał Kozik, and Michał Choraś. Evaluating the necessity of the multiple metrics for assessing explainable ai: A critical examination. *Neurocomputing*, 602:128282, 2024.

[82] Rosalind W. Picard. *Affective Computing*. The MIT Press, 09 1997.

[83] Juan D. Pinto and Luc Paquette. Towards a unified framework for evaluating explanations, 2024.

[84] Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, and Roman Klinger. Natural language inference prompts for zero-shot emotion classification in text across corpora. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6805–6817, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[85] Robert Plutchik. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553, 1982.

[86] JESSE PRINZ. 69Which emotions are basic? In *Emotion, Evolution, and Rationality*. Oxford University Press, 04 2004.

[87] Fan Qi, Huaiwen Zhang, Xiaoshan Yang, and Changsheng Xu. A versatile multimodal learning framework for zero-shot emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2024.

[88] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024.

[89] Dheeraj Rajagopal, Vidhisha Balachandran, Eduard Hovy, and Yulia Tsvetkov. Self-explain: A self-explaining architecture for neural text classifiers. 2021.

[90] Dheeraj Rajagopal, Vidhisha Balachandran, Eduard H Hovy, and Yulia Tsvetkov. SELFEXPLAIN: A self-explaining architecture for neural text classifiers. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 836–850, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[91] Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154, 2023.

[92] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[93] Rainer Reisenzein. What is a definition of emotion? and are emotions mental-behavioral processes? *Social Science Information*, 46:424–428, 09 2007.

[94] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[95] E. Saravia, H. T. Liu, Y. Huang, J. Wu, and Y. Chen. Carer: contextualized affect representations for emotion recognition. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[96] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[97] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[98] Klaus R. Scherer. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729, 2005.

[99] Muhammad Hammad Fahim Siddiqui, Diana Inkpen, and Alexander Gelbukh. Instruction Tuning of LLMs for Multi-label EmotionClassification in Social Media Content. *Proceedings of the Canadian Conference on Artificial Intelligence*, may 27 2024. https://caiac.pubpub.org/pub/lezimqvm.

[100] Z. Sijia, J. Zhao, and L. Zhang. Application of artificial intelligence on psychological interventions and diagnosis: an overview. *Frontiers in Psychiatry*, 13, 2022.

[101] Chaehan So. Understanding the prediction mechanism of sentiments by xai visualization. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '20, page 75–80, New York, NY, USA, 2021. Association for Computing Machinery.

[102] Rui Song, Xingbing Chen, Zelong Liu, Haining An, Zhiqi Zhang, Xiaoguang Wang, and Hao Xu. Label prompt for multi-label text classification, 2023.

[103] Stefan Stieglitz, Milad Mirbabaie, Jennifer Fromm, and Stefanie Melzer. The adoption of social media analytics for crisis management - challenges and opportunities. 06 2018.

[104] Sean Stilwell and Diana Inkpen. Explainable Prompt-based Approaches for Sentiment Analysis of Movie Reviews. *Proceedings of the Canadian Conference on Artificial Intelligence*, may 27 2024. https://caiac.pubpub.org/pub/oe1gma4v.

[105] Carlo Strapparava and Rada Mihalcea. SemEval-2007 task 14: Affective text. In Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[106] Jiazheng Sun, Xiaodong Zhang, and Shaojuan Lei. The evolution of public opinion and its emotion analysis in public health emergency based on weibo data. In Xiaopu Shang, Xiaowen Fu, Yixuan Ma, Daqing Gong, and Juliang Zhang, editors, *LISS 2022*, pages 415–434, Singapore, 2023. Springer Nature Singapore.

[107] Z. Sun, Y. Zhou, J. Hao, X. Fan, Y. Lu, C. Ma, W. Shen, and C. Guo. Improving contextual query rewrite for conversational ai agents through user-preference feed-

back learning. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2023.

[108] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.

[109] Jianhua Tao and Tieniu Tan. Affective computing: A review. pages 981–995, 10 2005.

[110] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Election forecasts with twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, 29(4):402–418, 2011.

