BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF COMPUTER SYSTEMS
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

INTERPRETATION OF EMOTIONS FROM TEXT ON SOCIAL MEDIA

INTERPRETACE EMOCÍ Z TEXTU NA SOCIÁLNÍCH SÍTÍCH

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTOR
AUTOR PRÁCE

Bc. VÍT TLUSTOŠ

SUPERVISOR
VEDOUCÍ PRÁCE

doc. AAMIR SAEED MALIK, Ph.D.

BRNO 2024

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Master's Thesis Assignment

153407

Institut:	Department of Computer Systems (DCSY)
Student:	**Tlustoš Vít, Bc.**
Programme:	Information Technology and Artificial Intelligence
Specialization:	Computer Vision
Title:	**Interpretation of emotions from text on social media**
Category:	Biocomputing
Academic year:	2023/24

Assignment:

1. Study and learn about the various emotions that can be interpreted/ recognized from the textual input on social media.
2. Get acquainted with text processing methods as well as machine learning techniques and their application to the interpretation/ recognition of emotions.
3. Find out the challenges for emotion interpretation/ recognition from text on social media as well as the limitations of the existing methods.
4. Design an algorithm for interpretation/ recognition of emotion from text input on social media.
5. Implement the designed algorithm.
6. Create a set of benchmark tasks to evaluate the quality of emotion interpretation/ recognition from text on social media as well as the corresponding computational performance and memory usage.
7. Conduct critical analysis and discuss the achieved results and their contribution.

Literature:

- According to supervisor's advice.

Requirements for the semestral defence:

- Items 1 to 4 of the assignment.

Detailed formal requirements can be found at https://www.fit.vut.cz/study/theses/

Supervisor:	**Malik Aamir Saeed, doc., Ph.D.**
Head of Department:	Sekanina Lukáš, prof. Ing., Ph.D.
Beginning of work:	1.11.2023
Submission deadline:	17.5.2024
Approval date:	30.10.2023

Abstract

Most human interactions are either text-based or can be converted to text using *speech-to-text* technologies. This thesis is dedicated to recognizing emotions from these texts. Despite extensive research in this domain, three significant challenges persisted: unexplored or limited cross-domain efficacy of the methods, superficial analysis of the result, and limited usability of the outcomes. We address these challenges by proposing two models based on the *RoBERTa* model, which we call *EmoMosaic-base* and *EmoMosaic-large*. These models were trained on the following datasets: *SemEval-2018 Task 1: Affect in Tweets*, *GoEmotions*, *XED*, and *DailyDialog* datasets. In contrast to prior studies, we trained our models on all the datasets simultaneously while preserving their original categories. This resulted in models that exhibit strong performance across diverse domains and are directly comparable to other methods. In fact, *EmoMosaic-large* outperforms recent single-domain *state-of-the-art* models on *SemEval-2018 Task 1: Affect in Tweets* and *GoEmotions* datasets, demonstrating outstanding cross-domain performance. To promote the usability and reproducibility of our research, we make all our code and models public, available at: https://huggingface.co/vtlustos.

Abstrakt

Většina lidských interakcí probíhá buď prostřednictvím textu, nebo může být na text převedena pomocí *speech-to-text* technologií. Tato práce je věnována rozpoznávání emocí z takovýchto textů. Navzdory rozsáhlému výzkumu v této oblasti tři významné problémy přetrvávaly: neprozkoumaná nebo omezená účinnost metod napříč doménami, povrchní analýza výsledků a omezená použitelnost výstupů. Tyto výzvy řešíme navržením dvou modelů založených na modelu *RoBERTa*, které nazýváme *EmoMosaic-base* a *EmoMosaic-large*. Tyto modely byly trénovány na následujicích datasetech: *SemEval-2018 Task 1: Affect in Tweets*, *GoEmotions*, *XED* a *DailyDialog*. Na rozdíl od ostatních studií jsme naše modely trénovali na všech uvedených datasetech současně, přičemž jsme zachovali jejich původní kategorie. Výsledkem jsou modely, které dobře fungují napříč různými doménami a jsou přímo porovnatelné s ostatními metodami. Model *EmoMosaic-large* dokonce překonává nedávné jedno-doménové *state-of-the-art* modely na datasetech *SemEval-2018 Task 1: Affect in Tweets* a *GoEmotions*, což dokazuje jeho vynikající schopnosti napříč různými oblastmi. Pro zvýšení využitelnosti a reprodukovatelnosti našeho výzkumu poskytujeme veškerý kód a modely veřejně na: https://huggingface.co/vtlustos.

Keywords

emotion classification from text, emotion recognition from text, cross-domain emotion recognition, GoEmotions, DailyDialog, XED, SemEval-2018 Task 1

Klíčová slova

klasifikace emocí z textu, rozpoznávání emocí z textu, rozpoznávání emocí napříč doménami, GoEmotions, DailyDialog, XED, SemEval-2018 Task 1

Reference

TLUSTOŠ, Vít. *Interpretation of emotions from text on social media.* Brno, 2024. Master's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor doc. Aamir Saeed Malik, Ph.D.

Rozšířený abstrakt

Většina lidských interakcí probíhá buď prostřednictvím textu, nebo může být na text převedena pomocí *speech-to-text* technologií. Tato práce je věnována rozpoznávání emocí z takovýchto textů. Začali jsme důkladným a rozsáhlým přehledem literatury. Tato analýza mimo jiné zdůraznila potřebu rozmanitých, kvalitních datasetů pro vývoj odolných a dobře generalizujících metod. Datasety *SemEval-2018 Task 1: Affect in Tweets*, *GoEmotions*, *XED* a *DailyDialog* jsme identifikovaly jako ty, které v případě sloučení splňují tato kritéria. Další podrobnosti najdete v Kapitole 2.2. Z analýzy také vyplynulo, že modely založené na Transformerech jako jsou *BERT*, *RoBERTa* a *BART* dosahují nejlepších výsledků na následujících datasetech *SemEval-2018 Task 1: Affect in Tweets* a *GoEmotions*. Na základě reportovaných výsledků na datasetu *DailyDialog*, jsme usoudili, že v oblasti zpracování konverzací nejlepších výsledků dosahují hybridní modely kombinující Transformer enkodéry jako je *BERT* s grafovými neuronovými sítěmi (GANs). Dalším výstupem z této analýzy, byla identifikace následujících klíčových oblastí pro zlepšení. Jednotlivé oblasti jsou podrobně popsány v Kapitole 2.5 a zahrnují:

- Účinnost napříč doménami: Většina současných metod je zaměřena na jednu konkrétní doménu, jako je nejčastěji Twitter a Reddit, a obvykle již není testována v rámci jiných domén. To výrazně omezuje aplikovatelnost takovýchto metod v reálných situacích, které jsou zpravidla rozmanité. Studie, které se snaží cílit na mezi-doménovou efektivitu, problém do značné míry zjednodušují. Za účelem sjednotit obecně se lišící kategorie (napříč datasety), přemapují typicky velký počet kategorií na výrazně menší. Jako příklad lze uvést přemapování 27 kategorií z datasetu *GoEmotions* na 6 základních emocí podle Ekmanova modelu. Takové to přemapování udělají obdobně i pro ostatní datasety. Tento přístup však vede ke značné ztrátě úrovně detailu a znemožní srovnání s ostatními metodami.

- Analýza výsledků: Posouzení většího počtu metrik na úrovni jednotlivých datasetů, ale také na úrovni kategorií/emocí je nezbytné pro komplexní analýzu výkonosti modelů. Nicméně většina studií typicky uvádí 1-3 globální metriky, což rozhodně není dostatečné pro komplexní posouzení výkonosti modelu. Dále pak žádná ze studií neposuzovala kalibraci jejich metody.

- Použitelnost výsledků: Výzkumníci často zveřejňují pouze články bez doprovodného kódu nebo natrénovaných modelů. To brání reprodukovatelnosti a praktickému využití jejich poznatků.

V rámci této studie navrhujeme metodu, která adresuje všechny uvedené výzvy. Začali jsme výběrem vhodných datasetů a pokračovali jejich převodem na jednotný formát. Konkrétně jsme zvolili *SemEval-2018 Task 1: Affect in Tweets*, *GoEmotions*, *XED* a *DailyDialog*. Kombinací těchto datasetů jsme vytvořili rozmanitý, kvalitní, *multi-label* dataset, který nazýváme *EmoMosaic-dataset*. Důležité je zmínit, že jsme u všech datasetů zachovali původní kategorie, narozdíl od ostatních studií. Další podrobnosti naleznete v Kapitole 3.2.1.

V rámci této studie představujeme dva modely, *EmoMosaic-base* a *EmoMosaic-large*, z nichž první je založen na modelu *RoBERTa-base* a druhý na modelu *RoBERTa-large*. Na rozdíl od ostatních studií jsme naše modely trénovali na všech uvedených datasetech současně, přičemž jsme zachovali jejich původní kategorie. Abychom tohoto cíle dosáhli, naše modely zpracovávají věty bez ohledu na konkrétní dataset či kategorii. Tím jsou nuceny předpovídat celé spektrum kategorií, které vzniklo sjednocením emocí z jednotlivých

datasetů. Jelikož nepřemapováváme kategorie, tak každá věta v našem sjednoceném datasetu obsahuje několik kategorií, pro které nemá anotace. Abychom předešli případným chybám, během tréninku tyto kategorie maskujeme. Výsledkem jsou modely, které dobře fungují napříč různými doménami a jsou přímo porovnatelné s ostatními metodami. Zároveň jelikož jsme neredukovali množinu kategorií žádného z použitých datasetů, naše modely si zachovávají původní úroveň detailu, což umožňuje pochopení i složitějších emocí. Další podrobnosti najdete v Kapitole 3.2.

Po natrénování modelů jsme přistoupili k jejich podrobnému otestování. Postupovali jsme podle postupů uvedených v Kapitole 3.2.4. Nejdříve jsme vyhodnotili jejich výkon na úrovni jednotlivých datasetů, následně pak na úrovni jednotlivých kategorií/emocí. Náš nejlépe fungující model *EmoMosaic-large* prokázal vynikající výsledky napříč doménami a předčil aktuální *state-of-the-art* (SOTA) modely na následujících datasetech: *SemEval-2018 Task 1: Affect in Tweets* a *GoEmotions*. *EmoMosaic-large* dosahuje na datasetu *SemEval-2018 Task 1: Affect in Tweets* makro-průměrovaného F1 skóre 60.72 % (nárůst o 0.42 % oproti SOTA) a v datasetu *GoEmotions* 53.93 % (nárůst o 0.13 % oproti SOTA). Jeho menší verze, *EmoMosaic-base*, sice nedosahuje výsledků SOTA modelů a v průměru (počítaném napříč všemi datasety) za ním zaostává o 1.94 % v makro-průměrovaném F1 skóre. Vzhledem k jeho přibližně třetinové velikosti stále však nabízí vynikající poměr výkonu a výpočetní náročnosti. Následně jsme prostřednictvím datasetu *DailyDialog* otestovali schopnost našich modelů pracovat s konverzacemi. Model *EmoMosaic-large* dosáhl mikro-průměrovaného F1 skóre 60.65% (3,56% pokles oproti SOTA). Vzhledem k absolutní hodnotě lze usuzovat, že si vedl poměrně dobře, ale nepřekonal nejlepší soudobé metody, které byly ve všech případech založeny na hybridních modelech kombinujících Transformer enkodéry a grafové neuronové sítě (GATs). Detailní porovnání a diskuzi výsledků nalzenete v Kapitole 4. Následovala analýza na úrovni jednotlivých kategorií, ze které vyplynuly konkrétní silné a slabé stránky našich modelů. Detailní analýzu naleznete v Kapitole 4.2. Poté jsme zhodnotili kalibraci našich modelů a zjistili, že oba navrhované modely jsou poměrně dobře kalibrovány, proto považujeme jejich predikce za důvěryhodné. Analýzu jsem zakončili empirickým testováním našich modelů v různých situacích. Ačkoli si oba naše modely obecně vedly dobře zjistili jsme, že špatně reagují na věty vykazující ironii. Kromě toho jsme učinili závěr, že model *EmoMosaic-base* není příliš vhodný pro zpracování konverzací. Další podrobnosti nalezente v Kapitole 4.3. Žádný z modelů však nevykazoval systematické chyby a oba modely se osvědčily pro zpracování textů z různých domén, což byl hlavní cíl této studie.

Tím, že jsme navrhli a podrobně otestovali modely, které dobře fungují napříč různými doménami, jsme posunuli oblast rozpoznávání emocí z textu vpřed. Dále zveřejněním všech našich modelů a kódu, dostupné na https://huggingface.co/vtlustos, jsme zvýšili reprodukovatelnost a využitelnost našeho výzkumu. Další výzkum by se mohl zaměřit na zvýšení přesnosti těchto systémů, i když předpokládáme, že dosažení makro-průměrovaných F1 skóre nad 75 % nemusí být dosažitelné zejména kvůli nejednoznačnosti emočních projevů. Další oblastí výzkumu by mohlo být vytvoření tzv. *human baseline*, která by umožnila srovnání mezi modely a lidským faktorem. Další oblast by mohla zahrnovat adaptaci velkých jazykových modelů (LLMs). Ačkoliv LLMs zatím nejsou vhodné pro rozpoznávání emocí, jejich neustálý vývoj naznačuje, že v budoucnu by mohly být. LLM pro rozpoznávání emocí by mohly zvýšit flexibilitu těchto systémů, protože nevyžadují zanesení konkrétních kategorií do jejich architektury.