[111] Orizu Udochukwu and Yulan He. A rule-based approach to implicit emotion detection in text. In Chris Biemann, Siegfried Handschuh, André Freitas, Farid Meziane, and Elisabeth Métais, editors, *Natural Language Processing and Information Systems*, pages 197–203, Cham, 2015. Springer International Publishing.

[112] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[113] Jianing Wang, Chengyu Wang, Fuli Luo, Chuanqi Tan, Minghui Qiu, Fei Yang, Qiuhui Shi, Songfang Huang, and Ming Gao. Towards unified prompt tuning for few-shot text classification, 2022.

[114] Meng-Jie Wang, Kumar Yogeeswaran, Sivanand Sivaram, and Kn Kn. Examining spread of emotional political content among democratic and republican candidates

during the 2018 us mid-term elections. *Humanities and Social Sciences Communications*, 8, 11 2021.

[115] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions, 2023.

[116] Zihan Wang, Peiyi Wang, Tianyu Liu, Binghuai Lin, Yunbo Cao, Zhifang Sui, and Houfeng Wang. Hpt: Hierarchy-aware prompt tuning for hierarchical text classification, 2022.

[117] Lisa Watson and Mark Spence. Causes and consequences of emotions on consumer behaviour. *European Journal of Marketing*, 41:487–511, 06 2007.

[118] Jingyuan Wen, Yutian Luo, Nanyi Fei, Guoxing Yang, Zhiwu Lu, Hao Jiang, Jie Jiang, and Zhao Cao. Visual prompt tuning for few-shot text classification. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5560–5570, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[119] Xun Xu. Examining the relevance of online customer textual reviews on hotels' product and service attributes. *Journal of Hospitality & Tourism Research*, 43(1):141–163, 2019.

[120] Seonghyeon Ye, Hyeonbin Hwang, Sohee Yang, Hyeongu Yun, Yireun Kim, and Minjoon Seo. Investigating the effectiveness of task-agnostic prefix prompt for instruction following, 2023.

[121] Wenhao Ying, Rong Xiang, and Qin Lu. Improving multi-label emotion classification by integrating both general and domain-specific knowledge. In Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi, editors, *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 316–321, Hong Kong, China, November 2019. Association for Computational Linguistics.

[122] Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. Rethinking cooperative rationalization: Introspective extraction and complement control. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4094–4103, Hong Kong, China, November 2019. Association for Computational Linguistics.

[123] Zhiqiang Yuan, Junwei Liu, Qiancheng Zi, Mingwei Liu, Xin Peng, and Yiling Lou. Evaluating instruction-tuned large language models on code comprehension and generation. *ArXiv*, abs/2308.01240, 2023.

[124] Y. Zhai, Y. Li, Z. Gao, X. Gong, K. Xu, D. Feng, B. Ding, and H. Wang. Optimistic model rollouts for pessimistic offline policy optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:16678–16686, 2024.

[125] Haoxing Zhang, Xiaofeng Zhang, Haibo Huang, and Lei Yu. Prompt-based meta-learning for few-shot text classification. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in*

Natural Language Processing, pages 1342–1357, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[126] Haoxing Zhang, Xiaofeng Zhang, Haibo Huang, and Lei Yu. Prompt-based meta-learning for few-shot text classification. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1342–1357, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[127] Junyi Zhang and Yuan Rao. Research on model and algorithm of multiview and multilabel classification based on nearest-neighbor model. *Mathematical Problems in Engineering*, 2022.

[128] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory, 2018.

1 Appendix

1.1 Prompts used

1.1.1 System prompts

Models	Prompt
Zero-shot multi-label, Zero-shot multi-class, Few-shot multi-label, Few-shot multi-class	You are an expert emotion classifier that accurately identifies emotions in tweets.
Self-explaining	You are an expert emotion classifier that accurately identifies emotions in Tweets and explains the reasoning behind them.

1.1.2 User prompts

1.1.2.1 Zero-shot

Multiclass:

Here is a tweet: [Insert Tweet Text Here]. Classify the following tweet into one of the following emotions depending on the presence of that emotion in tweet text. Emotions: ’anger’, ’fear’, ’joy’, ’love’, ’sadness’, ’surprise’. Limit the response to only the emotions.

Multilabel:

Here is a tweet: [Insert Tweet Text Here]. Classify the following tweet into one

or more of the following emotions depending on the presence of that emotion in tweet text. Emotions: ’anger’, ’anticipation’, ’disgust’, ’fear’, ’joy’, ’love’, ’optimism’, ’pessimism’, ’sadness’, ’surprise’, ’trust’. Limit the response to only the emotions.

1.1.2.2 Few-shot

#	Prompt
1	Your task is to analyze the content of the given tweet and classify it based on the following emotion labels: Emotion Labels: [anger, fear, love, surprise, sadness, joy, anticipation, disgust, optimism, pessimism, trust, neutral] Tweet: "{}" Instructions: 1. Analyze the Content: Thoroughly analyze the tweet's content, tone, and context. Look for linguistic cues, such as words, phrases, or implied sentiments that might indicate one or more emotions. 2. Chain of Thought Reasoning: Break down the tweet's components step by step. Consider the following questions as part of your reasoning: 2a. What is the speaker expressing? 2b. Does the speaker show a positive, negative, or neutral sentiment? 2c. Are there hints of longing, hope, frustration, or any other specific emotion? 2d. Is there any combination of emotions suggested by the words or tone? 3. Classify Emotions: Use the reasoning process to identify: 3a. The Most Prominent Emotion: Select the single most dominant emotion that captures the essence of the tweet. 3b. Other Present Emotions: List other emotions that might be present in the tweet, even if they are secondary. 4. If no emotions are present, then put neutral in both the most prominent emotion and other present emotions. Here are some examples of similar multilabel classifications: Examples: Example 1:

	Tweet: "{}" - multiple_emotions_present: "{}" Example 2: Tweet: "{}" - multiple_emotions_present: "{}" Example 3: Tweet: "{}" - multiple_emotions_present: "{}"
2	You are an expert in emotion recognition. Your task is to read the provided tweet and classify the emotions it conveys, following the instructions and using the emotion labels below: Emotion Labels: [anger, fear, love, surprise, sadness, joy, anticipation, disgust, optimism, pessimism, trust, neutral] Tweet: "{tweet}" Instructions: 1. Read the tweet carefully, noting the tone, context, and any emotional cues in the language. 2. Reflect on the following to guide your reasoning: - What feelings are being expressed or implied? - Is the speaker positive, negative, neutral, or experiencing mixed emotions? - Are multiple emotions apparent, or is one dominant? 3. Identify and output: - The **most prominent emotion** (the single emotion that stands out the most) - **All emotions present** (a list of any other relevant emotions from the set) 4. If the tweet does not contain any clear emotions, choose "neutral" for both fields. Examples: Tweet: "{}" multiple_emotions_present: {} Tweet: "{}" multiple_emotions_present: {} Tweet: "{}" multiple_emotions_present: {}
3	Classify the following tweet using these emotion labels: [anger, fear, love, surprise, sadness, joy, anticipation, disgust, optimism, pessimism, trust, neutral]. Tweet: "{tweet}" Instructions: - Identify the most prominent emotion.

	- List all emotions that apply (can be more than one; if none, use "neutral"). Examples: Tweet: "{}" multiple_emotions_present: {} Tweet: "{}" multiple_emotions_present: {} Tweet: "{}" multiple_emotions_present: {}
4	You are given a tweet and asked to identify its emotional content using the following labels: [anger, fear, love, surprise, sadness, joy, anticipation, disgust, optimism, pessimism, trust, neutral]. Tweet: "{tweet}" Steps: 1. Read the tweet and consider any emotional language, tone, or context. 2. Determine the single most prominent emotion the tweet expresses. 3. List all emotions that are present (there may be more than one). 4. If the tweet does not convey any particular emotion, choose "neutral" for both answers. Examples: Tweet: "{}" multiple_emotions_present: {} Tweet: "{}" multiple_emotions_present: {} Tweet: "{}" multiple_emotions_present: {}
5	Your task is to analyze the following tweet and classify its emotional content using a comprehensive set of emotion labels: [anger, fear, love, surprise, sadness, joy, anticipation, disgust, optimism, pessimism, trust, neutral]. Tweet: "{tweet}" Instructions: 1. Carefully read the tweet, considering not only explicit words but also the tone, context, and any implied emotions. 2. Reflect on which emotions the tweet may convey. Think about whether the sentiment is positive, negative, neutral, or mixed, and look for cues that suggest subtle feelings or