Interpretation of emotions from text on social media

Declaration

I hereby declare that this Master's thesis was prepared as an original work by the author under the supervision of doc. Aamir Saeed Malik Ph.D. I have listed all the literary sources, publications and other sources, which were used during the preparation of this thesis.

....
Vít Tlustoš
May 13, 2024

Acknowledgements

I would like to express my gratitude for the thorough guidance and support that I have received from my supervisor doc. Aamir Saeed Malik Ph.D.

Contents

Chapter 1

Introduction

Most human interactions are either text-based or can be transcribed to text using speech-to-text technologies. This thesis is dedicated to recognizing emotions from these texts. We can think of numerous practical applications for emotion recognition, ranging from mental health support to marketing. For instance, we can offer timely assistance to those in need or create a safer and better online space by filtering hateful and offensive content. Overall, emotion recognition can greatly improve our daily lives in many ways.

Deciphering human emotions from texts is a complex and multidisciplinary challenge, especially since human expressions are often ambiguous. The first hurdle is identifying the range of emotions people can experience. Several well-established psychological models describe human emotions, yet there is no agreement within the scientific community about a single, universally accepted model. [38] From a technical perspective, interpreting human emotions is also a significant challenge, as emotions are often expressed implicitly and indirectly. Moreover, ironic or sarcastic sentences complicate things even more, as they can be difficult for humans to understand and even more so for artificial models. [4] [38] Fortunately, with continuous developments in natural language processing, a research field that deals with text processing, machines are getting incrementally better at understanding written texts. As a result, they can now be used for various applications that require complex understanding, such as emotion recognition. [7] [25] [28] [10] [9] [32] [37] [12] [33] [8] [37] [30] [5] [30] [2] [16] [24] [14] [27] [40] [36] [22] [13] [18] [6] [11] [39]

Emotion recognition from text has been extensively researched for the past two decades, yet there is still significant room for improvement. For instance, no model yet demonstrates strong cross-domain performance, meaning that it would work well in diverse situations and contexts. Moreover, while relying on a limited number of metrics, most published works provide a superficial analysis of their results, lacking a comprehensive understanding of model behaviour. Furthermore, authors often do not share their models and code, which limits the practical applications of their outcomes. We aim to tackle some of these issues throughout this thesis.

Chapter 2 thoroughly reviews relevant literature (including psychological theories and models), setting the stage for this research. Chapter 3 clearly defines the research goals we addressed and outlines our methodology. We discuss dataset selection and model design, training and validation procedures. Chapter 4 presents our experimental results, highlighting the strengths and weaknesses of our models. Additionally, we compare our models with recent state-of-the-art methods. Chapter 5 summarizes our findings, emphasises the cross-domain efficacy of our proposed models and discusses possible directions for future research.

Chapter 2

Literature Review

This chapter introduces psychological models and theories relevant to this research. Additionally, it presents a thorough review of datasets and key literature, concluding with a summary of the limitations and gaps in current approaches, setting the stage for this research.

2.1 Psychological Models of Emotions

There isn't any consensus within the scientific community regarding a singular, universally accepted psychological model that completely and exclusively describes all human emotions. All existing models can be divided into two groups:

- Categorical models: represent emotions as a finite set of named entities that may have defined relationships. Generally, a set of basic/core emotions is defined. Additionally, relationships between the emotions may be formed, resulting in secondary emotions. These relationships describe what happens if two or more emotions are experienced simultaneously. For example, Plutchik's Wheel of Emotions Model defines a composition of joy and surprise as delight (secondary emotion). [1] [38]
- Dimensional models: define emotions as coordinates within a specific coordinate system. Emotions are then represented by varying degrees of its dimensions. The relationships are expressed implicitly by the nature of the system. An example of such taxonomy is Circumplex's Model of Affect, which represents emotions by varying degrees of valence and arousal dimensions. [38]

2.1.1 Paul Ekman's Model

Paul Ekman's Model, published in a study of facial expressions and emotions, is one of the simplest categorical models that is still, to some extent, accepted by the scientific community. It defines six primary/core emotions: *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*. The model does not define any relationships between the emotions. Therefore, adapting this model is relatively straightforward because annotating data precisely is simple due to the limited number of emotions. However, this can also be perceived as a disadvantage since a fine-graded emotion classification using this model is impossible. [38]

2.1.2 Wheel of Emotions Model

Another influential and widely adopted model is the Wheel of Emotions, also known as Plutchik's Model. It is based on eight so-called primary emotions: *joy*, *sadness*, *anger*, *fear*, *trust*, *disgust*, *surprise*, and *anticiation*. The emotions are carefully placed within the model, forming a structure depicted in Figure 2.1. Similar emotions are generally placed close to each other, whereas opposite emotions are placed 180 degrees apart. Intense emotions are placed near the model's centre, whereas mildly intense emotions are located closer to the edge. Furthermore, the model defines a relationship between emotions, as depicted by Figure 2.2, forming dyads (pairs of emotions) and triads (a mixture of three emotions). Dyads are also characterised based on the distance between those emotions: primary - 1 petal apart, secondary - two petals apart, tertiary - three petals apart and opposite dyad - 4 petals apart. The Wheel of Emotions model offers a comprehensive framework covering a wide spectrum of emotions. It can be adapted in various machine learning applications since precise data annotation is possible. However, due to its complexity, this can be quite challenging and may require a consensus across multiple annotators. [38] [31]

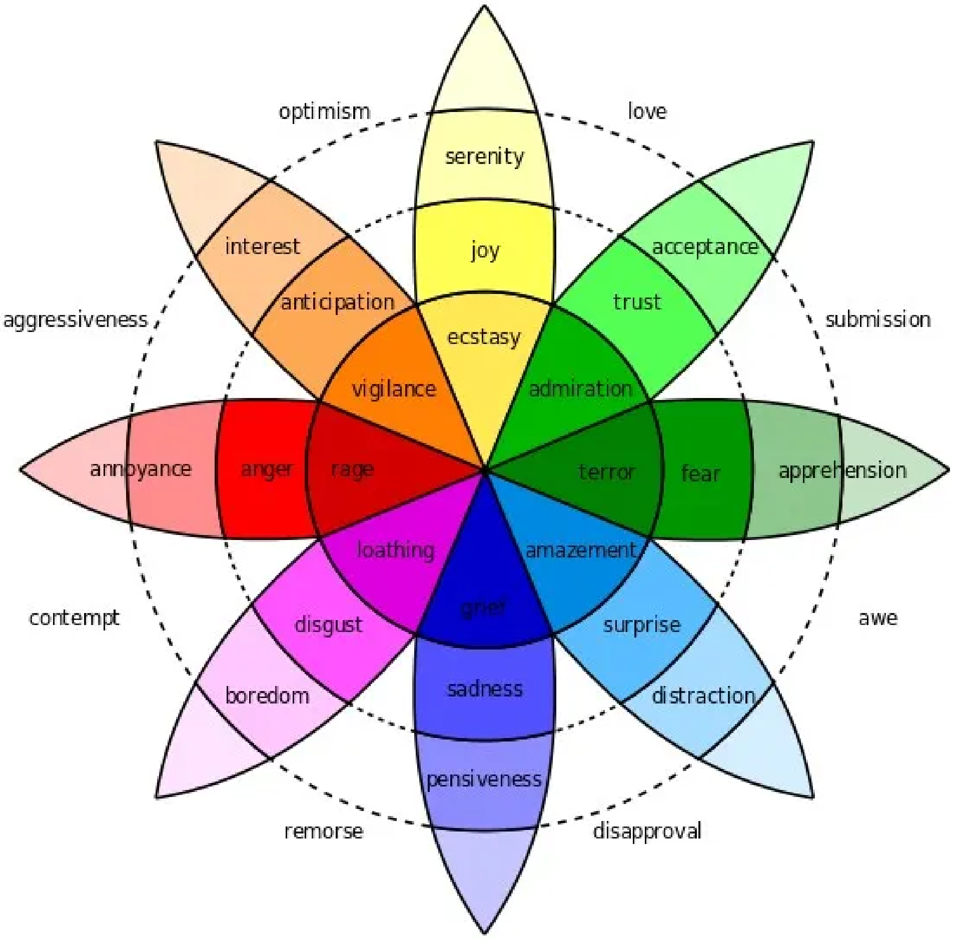


Figure 2.1: Depicts the Wheel of Emotions model. Taken from [1].

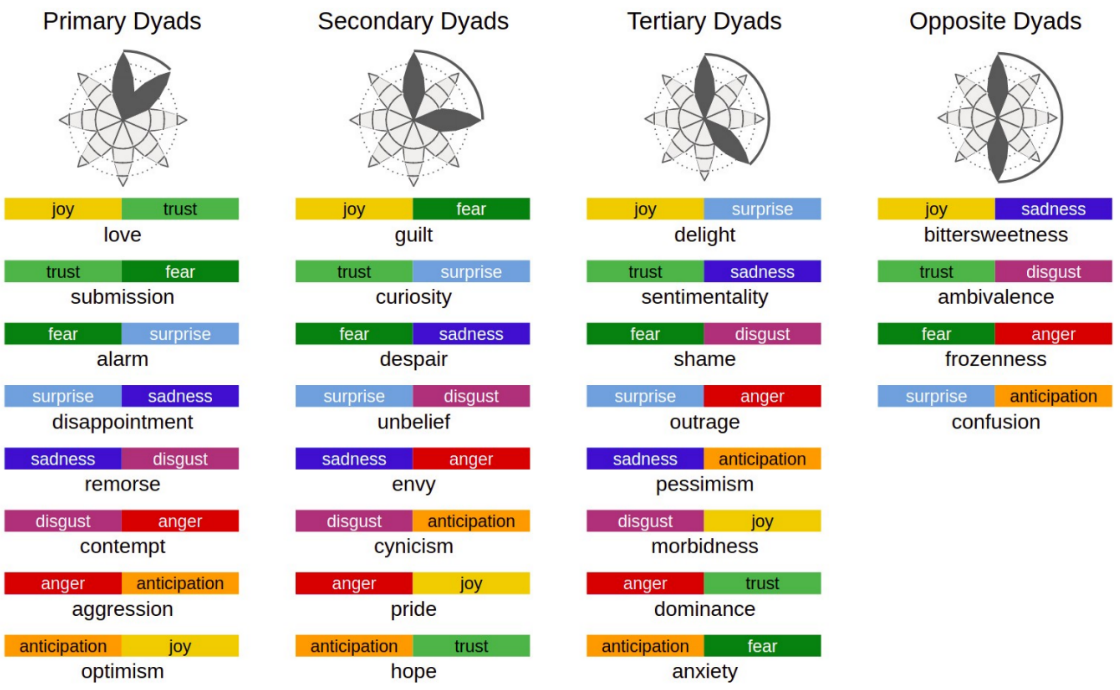


Figure 2.2: Shows dyads and triads as defined for the Wheel of Emotions Model. Taken from [31].

2.1.3 Parrot's Model

Parrot's model divides the emotions into layers, as depicted in Figure 2.3, forming primary, secondary and tertiary emotions. In total, this model defines 100 unique emotions. It is often visualized as a tree where a connection between nodes represents a relationship between emotions. This model can be used for nuanced emotion classification. However, precise data annotation can be challenging and sometimes somewhat subjective due to its complexity. [38]

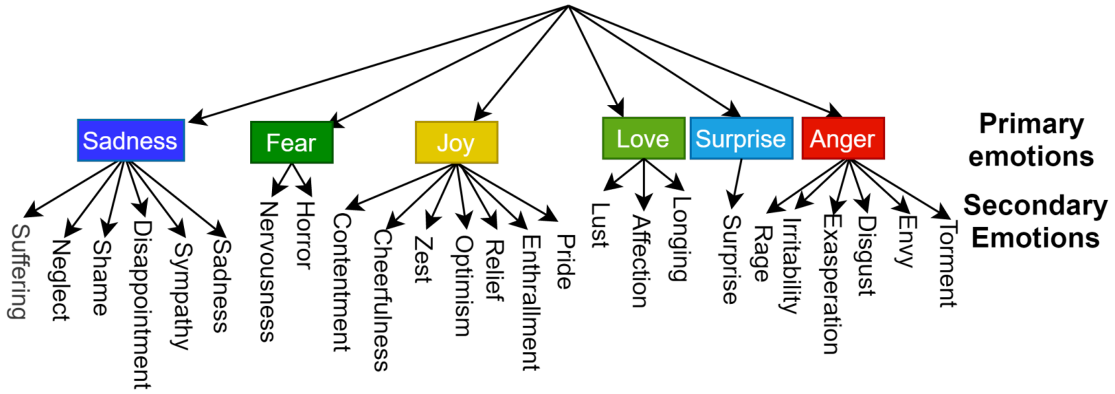


Figure 2.3: Displays the first two layers (primary and secondary emotions) of the Parrot model. Image taken from [38].