	combinations of emotions. 3. Select the most prominent emotion—the single emotion that is most clearly expressed or central to the tweet’s message. 4. Identify all other emotions that may be present in the tweet, beyond the most prominent one. If you believe only one emotion applies, list it for both. 5. If no emotion is detectable, or if the tweet is purely factual or neutral, use "neutral" for both the most prominent and the other emotions. Examples: Tweet: "{}" multiple_emotions_present: {} Tweet: "{}" multiple_emotions_present: {} Tweet: "{}" multiple_emotions_present: {}

1.1.2.3 Dynamic few-shot

Your task is to analyze the content of the given tweet and classify it based on the following emotion labels:

Emotion Labels: [anger, fear, love, surprise, sadness, joy, anticipation, disgust, optimism, pessimism, trust, neutral]

Tweet: "{}"

Instructions:

1. Analyze the Content: Thoroughly analyze the tweet's content, tone, and context. Look for linguistic cues, such as words, phrases, or implied sentiments that might indicate one or more emotions.

2. Chain of Thought Reasoning: Break down the tweet's components step by step. Consider the following questions as part of your reasoning:
  2a. What is the speaker expressing?
  2b. Does the speaker show a positive, negative, or neutral sentiment?
  2c. Are there hints of longing, hope, frustration, or any other specific emotion?
  2d. Is there any combination of emotions suggested by the words or tone?

3. Classify Emotions: Use the reasoning process to identify:
  3a. The Most Prominent Emotion: Select the single most dominant emotion that captures the essence

of the tweet.
  3b. Other Present Emotions: List other emotions that might be present in the tweet, even if they are secondary.

4. If no emotions are present, then put neutral in both the most prominent emotion and other present emotions.

Here are some examples of similar multilabel classifications:
Examples:

Example 1:
Tweet: "{}" - multiple_emotions_present: "{}"

Example 2:
Tweet: "{}" - multiple_emotions_present: "{}"

Example 3:
Tweet: "{}" - multiple_emotions_present: "{}"

1.1.2.4 Self-explaining

Your task is to analyze the content of the given tweet and classify it based on the following emotion labels:

Emotion Labels: [anger, fear, love, surprise, sadness, joy, anticipation, disgust, optimism, pessimism, trust, neutral]

Tweet: "{}"

Instructions:

1. Analyze the Content: Thoroughly analyze the tweet's content, tone, and context. Look for linguistic cues, such as words, phrases, or implied sentiments that might indicate one or more emotions.

2. Chain of Thought Reasoning: Break down the tweet's components step by step. Consider the following questions as part of your reasoning:
  2a. What is the speaker expressing?
  2b. Does the speaker show a positive, negative, or neutral sentiment?
  2c. Are there hints of longing, hope, frustration, or any other specific emotion?
  2d. Is there any combination of emotions suggested by the words or tone?

3. Classify Emotions: Use the reasoning process to identify:
  3a. The Most Prominent Emotion: Select the single most dominant emotion that captures the essence

of the tweet.
3b. Other Present Emotions: List other emotions that might be present in the tweet, even if they are secondary.

4. Explain the Classification: Explain your classification. Clearly articulate:
4a. Why did you choose the most prominent emotion?
4b. Why do you believe the secondary emotions are present?
4c. Reference specific parts of the tweet that support your reasoning.