2.1.4 Circumplex's Model of Affect

The Circumplex Model of Affect is widely respected by the scientific community, a continuous model having two dimensions: valence and arousal. Valence describes the emotional experience, ranging from pleasant to unpleasant. Arousal expresses the intensity of the experience, ranging from low/calm to high/excited. The centre of the graph represents a neutral emotion. Emotions are represented by varying degrees of valence and arousal. The model is shown using Figure 2.4. This model can be used when a continuous representation of emotions is needed. However, precise data annotation is impossible since the emotions have no clear boundaries. [38]

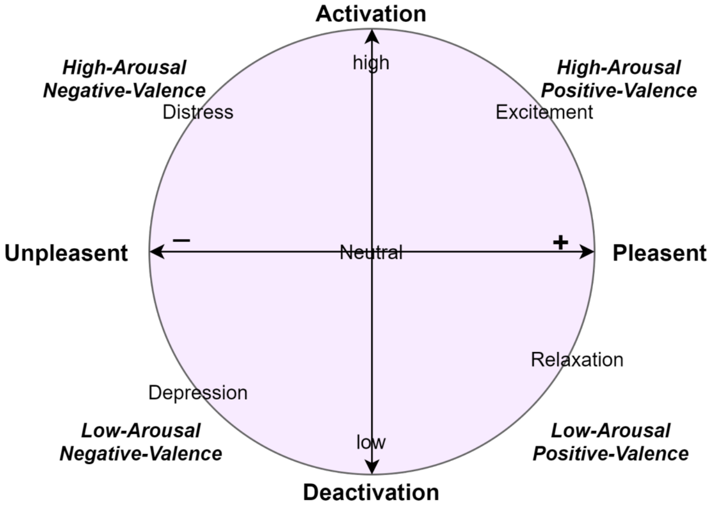

Activation
high
High-Arousal Negative-Valence
Distress
High-Arousal Positive-Valence
Excitement
Unpleasent
−
Neutral
+
Pleasent
Relaxation
Depression
Low-Arousal Negative-Valence
low
Low-Arousal Positive-Valence
Deactivation

Figure 2.4: Depicts the Circumplex's Model of Affect. Taken from [38].

2.1.5 Summary

Table 2.1 summarises the reviewed psychological models, providing valuable insights.

Model	Type	# Primary	# Total	Exact	Relations	Difficulty
Paul Ekman's Model [38]	categorical	6	6	✓	✗	1.
Parrot's Model [38]	categorical	6	100	✓	-	4.
Wheel of Emotions [38] [1]	categorical	8	92	✓	✓	3.
Circumplex's Model [38]	dimensional	5	∞	✗	✓	2.

Table 2.1: Summarises the reviewed psychological models. The column *# Primary* represents the number of primary emotions defined by the model, while *# Total* represents the total number of distinct emotions defined by the model. Column *Exact* highlights whether exact annotations are theoretically possible. Column *Relations* assesses whether the emotions have meaningful relationships defined. Column *Difficulty*, ranked in order of easiest to hardest, assesses the level of difficulty involved in generating annotations.

2.2 Datasets

Emotion classification is mostly approached as either a multi-class or multi-label problem. In multi-class scenarios, a single emotion is considered valid for each sentence, while in multi-label, combinations of multiple co-occurring emotions are valid for each sentence. This led to the creation of many datasets with varying characteristics. These include:

- Number of Samples: datasets vary greatly, with some having only a few hundred samples while others contain millions.
- Quality of Annotations: some datasets, especially large ones, have only weakly labelled samples, while others demonstrate rigorous annotation procedures involving multiple annotators.
- Complexity: some datasets contain only easy-to-classify sentences, while others comprise a mix of sentences featuring sarcasm, irony, and context-dependability in conversations (meaning is influenced by preceding utterances).

This chapter summarises only the most influential datasets in the domain.

2.2.1 SemEval-2018 Task 1: Affect in Tweets

SemEval-2018 Task 1: Affect in Tweets is a multi-task dataset developed for the 2018 Semantic Evaluation competition. It is dedicated to assessing an individual's emotional state based on their tweets. Since its publication in 2018, it has become a de facto standard for evaluating multi-label emotion classification models, garnering more than 750 citations. This chapter paraphrases the original paper [26]. The dataset comprises tweets annotated for the following tasks:

- Emotion intensity regression (*EI-reg*): Given an emotion and a tweet, the system is tasked to predict a number between 0 and 1 describing how intense the emotion is, with 1 being the most intense.
- Emotion intensity ordinal classification (*EI-oc*): similar to the *EI-reg*, this task requires the system to predict one of the following categories: *no emotion* (for tweets not conveying the emotion), *low emotion*, *moderate emotion* and *high emotion* describing the emotion intensity.
- Valence (sentiment) regression (*V-reg*): Given a tweet, the system is tasked to predict a real number between 0 and 1, describing how positive the sentiment of the sentence is, with 1 being the most positive.
- Valence ordinal classification (*V-oc*): unlike *V-reg*, this task involves classifying the sentiment of a tweet into one of these categories: *very negative*, *moderately negative*, *slightly negative*, *neutral or mixed*, *slightly positive*, *moderately positive*, and *very positive.*
- Emotion Classification (*E-c*): Given a tweet, the system is required to output one or more labels (multi-label classification) describing emotions present in the tweet.

Annotation Procedure

In total, the authors collected over 100 million English tweets. They gathered the data by pooling the Twitter API over the period spanning from July to September of 2017 using emotion-specific keywords. Subsequently, they employed a three-step process to get the final dataset:

1. Data Selection and Emotion Pre-Assignment: Specific keywords, emojis and emoticons were used to pre-select tweets rich in emotional content. Additionally, a word-embedding space selection was conducted to ensure the diversity of the dataset. At the end of this stage, the tweets have automatically pre-assigned emotions.

2. Data Annotation: A crowdsourcing platform, Figure Eight (formerly CrowdFlower), was used to annotate the tweets. The annotators were presented with a small set (4 to 8) of tweets having pre-assigned labels. Their task was to identify one tweet that most strongly represented that emotion and one that least represented it.

3. Data Aggregation: The authors compiled these rankings and executed particular post-processing steps to finalize the dataset.

Analysis

The SemEval-2018 Task 1 - Emotion Classification (E-c) dataset is explored through various tables, each providing unique perspectives on its structure and annotations. Table 2.2 presents five examples from the dataset, featuring tweets alongside corresponding annotations. Table 2.3 details how samples are divided into training, validation, and testing splits. Table 2.4 illustrates the distribution of samples annotated with varying co-occurring emotions. Table 2.5 shows the occurrence rates of individual emotions across the splits without considering co-occurrence. Conversely, Table 2.6 focuses on frequently co-occurring emotional combinations. For clarity, we present only those whose relative proportions are greater than 1%. These tables offer a nuanced understanding of the dataset's structure and emotional dynamics. The *SemEval-2018 Task 1: Affect in Tweets* is a rich corpus frequently containing sentences annotated with multiple co-occurring emotions.

Text	**Emotions**
My roommate: it's okay that we can't spell because we have autocorrect. #terrible #firstworldprobs	disgust
About 7 weeks till I can pick up my camera again. Though I think there is a group cemetery shoot in October, I can make it! #photography	joy, optimism
@NHLexpertpicks @usahockey USA was embarrassing to watch. When was the last time you guys won a game..? #horrible #joke	anger, disgust
@onefumi Oh, I see. I've seen so many people mourn the loss that I was surprised to see your tweet. I suppose same old here in SA	surprise, sadness
#smile every morning to a positive head start with your #clients relations	optimism, joy

Table 2.2: Presents five simplified examples from the *SemEval-2018 Task 1: Affect in Tweets* dataset. To see the full dataset's structure, please refer to Chapter 3.2.1, which describes it in detail.

Number of Samples				**Proportions (%)**		
train	validation	test	$\sum$	train	validation	test
6838	886	3259	10983	62.26	8.07	29.67

Table 2.3: Presents how the data in the *SemEval-2018 Task 1: Affect in Tweets* dataset is divided into training, validation, and test splits.

Number of Emotions	**Number of Instances**	**Proportions (%)**
0	293	2.67
1	1481	13.48
2	4491	40.89
3	3459	31.49
4	1073	9.77
5	170	1.55
6	16	0.15
$\sum$	10983	

Table 2.4: Displays the distribution of samples annotated with varying numbers of co-occurring emotions in the *SemEval-2018 Task 1: Affect in Tweets* dataset.

Emotion	**Frequency**			**Proportions (%)**		
	train	validation	test	train	validation	test
disgust	2602	319	1099	38.05	36.00	33.72
anger	2544	315	1101	37.20	35.55	33.78
joy	2477	400	1442	36.22	45.15	44.25
sadness	2008	265	960	29.37	29.91	29.46
optimism	1984	307	1143	29.01	34.65	35.07
fear	1242	121	485	18.16	13.66	14.88
anticipation	978	124	425	14.30	14.00	13.04
pessimism	795	100	375	11.63	11.29	11.51
love	700	132	516	10.24	14.90	15.83
surprise	361	35	170	5.28	3.95	5.22
trust	357	43	153	5.22	4.85	4.69

Table 2.5: Presents the distribution of emotions in the *SemEval-2018 Task 1: Affect in Tweets* dataset, providing frequency and percentages that reflect the proportion of sentences having assigned the emotion.

Emotion	Frequency			Proportions (%)		
	train	validation	test	train	validation	test
anger, disgust	865	107	366	12.65	12.08	11.23
joy, optimism	538	88	356	7.87	9.93	10.92
anger, disgust, sadness	446	56	201	6.52	6.32	6.17
joy, love, optimism	308	69	255	4.50	7.79	7.82
pessimism, sadness	174	18	82	2.54	2.03	2.52
anger, disgust, fear	157	18	49	2.30	2.03	1.50
joy, love	155	20	108	2.27	2.26	3.31
anticipation, joy, optimism	146	24	72	2.14	2.71	2.21
fear, sadness	113	9	46	1.65	1.02	1.41
disgust, sadness	93	14	42	1.36	1.58	1.29
anger, disgust, pessimism, sadness	87	24	61	1.27	2.71	1.87
anger, sadness	73	14	39	1.07	1.58	1.20

Table 2.6: Shows co-occurring emotions in the *SemEval-2018 Task 1: Affect in Tweets* dataset, providing frequency and percentages that reflect the proportion of sentences having assigned the combination. Only emotional tuples whose relative occurrence is >1% on all subsets are presented.

2.2.2 GoEmotions

GoEmotions is a large, carefully curated dataset of Reddit comments annotated for multi-label emotion classification. With 27 distinct emotions (not including combinations), it is one of the most challenging datasets. Similar to SemEval-2018 Task 1: Affect in Tweets, since its publication in 2020, it has also become a commonly used benchmark for evaluating multi-label classification models. This chapter paraphrases the original paper [8] and the official blog post [3].

Annotation Procedure

All the annotations were made by English speakers from India. The annotators were told to assign only emotions that they were reasonably confident about. Additionally, they could label the comment as difficult to classify if needed. The following pipeline was employed to curate the final dataset:

1. Data Collection: Reddit comments originating from subreddits with a minimum of 10,000 comments were used. Data was collected from the start of Reddit in 2005 through January 2019.

2. Data Curation: Reddit is known for its biases towards young male users (demographic bias) and toxic language. These biases were mitigated by using specific filters and performing manual inspections. Additionally, a machine-learning model was employed to weakly label the sentences. Subsequently, authors balanced the dataset, ensuring that no weakly labelled emotions and neither popular subreddits were overrepresented.

3. Data Annotation: Each pre-labelled sentence was initially assigned to three raters. Difficult samples without agreement on at least one label were assigned two additional raters. The final annotations were composed using a majority voting technique.

Analysis

The GoEmotions dataset is explored through various tables, each providing unique perspectives on its structure and annotations. Table 2.7 presents five examples from the dataset, featuring tweets alongside corresponding annotations. Table 2.8 details how samples are divided into training, validation, and testing splits. Table 2.9 illustrates the distribution of samples annotated with varying co-occurring emotions. Table 2.10 shows the occurrence rates of individual emotions across the splits without considering co-occurrence. Conversely, Table 2.11 focuses on frequently co-occurring emotional combinations. For clarity, we present only those whose relative proportions are greater than 0.1%. These tables offer a nuanced understanding of the dataset's structure and emotional dynamics. The GoEmotions dataset experiences a notable imbalance and a lower co-occurring emotion prevalence than the SemEval-2018 Task 1 - Emotion Classification (E-c) dataset. Specifically, GoEmotions has 3 emotion labels represented by fewer than 200 instances each, highlighting significant skewness in label distribution. Additionally, on the one hand, it features only 12% of instances annotated with two co-occurring and a mere 1% of three co-occurring emotions. On the other hand, its finely graded set of labels may potentially compensate for this, adding more depth.

Text	**Emotions**
We need more boards and to create a bit more space for [NAME]. Then we'll be good.	desire, optimism
that is what retardation looks like	anger
Thank you friend	gratitude
I miss them being alive	grief, sadness
I'm going to hold out hope for something minor, even though it looked really bad. Just going to wait for the official news.	optimism

Table 2.7: Presents five simplified examples from the *GoEmotions* dataset. To see the full dataset's structure, please refer to Chapter 3.2.1, where it is described in detail, along with specific pre-processing steps applied to the data.

Number of Samples				**Proportions (%)**		
train	validation	test	$\sum$	train	validation	test
43407	5426	5427	54260	80.00	10.00	10.00

Table 2.8: Presents how the data in the *GoEmotions* dataset is divided into training, validation, and test splits.

Number of Emotions	**Number of Instances**	**Relative**
0	16018	29.52
1	31096	57.31
2	6532	12.04
3	577	1.06
4	36	0.07
5	1	0.00
$\sum$	54260	

Table 2.9: Displays the distribution of samples annotated with varying numbers of co-occurring emotions in the *GoEmotions* dataset.

Emotion	**Frequency**			**Relative (%)**		
	train	validation	test	train	validation	test
admiration	4130	488	504	9.51	8.99	9.29
approval	2939	397	351	6.77	7.32	6.47
gratitude	2662	358	352	6.13	6.60	6.49
annoyance	2470	303	320	5.69	5.58	5.90
amusement	2328	303	264	5.36	5.58	4.86
curiosity	2191	248	284	5.05	4.57	5.23
love	2086	252	238	4.81	4.64	4.39
disapproval	2022	292	267	4.66	5.38	4.92
optimism	1581	209	186	3.64	3.85	3.43
anger	1567	195	198	3.61	3.59	3.65
joy	1452	172	161	3.35	3.17	2.97
confusion	1368	152	153	3.15	2.80	2.82
sadness	1326	143	156	3.05	2.64	2.87
disappointment	1269	163	151	2.92	3.00	2.78
realization	1110	127	145	2.56	2.34	2.67
caring	1087	153	135	2.50	2.82	2.49
surprise	1060	129	141	2.44	2.38	2.60
excitement	853	96	103	1.97	1.77	1.90
disgust	793	97	123	1.83	1.79	2.27
desire	641	77	83	1.48	1.42	1.53
fear	596	90	78	1.37	1.66	1.44
remorse	545	68	56	1.26	1.25	1.03
embarrassment	303	35	37	0.70	0.65	0.68
nervousness	164	21	23	0.38	0.39	0.42
relief	153	18	11	0.35	0.33	0.20
pride	111	15	16	0.26	0.28	0.29
grief	77	13	6	0.18	0.24	0.11

Table 2.10: Presents the distribution of emotions in the *GoEmotions* dataset, providing frequency and percentages that reflect the proportion of sentences having assigned the emotion.

Emotion	**Frequency**			**Proportions (%)**		
	train	validation	test	train	validation	test
anger, annoyance	233	48	26	0.54	0.88	0.48
admiration, gratitude	228	30	26	0.53	0.55	0.48
admiration, approval	199	25	17	0.46	0.46	0.31
confusion, curiosity	186	15	20	0.43	0.28	0.37
admiration, love	156	17	18	0.36	0.31	0.33
annoyance, disapproval	149	21	19	0.34	0.39	0.35
disappointment, sadness	106	12	16	0.24	0.22	0.29
admiration, joy	92	11	11	0.21	0.20	0.20
annoyance, disappointment	80	12	10	0.18	0.22	0.18
admiration, optimism	79	9	7	0.18	0.17	0.13
approval, optimism	73	9	10	0.17	0.17	0.18
amusement, joy	70	11	6	0.16	0.20	0.11
amusement, gratitude	65	7	7	0.15	0.13	0.13
excitement, joy	61	9	8	0.14	0.17	0.15
approval, caring	54	8	11	0.12	0.15	0.20
amusement, approval	51	6	10	0.12	0.11	0.18

Table 2.11: Shows cooccurring emotions in the *GoEmotions* dataset, providing frequency and percentages that reflect the proportion of sentences having assigned the combination. Only emotional tuples whose relative occurrence is >0.1% on all subsets are presented.

2.2.3 XED

XED is a multilingual dataset designated for multi-label emotion classification. Unlike previously introduced datasets, XED respects the well-established Plutchik's taxonomy featuring its eight core emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, trust and neutral emotion. The data comprises annotated movie subtitles obtained from OpenSubtitles [1] (OPUS). The XED presents a great challenge for the models to comprehend since subtitles are typically not self-sufficient and often rely on visual cues from associated movies. This chapter paraphrases the original paper [27].

Annotation Procedure

The dataset was annotated by university students. Expert annotators were also part of the annotation process to ensure quality. The students were instructed to favour quality over quantity and to assign the emotion from the speaker's point of view. Notably, no additional context was presented to the annotators (origin, visual cues, etc.). More than 108 annotators were present during the process, with 63 being active. An annotator who annotated more than 300 sentences is considered active. The following pipeline was employed to curate the final dataset:

1. Collection: The data comes from OPUS, a movie subtitle corpus. The subtitles were collected from movies covering a wide range of genres to cover a broad spectrum of human-spoken interactions. Other than that, the data was selected randomly.

[1] https://opus.nlpl.eu/OpenSubtitles-v2018.php

2. Cleaning: The authors applied a sequence of pre-processing steps to clean up the data, like entity removal of superfluous characters.

3. Annotations: Each sentence was assigned to three annotators. The final annotations were determined using a majority voting technique, where an annotation is assigned to the sentence only if at least two out of three annotators agree on it. Expert annotators check hard-to-annotate (no agreement within annotators) sentences and sometimes include them in the final corpus.

Analysis

The XED dataset is explored through various tables, each providing unique perspectives on its structure and annotations. Table 2.12 presents five examples from the dataset, featuring tweets alongside corresponding annotations. Table 2.13 details how samples are divided into training, validation, and testing splits. Table 2.14 illustrates the distribution of samples annotated with varying co-occurring emotions. Table 2.15 shows the occurrence rates of individual emotions across the splits without considering co-occurrence. Conversely, Table 2.16 focuses on frequently co-occurring emotional combinations. For clarity, we present only those whose relative proportions are greater than 0.1%. These tables offer a nuanced understanding of the dataset's structure and emotional dynamics. In summary, the XED dataset is an interesting and challenging dataset that exhibits a prevalence of labelled instances with multiple labels similar to the *GoEmotions*. However, its broader categorization may not adequately reflect the depth of human emotions. Additionally, the absence of training, validation, and test splits complicates comparisons of models on this dataset. Despite these shortcomings, the dataset is noteworthy for its focus on spoken language within a challenging environment (movies).

Text	**Emotions**
When I say run, run.	anticipation, fear
But, of course, she's lying.	disgust
When I came here, I thought this was gonna be a 30-day stretch, maybe 60.	surprise
You're welcome to have turkey with my husband and me.	trust
A hundred of these are produced every day and sent to sweatshops where urban slaves prepare this poison for our friends, our loved ones, and our children.	anger, disgust

Table 2.12: Presents five simplified examples from the *XED* dataset. To see the full dataset's structure, please refer to Chapter 3.2.1, where it is described in detail, along with specific pre-processing steps applied to the data.

Number of Smaples				**Proportions (%)**		
train	validation	test	$\sum$	train	validation	test
21762	2720	2721	27203	80.00	10.00	10.00

Table 2.13: Presents how the data in the *XED* dataset is divided into training, validation, and test splits. Since the authors did not provide the splits they used in their paper, we had to divide the corpus into training, validation, and testing by ourselves. See Chapter 3.2.1 for further details.

Number of Emotions	**Number of Instances**	**Proportions (%)**
0	9675	35.57
1	13655	50.20
2	3024	11.12
3	710	2.61
4	117	0.43
5	12	0.04
6	9	0.03
7	1	0.00
$\sum$	27203	

Table 2.14: Displays the distribution of samples annotated with varying numbers of co-occurring emotions in the *XED* dataset.

Emotion	**Frequency**			**Proportions (%)**		
	train	validation	test	train	validation	test
anger	3042	369	417	13.98	13.57	15.33
anticipation	2713	336	351	12.47	12.35	12.90
joy	2267	293	273	10.42	10.77	10.03
trust	2181	269	249	10.02	9.89	9.15
fear	1962	231	246	9.02	8.49	9.04
sadness	1960	251	253	9.01	9.23	9.30
surprise	1960	243	239	9.01	8.93	8.78
disgust	1836	240	241	8.44	8.82	8.86

Table 2.15: Presents the distribution of emotions in the *XED* dataset, providing frequency and percentages that reflect the proportion of sentences having assigned the emotion.

Emotion	**Frequency**			**Proportions (%)**		
	train	validation	test	train	validation	test
anger, disgust	314	49	44	1.44	1.80	1.62
joy, trust	279	44	31	1.28	1.62	1.14
anticipation, joy	249	37	30	1.14	1.36	1.10
anticipation, trust	161	18	21	0.74	0.66	0.77
anger, anticipation	161	16	19	0.74	0.59	0.70
sadness, surprise	144	16	17	0.66	0.59	0.62
fear, sadness	112	10	17	0.51	0.37	0.62
anger, fear	104	11	12	0.48	0.40	0.44
anger, sadness	92	9	9	0.42	0.33	0.33
disgust, sadness	92	9	12	0.42	0.33	0.44
anger, surprise	88	12	14	0.40	0.44	0.51
anticipation, joy, trust	88	10	11	0.40	0.37	0.40
anticipation, surprise	83	15	8	0.38	0.55	0.29
joy, surprise	79	12	14	0.36	0.44	0.51
anticipation, fear	78	4	16	0.36	0.15	0.59
fear, surprise	77	14	7	0.35	0.51	0.26
fear, trust	58	14	9	0.27	0.51	0.33
anger, disgust, sadness	44	6	8	0.20	0.22	0.29
sadness, trust	30	4	4	0.14	0.15	0.15
anger, disgust, surprise	26	3	3	0.12	0.11	0.11
anger, disgust, fear	25	4	3	0.11	0.15	0.11
surprise, trust	23	3	3	0.11	0.11	0.11
anticipation, sadness	23	3	3	0.11	0.11	0.11

Table 2.16: Shows cooccurring emotions in the *XED* dataset, providing frequency and percentages that reflect the proportion of sentences having assigned the combination. Only emotional tuples whose relative occurrence is >0.1% on all subsets are presented.

2.2.4 DailyDialog

DailyDialog is a dataset comprising dialogues sourced from websites designed to help English learners practice English in everyday situations, hence its name. Each dialogue involves two speakers and typically lasts for eight turns. These dialogues are split into utterances (single conversational turns) and annotated. Unlike the previously introduced datasets, it recognises the context preceding each utterance, thus minimizing ambiguity. This is important because the same utterance may be interpreted differently based solely on its context. See Table 2.18 for further details. This chapter paraphrases the dataset's paper [21]. The utterances are paired with their corresponding contexts and annotated for the following tasks:

- Act Identification: Given an utterance alongside its context, the model's task is to identify the function of the utterance, which represents its role in the conversation. The possible values are: *inform*, *question*, *directive*, and *commissive*.

- Emotion Classification: Given an utterance alongside its context, the model's task is to identify the emotion hidden in the utterance. The authors adopted the Ekman model, defining six core emotions.

Annotation Procedure

The dataset was annotated by three hired expert annotators. Notably, the annotators reached a high inter-annotator agreement of 78.9%. The full dataset annotation pipeline included the following steps:

1. Collection: The authors crawled several websites designed to help English learners practice conversations. Unfortunately, the exact websites are not mentioned.
2. Cleaning: Initially, the raw data was de-duplicated, and dialogues involving more than two parties were excluded. Subsequently, spelling errors were corrected using an auto-correction package.
3. Annotation: Initially, all three annotators collaboratively annotated 100 randomly selected dialogues, learning to annotate the data correctly. Following this, each annotator labelled the entire dataset independently. The final annotations were determined using a majority voting method.

Analysis

The DailyDialog dataset is examined through Tables 2.18, 2.17, 2.19 and 2.20, each assessing the dataset from a different point of view. Table 2.18 displays five examples from the dataset, each paired with relevant annotation. Table 2.17 shows the distribution of samples across training, validation, and testing splits. Table 2.19 details the prevalence of individual emotions in the dataset. Unlike the previously introduced datasets, we do not present the table for co-occurring emotions. This table is omitted because this dataset categorizes emotions into six broad, non-overlapping categories. In summary, the DailyDialog, with its long and context-dependent sentences, is a unique and challenging dataset. However, it also presents two significant downsides. Positive emotions are much more prevalent, and each utterance can be assigned only one emotion from six possible categories, limiting the depth of recognised emotional expression. Despite these limitations, the dataset remains a valuable and widely used resource for recognising emotions in dialogues.

Number of Samples				**Proportions (%)**		
train	validation	test	$\sum$	train	validation	test
79580	7475	7089	94144	84.53	7.94	7.53

Table 2.17: Presents how the data in the *DialyDialog* dataset is divided into training, validation, and test splits.

Text	**Emotion**
I suggest a walk over to the gym where we can play singsong and meet some of our friends. That's a good idea. I hear Mary and Sally often go there to play ping-pong. Perhaps we can make a foursome with them.	happiness
Oh, my God! I've been cheated! What? What did you buy? It's a brick!	anger
What happened, John? Nothing. Why do you look unhappy? I'm rather disappointed at not being able to see my best friend off.	saddens
Can you do push-ups? Of course, I can. It's a piece of cake! Believe it or not, I can do 30 push-ups a minute. Really? I think that's impossible!	surprise
Are you excited about your trip next month? Yes and no. I can't wait to go to Europe. Well, I have acrophobia.	fear

Table 2.18: Presents five simplified examples from the *DialyDialog* dataset. To see the full dataset's structure, please refer to Chapter 3.2.1, which describes it in detail. The context preceding the classified utterance is italicized.