5. If no emotions are present, then put neutral in both the most prominent emotion and other present emotions.

Here are some examples of similar multilabel classifications:
Examples:

Example 1:
Tweet: "{}" - multiple_emotions_present: "{}"

Example 2:
Tweet: "{}" - multiple_emotions_present: "{}"

Example 3:
Tweet: "{}" - multiple_emotions_present: "{}"

1.2 Datasets being used:

1. DAIR AI emotion classification dataset (20k | 16k: 2k: 2k) - choose one class out of 6
2. SemEval 2018 multilabel emotion classification (11k | 6.8k: 0.9k: 3.2k)

Emotions present:

DAIR AI	SemEval 2018
anger	anger
fear	fear
love	love
surprise	surprise
sadness	sadness

joy	joy
	anticipation
	disgust
	optimism
	pessimism
	trust

All the emotions present in the DAIR AI dataset are present in the SemEval dataset. Both datasets consist of tweets.

In the thesis proposal document, we committed to creating a model that can perform single-class and multi-label emotion classifications, as well as provide explanations. We aim to get the following output format from the model.

Tweet:

"I just want to fast forward to the part of my day where I take a nap between work and cheer."

Sample output (after appending tweet to the prompt):

{

"**explanation**":

"The tweet expresses a mix of emotions, primarily revolving around the desire for rest and relief. The phrase "I just want to fast forward" conveys impatience and anticipation as the speaker looks forward to a specific part of their day. The mention of a "nap" implies a longing for comfort and relaxation, suggesting a sense of tiredness that might align with sadness or pessimism about the current state of their day. However, the inclusion of "cheer" indicates a positive undertone, reflecting optimism or hope that the later part of the day will be more enjoyable. The overall tone balances between frustration and hopefulness, highlighting a combination of anticipation, optimism, and a touch of sadness."

"**most_prominent_emotion**": "anticipation,"

"**multiple_emotions_present**": ["anticipation," "sadness," "optimism," "pessimism"]

}

1.3 Annotation guidelines for the preference dataset

1.3.1 Introduction:

1.3.1.1 Problem definition:

In the domain of multi-label emotion classification on Twitter, we aim to develop models that predict relevant emotions and provide human-readable explanations for those predictions. Since multiple valid ways exist to explain or present the classification, we collect pairs of model-generated responses for each tweet. Annotators are tasked with evaluating these responses based on specific dimensions (e.g., clarity, correctness, helpfulness, verbosity), ultimately selecting which response they find more suitable overall.

1.3.1.2 Need for the dataset:

Explainable AI: The dataset helps us assess how well models explain their reasoning for emotion classification.
Quality Control: By collecting human preferences between two responses, we can better train and refine model responses to align with human expectations.
Multi-Label Complexity: Tweets may contain multiple overlapping emotions. Evaluating correctness in this context (matching the gold labels) is crucial to producing more trustworthy AI systems.

1.3.2 Dataset Fields (Already Provided)

Each record in the dataset contains the following **fixed fields** (annotators do **not** fill these, they are already given):

1. **ID**: The unique identifier (Tweet ID).
2. **Tweet**: The original Tweet text.
3. **Prompt**: The exact prompt used to generate the two responses.
4. **Gold_multilabel**: The gold (human-annotated) set of emotions present in the Tweet (from the set {anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust}).

5. **Response1**: The first model-generated response.
6. **Response2**: The second model-generated response.
7. **Explanation1**: The explanation given by the first model.
8. **Predicted_multilabel_1**: The set of predicted emotions (multiple) by the first response.
9. **Predicted_single_1**: The single major emotion predicted by the first response (from the set {anger, anticipation, disgust, fear, joy, love}).
10. **Explanation2**: The explanation given by the second model.
11. **Predicted_multilabel_2**: The set of predicted emotions (multiple) by the second response.
12. **Predicted_single_2**: The single major emotion predicted by the second response (from the set {anger, anticipation, disgust, fear, joy, love}).

1.3.3 Fields for Annotators to Fill

Annotators will provide evaluations for each record by filling out the following **annotation fields**:

1. **Chosen response**: Indicate which response (Response1 or Response2) you prefer.
2. **Clarity_response_1** (1–5)
3. **Correctness_response_1** (1–5)
4. **Helpfulness_response_1** (1–5)
5. **Verbosity_response_1** (1–5)
6. **Clarity_response_2** (1–5)
7. **Correctness_response_2** (1–5)
8. **Helpfulness_response_2** (1–5)
9. **Verbosity_response_2** (1–5)
10. **Overall_rating_response_1** (1–5)
11. **Overall_rating_response_2** (1–5)

1.3.4 Dimension Definitions

Below are detailed definitions of each dimension and instructions for scoring them on a scale from 1 to 5 (where **1** is the lowest score and **5** is the highest). A more detailed **rating reference table** is provided in Table 1.