Number of Emotions	**Number of Instances**	**Proportions (%)**
0	78635	83.53
1	15509	16.47
$\sum$	94144	

Table 2.19: Displays the distribution of samples annotated with varying numbers of co-occurring emotions in the *DialyDialog* dataset and their respective percentages.

Emotion	**Frquency**			**Proportions (%)**		
	train	validation	test	train	validation	test
happiness	9871	598	872	12.40	8.00	12.30
surprise	1483	101	111	1.86	1.35	1.57
sadness	920	79	91	1.16	1.06	1.28
anger	729	65	101	0.92	0.87	1.42
disgust	282	3	46	0.35	0.04	0.65
fear	131	11	15	0.16	0.15	0.21

Table 2.20: Presents the distribution of emotions in the *DialyDialog* dataset, providing frequency and percentages that reflect the proportion of sentences having assigned the emotion.

2.2.5 Summary

The summary of all the reviewed datasets is presented in Table 2.21. Moreover, all the datasets that were reviewed possess certain properties such as a well-defined and rigorous annotation process, and they are of a reasonable size. By „reasonable size“, we mean that they are neither too small to facilitate the training of deep learning models, nor too large to affect the ability to annotate them in a rigorous way.

Dataset	Domain	# emotions	# train	# validation	# test
Daily Dialog [21]	conversations	6 + 1	87 170	8 069	7 740
GoEmotions [8]	Reddit posts	27 + 1	43 410	5 426	5 427
XED [27]	movie subtitles	8 + 1	21 762	2 720	2 721
SemEval [26]	Twitter posts	11 + 1	6 838	886	3 259

Table 2.21: Sumarises all reviewed datasets. # emotions denotes the number of distinct emotions in the annotations, with +1 representing neutral/no emotion. # train, # validation, and # test denote the number of samples used for training, validation, and testing, respectively.

2.3 Metrics

This chapter introduces metrics commonly used for assessing the performance of classification models.

2.3.1 Confusion Matrix

The confusion matrix, as per Equation 2.1, is an important measure that visualizes the degree to which the classes are being confused. By analyzing the matrix, we can compute the number of true positives, true negatives, false positives, and false negatives for each class, enabling the calculation of key metrics like accuracy, precision, recall/sensitivity, specificity, and F1-score. Additionally, it provides details about commonly mistaken classes.

$$M = \begin{pmatrix} M_{11} & M_{12} & \cdots & M_{1n} \\ M_{21} & M_{22} & \cdots & M_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ M_{n1} & M_{n2} & \cdots & M_{nn} \end{pmatrix} \tag{2.1}$$

where:

– M_{ij} indicates the number of instances of class i predicted as class j.

The key components that can be calculated using the matrix are:

- True Positives (TP_i): defined in Equation 2.2, represents the amount of correctly identified instances of class i.

$$TP_i = M_{ii} \tag{2.2}$$

- True Negatives (TN_i): defined in Equation 2.3, represent instances correctly identified as not belonging to class i.

$$TN_i = \sum_{\substack{k=1 \\ k \neq i}}^{C} \sum_{\substack{l=1 \\ l \neq i}}^{C} M_{kl} \tag{2.3}$$

- False Positives (FP_i): defined in Equation 2.4, are instances of other classes incorrectly identified as class i.

$$FP_i = \sum_{\substack{k=1 \\ k \neq i}}^{C} M_{ki} \tag{2.4}$$

- False Negatives (FN_i): defined in Equation 2.5, occur when instances of class i are wrongly classified as other classes.

$$FN_i = \sum_{\substack{l=1 \\ l \neq i}}^{C} M_{il} \tag{2.5}$$

2.3.2 Exact Match Ratio

The Exact Match Ratio metric, as defined in Equation 2.6, measures the proportion of instances where the model's prediction precisely aligns with the ground truth. A prediction is considered correct when it perfectly matches the ground truth as if every class aligns precisely.

$$ExactMatchRatio = \frac{1}{N} \sum_{i=1}^{N} I(\hat{Y}_i = Y_i) \tag{2.6}$$

where:

- N is the total amount of test data,
- I is an indicator function,
- $\hat{Y}_i$ is predicted sequence of labels
- Y_i is the ground truth sequence.

2.3.3 Accuracy

Accuracy is a metric that evaluates the overall performance of a classification model. It is calculated as the ratio of correctly predicted instances to the total instances in the dataset. As per Equation 2.7, the accuracy can be derived from the Confusion matrix. This metric is particularly useful when the datasets are well-balanced, as it considers both True Positives and True Negatives. However, it may not be as indicative of the model's performance in imbalanced scenarios.

$$Accuracy = \frac{\sum_i^C (TP_i + TN_i)}{\sum_i^C (TP_i + TN_i + FN_i + FP_i)} \tag{2.7}$$

2.3.4 Precision

The precision metric, defined by Equation 2.8, measures the proportion of instances that the model correctly identifies as belonging to a particular category i out of all instances that the model predicts as category i. In other words, it measures how many instances predicted as label i are correct. It provides important insights into the reliability of the model's positive predictions. The micro-averaged precision metric, as defined in Equation 2.9, measures overall precision across all classes.

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{2.8}$$

$$Precision_{micro} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i + FP_i)} \tag{2.9}$$

2.3.5 Recall / Sensitivity / True Positive Rate

The recall, also known as the sensitivity or True Positive Rate, defined by Equation 2.10, measures the proportion of instances the model correctly identifies as belonging to a particular category i out of all instances belonging to the class i. In other words, Recall measures how well the model can identify the true instances of a particular category among all the instances of that category. This metric is important for understanding the model's ability to capture all relevant cases of a certain category. The micro-averaged recall metric, as defined in Equation 2.11, measures overall recall across all classes.

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{2.10}$$

$$Recall_{micro} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i + FN_i} \tag{2.11}$$

2.3.6 Specificity / True Negative Rate

The specificity, also known as the true negative rate, defined in Equation 2.12, measures the proportion of true negative instances of class i correctly identified by the model, essentially assessing the model's ability to distinguish non-members of class i accurately. This metric offers a comprehensive view of the model's performance in conjunction with precision and recall. However, when negative instances are predominant, relying on specificity could lead to a skewed assessment. In such cases, there's a risk that the model, despite achieving a high specificity score, might fail to identify the positive class correctly. The micro-averaged specificity metric, as defined in Equation 2.13, measures overall specificity across all classes.

$$Specificity_i = \frac{TN_i}{TN_i + FP_i} \tag{2.12}$$

$$Specificity_{micro} = \frac{\sum_{i=1}^{C} TN_i}{\sum_{i=1}^{C}(TN_i + FP_i)} \tag{2.13}$$

2.3.7 False Negative Rate

The False Negative Rate, also known as the Miss Rate, defined in Equation 2.14, measures the proportion of positive instances that are incorrectly classified as negative by the model. In other words, this metric measures how often the model misses a positive instance. Given its close relationship to recall, Equation 2.15 can determine the False Negative Rate. The micro-averaged False Negative Rate, as defined in Equation 2.16, measures overall specificity across all classes.

$$FNR_i = \frac{FN_i}{FN_i + TP_i} \tag{2.14}$$

$$FNR_i = 1 - Recall_i \tag{2.15}$$

$$FNR_{micro} = \frac{\sum_{i=1}^{C} FN_i}{\sum_{i=1}^{C}(FN_i + TP_i)} \tag{2.16}$$

2.3.8 F1-score

The macro-averaged F1 score, as defined in Equation 2.17, provides an average measure of the model's performance across all **labels**. It assigns equal importance to all labels, even those that are underrepresented. The micro-averaged F1 score, defined in Equation 2.18, provides an average measure of the model's performance across all **instaces**. Hence, overrepresented labels have a proportionally greater impact on this measure.

$$F1_{macro} = \frac{1}{C}\sum_{i=1}^{C} 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{2.17}$$

$$F1_{micro} = 2 \times \frac{Precision_{micro} \times Recall_{micro}}{Precision_{micro} + Recall_{micro}} \tag{2.18}$$

2.3.9 Expected Calibration Error (ECE)

The Expected Calibration Error (ECE) is a score commonly used to assess model calibration. It measures how well a model's predicted probabilities align with actual outcomes. Initially, the predictions are divided into bins based on their predicted probability scores. In each bin, the observed accuracy is compared with the average predicted probability of that bin. For more details refer to Equation 2.19.

$$ECE = \sum_{k=1}^{K} \frac{|B_k|}{N} \times \left| \frac{\sum_{i \in B_k} \mathbf{I}(y_i = \hat{y}_i)}{|B_k|} - \frac{\sum_{i \in B_k} \hat{y}_i}{|B_k|} \right| \tag{2.19}$$

where:

- B_k: represents the set of samples falling within the k^{th} bin.
- N: represents the total number of samples.
- $\mathbf{I}(\cdot)$: denotes an indicator function.
- y_i: denotes the true label for sample i.
- $\hat{y}_i$: denotes predicted probability for sample i.

2.3.10 Brier Score

The Brier Score, as defined in Equation 2.20, is another score used for assessing model calibration. It measures a mean squared error between predicted probabilities and actual outcomes.

$$BS = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}(y_{ij} - \hat{p}_{ij})^2 \tag{2.20}$$

where:

- N: represents the total number of samples.
- C: represents the total number of classes/labels.
- y_i: denotes the true label for sample i.
- $\hat{y}_i$: denotes predicted probability for sample i.

2.4 Approaches in Emotion Recognition

Initially, the task of emotion recognition has been approached using keyword-spotting approaches. These systems inferred the emotions by matching pre-processed input sequences with predefined sets of keywords corresponding to each emotion, optionally considering negations. While straightforward and computationally inexpensive, these methods failed to recognise implicit emotions, as they did not consider sentence semantics. The field then progressed to rule-based approaches. These systems determined emotions by applying linguistic, statistical, and computational rules to pre-processed input sequences. While they considered sentence semantics and yielded better results over keyword-spotting methods, they still struggled with handling complex sentences. [4]

Currently, emotion recognition is approached either through traditional machine learning or deep learning. Approaches based on traditional machine learning have shown significantly better results than prior methods. However, their success heavily depends on the quality of manually extracted features, making their development challenging. On the other hand, deep learning approaches do not require any feature engineering, as the features are automatically extracted during training. Moreover, deep learning methods generally have better semantic understanding, dominating the field of emotion recognition.

2.4.1 Relevant Literature

After reviewing over 50 works, we selected five that applied different approaches to showcase various methodologies. Additionally, we highlighted their strengths and weaknesses.

Recurrent Neural Networks (RNNs)

As shown in Figure 2.5, NTUA-SLP is a model employing a Bidirectional LSTM (Bi-LSTM) with a multi-layer deep self-attention mechanism. This scores first and second place in *E-c* and *EI-reg* in the *SemEval-2018 Task 1: Affect in Tweets* competition, respectively. First, they pre-trained *word2vec* word embeddings using unsupervised learning on a large corpus of tweets (about 550 million tweets). Then, they utilized a transfer learning scheme by pre-training their model on the*SemEval 2017 Task 4A* (sentiment analysis in Twitter) and fine-tuning the *SemEval-2018 Task 1: Affect in Tweets* datasets (multi-task). Like many reviewed methods, this approach faces challenges in generalization outside the Twitter domain due to the specialized training data, which consists only of tweets. Conversely, the authors promoted the interpretability of the results by visualizing the self-attention scores for some sentences. Additionally, the authors efficiently compensated for the lack of labelled data by utilizing an unlabeled dataset of tweets along with two other supervised datasets. [7]

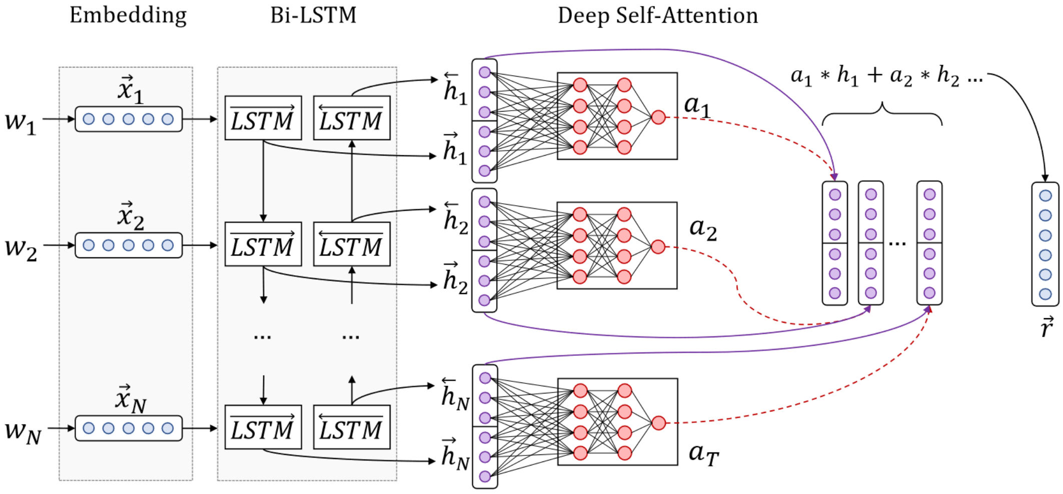


Figure 2.5: Displays the *NTUA-SLP* model. Image taken from [7].

Transformer Encoders (Encoders)

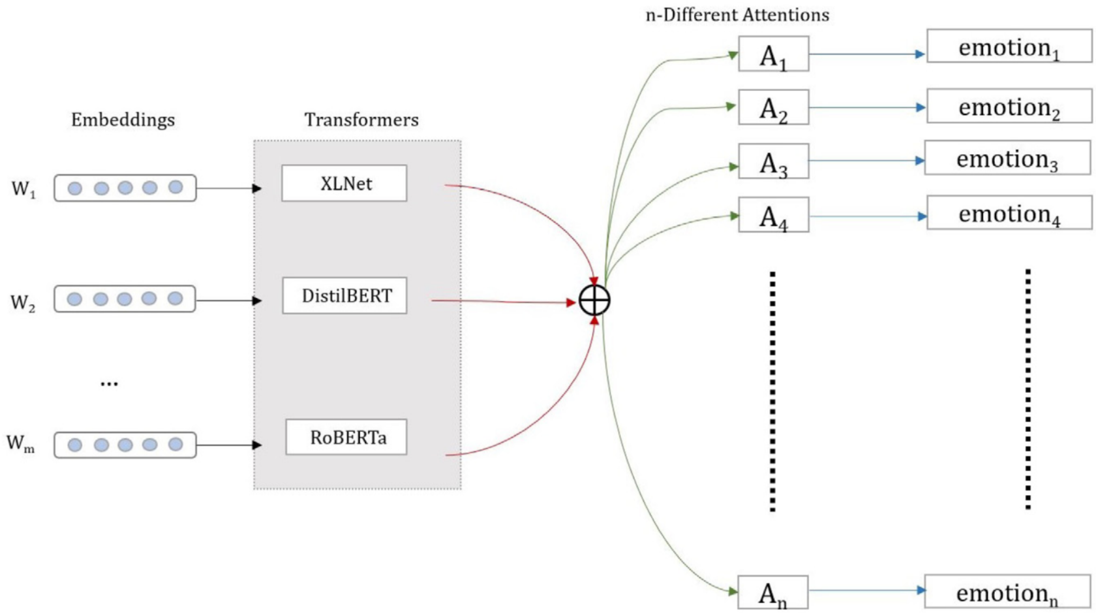


Figure 2.6: Displays the architecture of *RoBERTA-MA*, *DistilBERT-MA* and *XLNet-MA* models. Image taken from [5].

As shown in Figure 2.6, *RoBERTA-MA*, *DistilBERT-MA*, *XLNet-MA* are models based on the Transformer encoder architecture. This work significantly contributed to the field as the *RoBERTA-MA* achieved state-of-the-art performance. The models were trained and evaluated on the *SemEval-2018 Task 1: Affect in Tweets* and *Ren-CECps* (Chinese corpus consisting of blog posts). As part of their work, the authors demonstrated that Transformer-based models outperform the RNN-based approaches significantly. Additionally, they improved the results slightly by introducing additional multiple attention (MA) layers before the output layer. [5]

Transformer Encoder-Decoders (Encoder-Decoders)

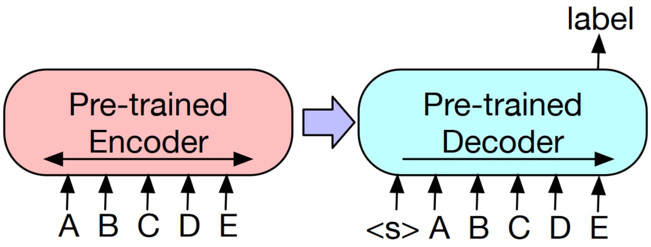


Figure 2.7: Presents a high-level overview of the BART model utilized for classification. The same input sentence is fed to both the encoder and decoder. Image taken from [17].

Madaan et al. focused on improving the performance of sequence-to-sequence models like *BART-base* and *T5-11B* for set generation tasks, including emotion classification. Both of these models are based on encoder-decoder Transformer architecture. Figure 2.7 provides a high-level overview of the BART model and explains how it can be used for classification. The proposed *SETAUG* data augmentation method improves sequence-to-sequence model performance by an average of 20% for multiple set generation tasks. Consequently, the authors demonstrated that sequence-to-sequence models can be effectively utilized for emotion classification.

Graph Attention Networks (GATs)

As shown in Figure 2.8, *UCCA-GAT* and *Dep-GAT* are models based on graph attention networks (*GATs*). They achieved almost state-of-the-art results, only about 0.3% worse than the best-performing models, on the *SemEval-2018 Task 1: Affect in Tweets* dataset. The authors proposed two models: Universal Conceptual Cognitive Annotation Graph Attention Network (*UCCA-GAT*) and Dependency Graph Attention Network (*Dep-GAT*). Their methodology involved extracting semantic/syntactic representation of the sentence, followed by applying a graph attention network. The *UCCA-GAT* model utilizes adjacency and feature matrices extracted from the Universal Conceptual Cognitive Annotation (*UCCA*), while the *Dep-GAT* is based on syntactic representation extracted from dependency trees (Dep). In both cases, they utilized the output of the pre-trained *BERT* model as the feature matrices. *UCCA-GAT* tends to produce better results than *Dep-GAT*. [6]

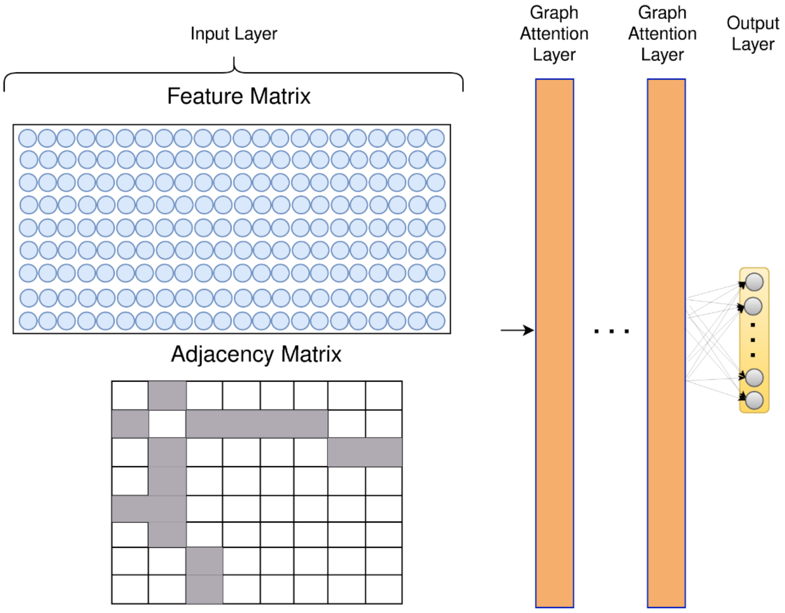
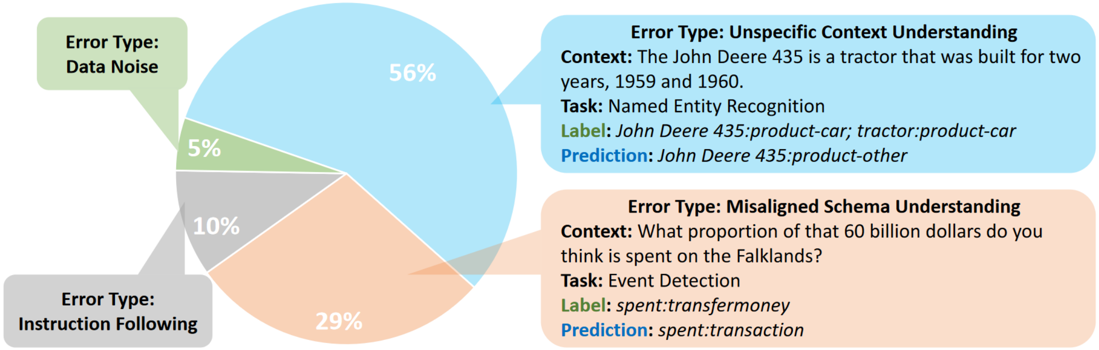


Figure 2.8: Displays the architecture of *UCCA-GAT* and *Dep-GAT* models. Image taken from [6].

Large Langauge Models (LLMs)



Error Type: Unspecific Context Understanding
Context: The John Deere 435 is a tractor that was built for two years, 1959 and 1960.
Task: Named Entity Recognition
Label: *John Deere 435:product-car; tractor:product-car*
Prediction: *John Deere 435:product-other*

Error Type: Misaligned Schema Understanding
Context: What proportion of that 60 billion dollars do you think is spent on the Falklands?
Task: Event Detection
Label: *spent:transfermoney*
Prediction: *spent:transaction*

Figure 2.9: Displays error types with proportions identifying why in-context fails in specification-heavy tasks as uncovered by the authors. Image taken from [29].

Peng et al. published a study on Specification-Heavy tasks for in-context learning (*ICL*). The in-context learning allows large language models (*LLMs*) to be adapted to specific tasks (like emotion classification) without altering their parameters. Instead, the models are given a carefully crafted prompt that specifies the task it needs to perform, relying on pre-existing knowledge of the models and the context provided by the prompt. However, this study identifies Specification-Heavy tasks that require substantial training to master, even for humans. One such task is emotion recognition. Using the *GoEmotions* dataset, the authors demonstrated that current LLMs, including *FLAN-UL2*, *Alpaca*, *Vicuna*, *Chat-*

GPT, *DaVinci*, and *GPT-4*, perform below half the performance of state-of-the-art (SOTA) models. Interestingly, when fine-tuned, these models achieve results comparable to those of the *SOTA* models. Three key reasons, as shown in Figure 2.9, for *ICL* failing were identified: inability to specifically understand context, misalignment with humans, and inadequate long-text understanding. The inability to specifically understand context stands for the model's inability to comprehend all details, including small nuances from the provided context, which leads to inaccurate responses. Misalignment with humans indicates that the model might not interpret the task as humans do, leading to unexpected responses. Inadequate long-text understanding refers to the model's limitations in efficiently utilizing information over longer contexts. Experiments with both aligned and unaligned models showed that human-aligned models performed consistently better in ICL scenarios. However, even smaller, fine-tuned models outperformed larger, aligned models using ICL. [29]

2.4.2 Comparative Summary

This chapter presents a summary Table 2.22 that introduces all research papers (organized by the year publication) and details their methodologies, types of methods employed alongside specific models used and datasets utilized for evaluation (only datasets for which results are reported are presented). This table serves as a quick reference, offering a quick overview of methods used in emotion recognition, highlighting studies that achieve *state-of-the-art* (SOTA) results, maintain dataset integrity (Integrity), and are trained/evaluated across multiple domains (Cross-Domain). By *state-of-the-art* results, we refer to a method achieving the highest score based on the main metric widely recognised by the research community. *For SemEval-2018 Task 1: Affect in Tweets*, *GoEmotions*, and *XED*, this metric is the macro-averaged F1 score. For *DailyDialog*, it is the micro-averaged F1 score.

Year	**Model**	**Approach**	**Data**	**SOTA**	**Cross-Domain**	**Integrity**
2018	NTUA-SLP [7]	RNN: word2vec, Bi-LSTM	SemEval 2017 - 4A, SemEval 2018 1-Ec	✗	✗	✓
2018	TCS Research [25]	Hybrid: GloVe, emoji2vec, Sentiment Neuron, Bi-LSTM	SemEval 2018 1-Ec	✗	✗	✓
2018	PlusEmo2Vec [28]	Hybrid: GloVe, DeepMoji, Bi-LTSM, SVR, logistic regression	SemEval 2018 1-Ec	✗	✗	✓
2019	KET [40]	Hybrid: graph attention, knowledge base, cross-attention	DailyDialog and other emotion dialogues datasets	✗	✗	✓
2019	BERT-large+DK [37]	Multiple: BERT, BiLSTM, CNN	SemEval 2018 1-Ec	✗	✗	✓
2020	EmoGraph: BERT-GAT [36]	GAT: BERT features	IEMOCAP, SemEval 2018 1-Ec (a)	✗	✗	✓
2020	GoEmotions:BERT-base [8]	Encoders: BERT	GoEmotions	✗	✗	✓
2020	XED:BERT [27]	Encoders: BERT	XED	✓	✗	✓
2020	CESTa [34]	Hybrid: LSTM, self-attention, CRF	IEMOCAP, DailyDialog, MELD	✗	✗	✓