1.3.4.1 Clarity

- **Definition**: How clear and unambiguous the explanation is. Does the explanation use concise language and avoid vague or contradictory statements?
- **Scoring Considerations**:
  - A score of **1** means the explanation is almost completely unclear or riddled with contradictions.

    - A score of **5** means the explanation is perfectly clear, with no ambiguity or unclear references.

1.3.4.2 Correctness

- **Definition**: How accurately the response's predicted emotions match the gold labels, and how well the explanation supports those labels.
- **Scoring Considerations**:
    - **Highest priority** dimension in deciding the final chosen response.
    - If the **Predicted_multilabel** exactly matches **Gold_multilabel**, it automatically leans towards a higher correctness score (potentially a **5** if explanation is also coherent).
    - Partial mismatches lower the score.
    - Major mistakes or complete mismatch should be scored low (e.g., **1** for entirely wrong).
    - Consider also if the explanation logically supports the predicted labels.

1.3.4.3 Helpfulness

- **Definition**: How effectively the explanation helps a human understand why certain emotions were predicted.
- **Scoring Considerations**:
    - A helpful explanation should highlight relevant parts of the tweet or reasoning that led to the classification.
    - It should give the user actionable or understandable insight into the reasoning process.
    - Overly generic or repetitive statements are less helpful.

1.3.4.4 Verbosity

- **Definition**: How the explanation balances detail with conciseness—avoiding being too short (missing information) or too long (unnecessarily repetitive).
- **Scoring Considerations**:
    - We do **not** want very short (lack of detail) or overly lengthy explanations (risk of filler content or hallucination).
    - A "good" explanation should be direct, sufficiently detailed, but not verbose to the point of including irrelevant information.

1.3.5 Overall Rating

Overall_rating_response_X (1–5): An overall assessment of each response, considering the four specific dimensions (Clarity, Correctness, Helpfulness, Verbosity).

Both responses can have the same overall rating (e.g., 5). However, annotators must still **choose** their preferred response (see **Chosen response** below).

1.3.6 Chosen Response

- **Definition**: Indicate whether you prefer **Response1** or **Response2**.
- **Important**: **The correctness** dimension should have the highest weight in this decision. That is, if one response is more accurate in matching the gold labels, that response typically should be chosen unless its other dimensions are extremely poor.

1.3.7 Step-by-Step Annotation Procedure

For each record, follow these steps:

1. **Read the Tweet** and **Gold_multilabel**: Understand the ground-truth emotions.
2. **Review Response1**: Examine its **Explanation1**, **Predicted_multilabel_1**, and **Predicted_single_1**.
    - Fill in **Clarity_response_1**, **Correctness_response_1**, **Helpfulness_response_1**, **Verbosity_response_1**, and **Overall_rating_response_1** according to the criteria below.
3. **Review Response2**: Examine its **Explanation2**, **Predicted_multilabel_2**, and **Predicted_single_2**.
    - Fill in **Clarity_response_2**, **Correctness_response_2**, **Helpfulness_response_2**, **Verbosity_response_2**, and **Overall_rating_response_2** according to the criteria below.
4. **Compare** the two responses, focusing mainly on **Correctness** (does the predicted multilabel match the gold multilabel?). Then, review clarity, helpfulness, verbosity, and overall rating.
5. **Choose** which response you prefer in the field. Choose your **response**.

1.3.8 Detailed Scoring Guidelines

Below is a reference table providing examples for each rating (1–5) across the four dimensions. Use this table as a guide to maintain consistency.

Note (important): If the chosen model response explanation lacks some dimensions or overall structure, please revise it to the best of your ability and highlight that row in yellow so the final reviewer can verify your changes.