2020	COSMIC [9]	Hybrid: RoBERTa, COMET, GRU, attention	IEMOCAP, MELD, EmoryNLP, DailyDialog	✗	✗	✓
2021	Seq2Emo [10]	RNN: BiLSTM with attention	SemEval 2018 1-Ec, GoEmotions	✗	✗	✓
2021	SpanEmo [2]	Encoders: BERT	SemEval 2018 1-Ec	✗	✗	✓
2021	TUCORE-GCN [13]	Hybrid: BERT, multi-headed attention, LSTM, GCN	DialogRE, MELD, EmoryNLP, DailyDialog	✗	✗	✓
2021	SKAIG [18]	Hybrid: RoBERTa, COMET, Graph Transformer	IEMOCAP, MELD, EmoryNLP, DailyDialog	✗	✗	✓
2021	TODKAT [41]	Hybrid: LSTM (encoder-decoder), Transformer (encoder-decoder), attention, SBERT, COMET	IEMOCAP, MELD, EmoryNLP, DailyDialog	✗	✗	✓
2021	KI-Net [35]	Hybrid: Context- and Dependency-Aware Encoders, word embeddings, knowledge base	IEMOCAP, MELD, EmoryNLP, DailyDialog	✗	✗	✓
2022	S + PAGE [22]	Hybrid: Transformer, GAT, CRF	IEMOPCAP, MELD, DailyDialog (a), EmoryNLP	✓ (a)	✗	✗
2022	MADAAN:T5-11B, MADAAN:BART [24]	Encoder-Decoders: BART, T5-11B	GoEmotions, other non-emotion datasets	✗	✗	✓
2022	RoBERTa + FSA [14]	Encoders: RoBERTa	IEMOCAP, MELD, EmoryNLP, DailyDialog	✗	✗	✓
2022	CoMPM [15]	Hybrid: RoBERTa, GRU	IEMOCAP, MELD, EmoryNLP, DailyDialog	✗	✗	✓
2023	RoBERTA-MA, DistilBERT-MA, XLNet-MA [5]	Multiple: GloVe (for RNN), LSTM, Bi-LSTM, XLNet, DistilBERT, RoBERTa	SemEval 2018 1-Ec (a), RenCECps	✓ (a)	✗	✓
2023	UCCA-GAT, Dep-GAT [6]	GAT: BERT features, dependency trees	GoEmotions (a), SemEval 2018 1-Ec (b)	✗	✗	✗ (a) / ✓ (b)

2023	EmoLit:RoBERTa [30]	Encoders: RoBERTa, BERT	EmoLit, Tales, ISEAR, EMOINT, GoEmotions	✗	✓	✓
2023	REMTA [42]	Hybrid: BERT, LSTM, attention	GoEmotions	✗	✗	✓
2023	TTL [16]	Encoders: RoBERTa	11 datasets including GoEmotions, SemEval 2018 1-Ec	✗	✓	✗
2023	DualGATs [39]	GAT: RoBERTa features, GAT, cross-attention	IEMOCAP, MELD, EmoryNLP, DailyDialog	✗	✗	✓
2023	COSMIC + CKCL [9]	RNN: custom RNN based	IEMOCAP, MELD, EmoryNLP, DailyDialog	✗	✗	✓
2023	Mtl-ERC-ES [32]	RNN: BiLSTM, attention	IEMOCAP, MELD, EmoryNLP, DailyDialog	✗	✗	✓
2024	MIP-GAT [11]	GAT: BERT features	GoEmotions (a), CMU-MOSEI	✓ (a)	✗	✓
2024	Li:BART, Li:GPT-3.5-Turbo [19]	LLMs, Encoder-Decoders: GPT-3.5-Turbo, BART (a)	GoEmotions (a), other non-emotion dataset	✓ (a)	✗	✓
2024	CLED [12]	Encoders: RoBERTa	IEMOCAP, MELD, EmoryNLP, DailyDialog	✗	✗	✓
2024	Wang:BERT, Wang:RoBERTa [33]	LLMs, Encoders: RoBERTa, BERT, ChatGPT 4	GoEmotions, CARER	✗	✗	✓

Table 2.22: Summarises all reviewed papers (ordered chronologically by their year of publication). Each paper's approach is outlined, specifying the type of method used, such as Recurrent Neural Networks (RNN), Hybrid (a combination of methods), Graph Attention Networks (GAT), Transformer Encoders, Large Language Models (LLMs), or Encoder-Decoders (sequence-to-sequence transformers). Specific models used are also detailed alongside their types. Additionally, the datasets used for evaluation are displayed for comparison purposes. Finally, each paper is assessed to determine whether it achieves *state-of-the-art* results on any specific dataset (SOTA), was trained across different domains (Cross-Domain), and preserves the original categories of the datasets (Integrity).

2.4.3 Reported Results

The results reported by individual studies are presented in Tables 2.23, 2.24, 2.25 and 2.26. Researchers often evaluate their models using only a few (typically 1 to 3) established metrics for specific datasets. This complicates the comparison since these metrics vary across datasets.

SemEval-2018 Task 1: Affect in Tweets

Models	Year	Accuracy	F1 micro	F1 macro
RoBERTa-MA [5]	2023	**62.4**	**74.2**	**60.3**
UCCA-GAT [6]	2023	61.2	66.1	60.0
DistilBERT-MA [5]	2023	61.3	72.5	58.9
XLNet-MA [5]	2023	60.5	70.4	58.4
Dep-GAT [6]	2023	59.7	63.5	57.8
SpanEmo [2]	2021	-	-	57.8
EmoGraph: BERT-GAT [36]	2020	58.3	69.9	56.9
BERT-large+DK [37]	2019	59.5	71.6	56.3
TCS Research [25]	2018	58.2	69.3	53.0
NTUA-SLP [7]	2018	58.8	70.1	52.8
Seq2Emo [10]	2021	58.67	70.02	51.92
PlusEmo2Vec [28]	2018	57.6	69.2	49.7

Table 2.23: Compares reported results on the *SemEval-2018 Task 1: Affect in Tweets* dataset. In this instance, we compare accuracy, macro-averaged and micro-averaged F1 scores. These are commonly the only results reported in those studies.

GoEmotions

Model	Year	P macro	R macro	F1 macro
Li:BART-base [19]	2024	56.3	**53.9**	**53.8**
MIP-GAT [11]	2024	56.4	51.7	**53.8**
REMTA [42]	2023	52.12	54.08	52.27
EmoLit:RoBERTa [30]	2023	-	-	52
Wang:BERT [33]	2024	**57.27**	49.18	51.83
MADAAN:T5-11B [24]	2022	-	-	50.9
Seq2Emo [10]	2021	-	-	47.28
GoEmotions:BERT-base [8]	2020	40	63	46
Li:GPT-3.5-Turbo [19]	2024	53.1	40.9	42.1
MADAAN:BART [24]	2022	-	-	30

Table 2.24: Compares reported results on the *GoEmotions* dataset using the full taxonomy. P, R and F1 denote macro-averaged precision, recall and F1 scores, respectively.

DailyDialog

Model	Year	F1 macro	F1 micro
S + PAGE [22]	2022	-	**64.18**
CESTa [34]	2020	-	63.12
TUCORE-GCN [13]	2021	-	61.91
DualGATs [39]	2023	-	61.84
RoBERTa + FSA [14]	2022	**55.84**	61.67
CLED [12]	2024	-	61.23
COSMIC + CKCL [9]	2023	53.09	60.96
CoMPM [15]	2022	53.15	60.34
Mtl-ERC-ES [32]	2023	53.06	60.10
SKAIG [18]	2021	51.95	59.75
TODKAT [41]	2021	52.56	58.47
COSMIC [9]	2020	51.05	58.48
KI-Net [35]	2021	-	57.3
CoG-BART [20]	2022	-	56.29
KET [40]	2019	-	53.37

Table 2.25: Compares reported results on the *DailyDialog* dataset. In this instance, we compare macro-averaged and micro-averaged F1 scores. These are commonly the only results reported in those studies.

XED

Models	Year	Accuracy	F1 macro
XED:BERT [27]	2020	**54.4**	**53.6**

Table 2.26: Compares reported results on the *XED* dataset. We found other results beyond those provided as baselines in the dataset paper. Additionally, this dataset comes without predefined training, validation, and test splits. The authors utilized a 5-fold cross-validation with a stratified splitting method of 70:20:10 for training, development, and test data.

2.5 Summary and Limitations of Existing Methods

After a thorough and extensive review of existing research, we have made several observations and identified key areas for improvement. These areas can be grouped into five main categories: datasets, methods, cross-domain efficiency, evaluation and usability. In the subsequent chapters, we delve deeper into each of these categories.

2.5.1 Datasets

There is a wide range of datasets available for emotion classification. However, these datasets often do not adhere to any well-established psychological model, making them difficult to interpret. Moreover, many of these datasets only recognise one emotion at a time, contradicting established psychological theories that suggest multiple emotions can be experienced simultaneously. Furthermore, the quality of annotations varies widely. For instance, datasets like [30] utilize distantly labelled data, suffering from low-quality annotations. Additionally, the lack of clearly described data collection and annotation procedures in some datasets makes it challenging to assess their demographics and the quality of their annotations. Therefore, we have set the following criteria for selecting high-quality multi-label datasets:

- Number of Samples: Datasets must have a reasonable number of samples to ensure annotation quality isn't compromised by excessively large datasets while still being large enough to train deep learning models effectively. We reviewed only those containing thousands or tens of thousands of samples.
- Multi-label Annotations: Datasets must support multi-label annotations, adhering to well-established psychological theories.
- Rigorous annotation process: Datasets must be subject to a comprehensive annotation process, ensuring high quality of the annotations. We included only those where multiple annotators reviewed each sentence, did not include distantly labelled data, and outlined their validation procedure.
- Diversity: Each dataset should come from a different domain, providing comprehensive coverage.

As a result, we have identified four high-quality multi-label datasets that adhere to our criteria. These datasets include:

- GoEmotions dataset [8]: comprising Reddit posts,
- SemEval-2018 Task 1: Affect in Tweets dataset [26]: comprising tweets,
- XED dataset [27]: composed of movie subtitles,
- DailyDialog dataset [21]: composed of conversations between two individuals.

2.5.2 Methods

Studies such as [7], [25], [28], [10], [9], [32] and [37] utilized Recurrent Neural Networks (RNNs) with embeddings trained using various methods and datasets. These studies showed that the quality of the embeddings used significantly influenced the performance of their models. However, these approaches have been consistently outperformed by newer studies, such as [12], [33], [8], [37], [30], [5], [30], [2], [16], [24], [14] and [27], that incorporate transformer-based models. The researchers have already tried many models, including encode-only (BERT, RoBERTa, DistlBERT, XLNet), encode-decoder (BART, T5) and decoder-only (Chat-GPT 3.5, ChatGPT 4, DaVinci) models, with BERT and RoBERTa showing superior performance. A recent trend in adapting large language models to specific problems involves in-context learning when the model is presented with a carefully crafted prompt, removing the need to alter the model's weights. However, as addressed by studies [33], [29], [24] and [19], this approach has proved to produce mediocre results for emotion classification. The latest trend in emotion classification, especially in processing dialogues, leverages the strengths of hybrid models that combine Transformers with Graph Attention Networks (GANs). This approach can be seen in studies such as [40], [36], [22], [13], [18], [6], [11] and [39]. Transformers, such as BERT and RoBERTa, are designed to capture the contextual meaning of words, while Graph Attention Networks (GANs) excel at understanding relationships between data points, which, in this case, represent embeddings of individual utterances (part of the conversations). On the DailyDialog dataset, these models have outperformed approaches based solely on Transformers.

2.5.3 Cross-Domain Efficacy

All the reviewed studies, except [16] and [30], relied exclusively on data from a single domain like Twitter or Reddit. This leads to models that excel within a specific domain but may struggle to generalize across domains. Notably, the study [30] demonstrated the difficulty of the domain transfer. Their model fine-tuned on the EmoLit dataset (a large dataset designated for multi-label emotion classification in literature) demonstrated weak performance when applied to the GoEmotions dataset. The study [16] trained and evaluated their model using multiple datasets, yet they restricted the range of emotions to Ekman's taxonomy (six core emotions). This remapping resulted in a loss of detail, making a comparative assessment with other models that utilize the original labelling of the datasets impossible. In conclusion, achieving strong multi-domain performance while preserving the dataset's original labelling still remains a significant challenge.

2.5.4 Evaluation

Existing research, presented in Tables 2.24, 2.23, 2.25, and 2.26, often utilised only a limited number of evaluation metrics. These studies typically reported fewer than three metrics, with approximately one-third reporting only a single F1 score (either micro or macro-averaged). Calculating metrics at both dataset and label levels is important to comprehensively evaluate the model's performance. Most studies neglect the label level evaluation. Assessing metrics at the dataset level helps us to understand how well the model performs overall. This is crucial, but it's also important to go deeper and see how the model performs for individual labels/emotions. This way, we can identify the model's individual strengths and weaknesses. Additionally, none of the studies evaluate the calibration of their models, which complicates the assessment of how trustworthy the model's outputs are.

2.5.5 Usability

While many papers have been published, most authors have not included the corresponding code and models. Releasing the models publicly would improve practical applications and facilitate better reproducibility.

Chapter 3

Proposed Methodology & Implementation

This chapter outlines our methodology. We start by defining our research goals. Then, we explain how we select and preprocess the data. Next, we discuss the selection and implementation of our model. Finally, we provide a detailed explanation of how we train and evaluate our models.

3.1 Research Questions and Goals

This thesis tackles three pivotal challenges in emotion recognition. For a discussion on the existing limitations of emotion recognition systems, refer to Chapter 2.5. Specifically, we explore:

- **Cross-Domain Efficacy**: How can we build emotion recognition models that generalise well across domains, communication styles and contexts?
- **Preserving Labeling in Cross-Domain Settings**: How can we maintain the integrity of labelling when applying emotion recognition models to datasets with varying emotional categories?
- **Comprhensive Evaluation**: How can we thoroughly assess emotion recognition models to provide in-depth insights into their effectiveness and limitations?

The main objective of this thesis is to design and propose a deep learning-based model that addresses all these three research questions. Additionally, to facilitate usability and promote the reproducibility of our research, we make all our models and code public, available at: https://huggingface.co/vtlustos.