Dimension	Score = 1	Score = 2	Score = 3	Score = 4	Score = 5

	(Poor)	(Fair)	(Average)	(Good)	(Excellent
Clarity	- Explanation is confusing or incoherent. - Uses ambiguous or contradictory language. - Impossible to understand.	- Somewhat unclear or partially incoherent. - Leaves important reasoning points ambiguous.	- Moderately clear but contains minor ambiguities. - The main idea is understandable, but details may be fuzzy.	- Mostly clear and well-structured. - Minor ambiguities may exist but do not hinder overall comprehension.	- Extremely clear, direct, and unambiguous. - No contradictions. - Easy to follow and coherent throughout.
Correctness	- Predicted emotions are completely mismatched from **Gold_multilabel**. - Explanation contradicts the gold labels.	- Some predicted emotions overlap but critical emotions are missing or incorrect. - Explanation partially contradicts the gold.	- Around half of the predicted emotions are correct or major label is correct but minor labels are missed. - Explanation is somewhat aligned with correct labels.	- Predicted multilabel nearly matches the gold, with minor omissions or extraneous labels. - Explanation mostly supports them.	- Predicted multilabel is an exact match with **Gold_multilabel**. - Explanation fully supports the predicted labels accurately. - *Highest priority dimension for final choice.*
Helpfulness	- Explanation provides no insight into the tweet's content. - Mostly generic or irrelevant statements.	- Explanation is partially useful but lacks depth. - Some relevant points are mentioned, but not well detailed.	- Explanation is somewhat helpful, covers some relevant details about the tweet. - Partial or inconsistent rationale.	- Explanation is helpful, referencing key words or phrases from the tweet. - Clear rationale for at least major emotions.	- Explanation is highly informative, pointing directly to relevant cues from the tweet. - Provides strong rationale that is easy to understand.
Verbosity	- Explanation is overly long or too short.	-Explanation's length is somewhat	- Explanation length is moderate but	- Explanation is generally concise.	- Explanation is well-structured

	- Rambles without focus or misses core details.	imbalanced (too brief or somewhat wordy). - Contains repetitive or irrelevant details.	could improve on conciseness or detail. - Mild redundancy or brevity issues.	- Provides a sufficient level of detail without unnecessary information.	, concise yet complete. - No excessive repetition or irrelevant content. - Very balanced presentation.

1.3.9 Choosing the Final Preferred Response

1. **Start with Correctness**: Identify the response that best matches the **Gold_multilabel**. If one response is entirely correct and the other is not, typically choose the correct one.
2. **Consider Other Dimensions**: If both responses are equally correct or differ only slightly in correctness, use the other dimensions (Clarity, Helpfulness, Verbosity) to decide which is superior.
3. **Overall Ratings**: After considering each dimension, check the **Overall_rating_response_X** you gave to each response. In rare cases, both might be perfect (score of 5). You can still prefer one over the other based on subtle differences—e.g., maybe one is slightly clearer or more helpful.
4. **Document Your Choice**: Select **"response1"** or **"response2"** in **Chosen response** field.

1.3.10 Summary of the Annotation Fields to Complete

For each record (one tweet with two model responses):

1. **Chosen response**: "response1" or "response2".
2. **Clarity_response_1**: 1–5
3. **Correctness_response_1**: 1–5
4. **Helpfulness_response_1**: 1–5
5. **Verbosity_response_1**: 1–5
6. **Clarity_response_2**: 1–5
7. **Correctness_response_2**: 1–5
8. **Helpfulness_response_2**: 1–5
9. **Verbosity_response_2**: 1–5
10. **Overall_rating_response_1**: 1–5
11. **Overall_rating_response_2**: 1–5

1.3.11 Final Notes and Best Practices

- **Be Consistent**: Try to apply the same logic and thresholds across all tweets and responses.
- **Focus on Correctness**: Remember, correctness (matching the gold labels) is the most critical factor. A perfectly written explanation that misclassifies the tweet should not outweigh a correct classification with a slightly less polished explanation.
- **No Overthinking**: If two responses are extremely similar, having the same overall rating is acceptable. However, **still pick** one as the "Chosen response" (possibly by a minor difference in clarity or helpfulness).