3.2 Proposed Method

We propose a method that successfully addresses all outlined goals. We accomplish this by simultaneously training our models on four datasets systematically chosen to cover a wide range of scenarios. However, unlike the existing approaches that typically remap the original finely detailed categorizations into broader categories, our approach preserves the original labelling of each dataset. Furthermore, we conduct a thorough analysis of our models. Initially, we evaluate the performance individually on each dataset to facilitate comparison with other models. Then, we assess performance at the level of individual emotions to identify the strengths and weaknesses of each model. Subsequently, we evaluate the calibration of our models to verify the outcome's trustworthiness. Finally, we complement our results with a qualitative assessment to empirically validate our models.

3.2.1 Proposed Dataset

To ensure our models generalise well across domains, communication styles and contexts, we composed the *EmoMosaic-dataset* dataset by integrating four previously established datasets, namely the *SemEval-2018 Task 1: Affect in Tweets* (2.2.1), the *GoEmotions* (2.2.2), the *XED* (2.2.3) - only the English subset, and the *DailyDialog* (2.2.4).

The *EmoMosaic-dataset* is analysed through various tables, each providing unique insights. Table 3.1 details the distribution of data across training, validation, and test splits, while Table 3.3 explores the distribution of samples annotated with varying numbers of co-occurring emotions. Table 3.2 displays the occurrence rates of individual emotions across splits, while Table 3.4 focuses on combinations of emotions that frequently co-occur. The EmoMosaic-dataset is a large corpus that combines four challenging datasets, covering a broad spectrum of situations, contexts and domains. Overall, the dataset has a balanced mix of positive and negative emotions, with a significant portion of the sentences, 56.07% to be precise, classified as neutral (having no emotions assigned). Additionally, 33.09% of the sentences are annotated with one emotion and 7.53% with two emotions. Sentences annotated with three or more emotions are less common but still significant, making up less than 4% of the dataset. The EmoMosaic-dataset is imbalanced in terms of emotion frequency, with most emotions occurring reasonably frequently (appearing in more than 500 instances). Additionally, the dataset frequently features co-occurring emotions such as (anger, disgust), (joy, optimism) and (anger, disgust, sadness). However, it also contains rare emotions such as grief, pride, relief, nervousness, and embarrassment, represented by less than 300 instances.

Number of Samples				**Proportions (%)**		
train	validation	test	$\sum$	train	validation	test
151587	16507	18496	186590	81.24	8.85	9.91

Table 3.1: Presents how the data in the *EmoMosaic-dataset* is divided into training, validation, and test splits.

Emotion	**Frequency**			**Proportions (%)**		
	train	validation	test	train	validation	test
happiness	9871	598	872	6.51	3.62	4.71
anger	7882	944	1817	5.20	5.72	9.82
sadness	6214	738	1460	4.10	4.47	7.89
joy	6196	865	1876	4.09	5.24	10.14
disgust	5513	659	1509	3.64	3.99	8.16
surprise	4864	508	661	3.21	3.08	3.57
admiration	4130	488	504	2.72	2.96	2.72
fear	3931	453	824	2.59	2.74	4.46
anticipation	3691	460	776	2.43	2.79	4.20
optimism	3565	516	1329	2.35	3.13	7.19
approval	2939	397	351	1.94	2.41	1.90
love	2786	384	754	1.84	2.33	4.08
gratitude	2662	358	352	1.76	2.17	1.90
trust	2538	312	402	1.67	1.89	2.17
annoyance	2470	303	320	1.63	1.84	1.73
amusement	2328	303	264	1.54	1.84	1.43
curiosity	2191	248	284	1.45	1.50	1.54
disapproval	2022	292	267	1.33	1.77	1.44
confusion	1368	152	153	0.90	0.92	0.83
disappointment	1269	163	151	0.84	0.99	0.82
realization	1110	127	145	0.73	0.77	0.78
caring	1087	153	135	0.72	0.93	0.73
excitement	853	96	103	0.56	0.58	0.56
pessimism	795	100	375	0.52	0.61	2.03
desire	641	77	83	0.42	0.47	0.45
remorse	545	68	56	0.36	0.41	0.30
embarrassment	303	35	37	0.20	0.21	0.20
nervousness	164	21	23	0.11	0.13	0.12
relief	153	18	11	0.10	0.11	0.06
pride	111	15	16	0.07	0.09	0.09
grief	77	13	6	0.05	0.08	0.03

Table 3.2: Presents the distribution of emotions in the *EmoMosaic-dataset* dataset, providing frequency and percentages that reflect the proportion of sentences having assigned the emotion.

Number of Emotions	**Number of Instances**	**Proportions (%)**
0	104621	56.07
1	61741	33.09
2	14047	7.53
3	4746	2.54
4	1226	0.66
5	183	0.10
6	25	0.01
7	1	0.00
$\sum$	186590	

Table 3.3: Displays the distribution of samples annotated with varying numbers of co-occurring emotions in the *EmoMosaic-dataset*.

Emotion	**Frequency**			**Proportions (%)**		
	train	validation	test	train	validation	test
anger,disgust	1219	161	417	0.80	0.98	2.25
joy,optimism	556	91	362	0.37	0.55	1.96
anger,disgust,sadness	491	62	209	0.32	0.38	1.13
anticipation,joy	323	46	61	0.21	0.28	0.33
joy,love,optimism	308	69	255	0.20	0.42	1.38
joy,trust	294	46	34	0.19	0.28	0.18
anger,annoyance	233	48	26	0.15	0.29	0.14
fear,sadness	232	21	65	0.15	0.13	0.35
admiration,gratitude	228	30	26	0.15	0.18	0.14
joy,love	208	23	115	0.14	0.14	0.62
disgust,sadness	192	24	54	0.13	0.15	0.29
anger,anticipation	183	18	26	0.12	0.11	0.14
anger,disgust,fear	182	22	52	0.12	0.13	0.28
anger,sadness	176	24	49	0.12	0.15	0.26
pessimism,sadness	174	18	82	0.11	0.11	0.44
anticipation,trust	169	18	22	0.11	0.11	0.12
sadness,surprise	153	18	23	0.10	0.11	0.12

Table 3.4: Shows co-occurring emotions in the *EmoMosaic-dataset*, providing frequency and percentages that reflect the proportion of sentences having assigned the combination. Only emotional tuples whose relative occurrence is >1% on all subsets are presented.

To enable comparative analysis, we made minimal adjustments to the original datasets, ensuring that the changes were reversible. These transformations typically included converting individual datasets into a consistent format as processed by our model while maintaining the original labelling and other properties. However, since the authors of the *XED* dataset did not provide predefined splits for training, validation, and testing, we had to create these ourselves. As a result, our findings on this dataset are not directly comparable with other models. Subsequent chapters provide in-depth descriptions of our pre-processing steps applied to each individual dataset.

SemEval-2018 Task 1: Affect in Tweets

We use the *subtask5.english* subset of the *sem_eval_2018_task_1* dataset, available at: https://huggingface.co/datasets/sem_eval_2018_task_1. This dataset is offered in three versions: *subtask5.arabic*, *subtask5.english* and *subtask5.english*. Given that our model is monolingual, focusing solely on English, we utilize the *subtask5.english*.

Table 3.5 presents a single-row example from the *subtask5.english* subset. We remove the *ID* column and rename the *Tweet* column to *text* to maintain consistency. The training, validation, and test splits are preserved. Table 3.6 displays the sentence after pre-processing.

ID	Tweet	anger	anticipation	...	trust
2017-En-31527	@enews #breezy deserve it..	False	True	False	False

Table 3.5: Presents a single-row example from the *simplified* subset of the *sem_eval_2018_task_1* dataset. Certain columns/labels are omitted to ease the illustration.

text	anger	anticipation	...	trust
@enews #breezy deserve it.	False	True	False	False

Table 3.6: Displays the pre-processed row, simplified for clarity.

Go Emotions

We use *go_emotions* dataset, more specifically its *simplified* subsets, available at: https://huggingface.co/datasets/go_emotions. This dataset is offered in two versions: *raw* and *simplified*. The *raw* provides annotations from individual annotators, while the *simplifed* version aggregates these annotations and includes predefined training, validation, and testing splits. Moreover, the *simplified* subset is commonly used to compare models.

Table 3.7 displays a single-row example from the *simplified* subset. We remove the *id* column to maintain a consistent format within our dataset, as it is unnecessary for our purposes. Additionally, we convert the *labels*, provided as indexes, into a multi-hot representation. This format assigns a boolean value to each label: *True* if the label is present in the label indexes, and *False* if it is not. The training, validation, and test splits are preserved. Table 3.8 displays the sentence after pre-processing.

text	labels	id
I miss them being alive.	[16, 25]	ee8mzwa

Table 3.7: Presents a single-row example from the *simplified* subset of the *go_emotions* dataset

text	admiration	...	grief	...	sadness	...	neutral
I miss them being alive.	False	False	True	...	True	False	False

Table 3.8: Displays the pre-processed row, simplified for clarity. Certain columns/labels are omitted to ease the illustration. However, in the *emo-mosaic* dataset, every label is always represented by *True* or *False*.

XED

We use the *en_annotated* and *en_neutral* subsets of the *xed_en_fi* dataset, available at: https://huggingface.co/datasets/xed_en_fi. The dataset is provided in four versions: *en_annotated*, *en_neutral*, *fi_annotated* and *fi_neutral*. Since our model operates solely in English, we utilize the English subsets exclusively. The annotated subsets, suffixed *_annotated*, include multi-label annotations, whereas subsets suffixed *_neutral* contain only samples that do not display any emotion.

Tables 3.9 and 3.9 present a single-row example from the *en_annotated* and *en_neutral* subsets, respectively. Initially, we renamed the *sentence* column to *text* to maintain consistency. Moreover, we unified the formats of the *en_annotated* and *en_neutral* subsets by substituting the integer labels (always 0) in the *en_neutral* subset with [0]. Furthermore, as shown in Figure 3.1, we concatenated and split the datasets, creating training, validation and test splits in a 90:10:10 ratio, respectively. Finally, similar to our approach with the GoEmotions dataset, we converted the indexes into a multi-hot representation. Table 3.11 displays the pre-processed rows.

```
# concatenate the datasets
dataset = concatenate_datasets([
    dataset_no_neutral,
    dataset_only_neutral
])

# split the dataset
train_test_split = dataset.train_test_split(
    test_size=0.2,
    shuffle=True,
    seed=42
)
val_test_split = train_test_split['test'].train_test_split(
    test_size=0.5,
    shuffle=True,
    seed=42
)

train_split = train_test_split['train']
validation_split: val_test_split['train'],
test_split: val_test_split['test'],
```

Figure 3.1: Displays the code used for transforming the *en_neutral* and *en_annotated* datasets into the consistent format as used by our *EmoMosaic-dataset*.

sentence	labels
All happened on your watch .	[1, 3]

Table 3.9: Presents a single-row example from the *en_annotated* subset.

sentence	labels
One moment please.	0

Table 3.10: Presents a single-row example from the *en_neutral* subset.

text	neutral	anger	...	disgust	...	trust
All happened on your watch.	False	True	False	True	False	False
One moment, please.	True	False	False	False	False	False

Table 3.11: Displays the pre-processed rows, simplified for clarity.

DailyDialog

This dataset distinguishes itself from others by providing per-utterance annotations, meaning that every dialogue between two individuals is divided into separate utterances, each with its preceding context (previous utterances). We use the *daily_dialog* dataset, available at: `https://huggingface.co/datasets/daily_dialog`.

Reflecting on the dataset's characteristics, we process the dialogues at the level of individual *utterances*, increasing rows in the pre-processed dataset. For each *utterance*, we consider all preceding utterances as its *context*. The *context* and *utterance* are then formatted as *<s>context</s><s>utterance</s>*. It's important to note that the annotation applies specifically to the *utterance* and not to the *context*. Finally, similar to our approaches with the GoEmotions and XED datasets, we converted the label indexes into a multi-hot representation. The training, validation, and test splits are preserved.

Table 3.12 presents a single row from the *daily_dialog* dataset. Table 3.13 displays the sentence after pre-processing.

dialog	emotion
[„How do you like the pizza here ? ", „ Perfect. It really hits the spot. "]	[4, 4]

Table 3.12: Presents a single-row example from the *simplified* subset of the *daily_dialog* dataset. Certain columns/labels are omitted to ease the illustration.

text	...	happiness	...
<s></s><s>How do you like the pizza here?</s>	False	True	False
<s>How do you like the pizza here ?</s><s>Perfect. It really hits the spot.</s>	False	True	False

Table 3.13: Displays the pre-processed row, simplified for clarity. In this case, the transformation of a single dialogue led to the creation of two rows in the pre-processed dataset.

3.2.2 Proposed Model

As demonstrated by [12], [33], [8], [37], [30], [5], [30], [2], [16], [24], [14] and [27], among others, models based on the Transformer architecture have repeatedly outperformed other methods for emotion recognition. Additionally, the *RoBERTa* model has consistently demonstrated strong performance across various tasks, including emotion recognitions, as highlighted by [5], [18], [14] and [12]. Therefore, our proposed *EmoMosaic-base* and *EmoMosaic-large* models are based on the *RoBERTa-base* and *RoBERTa-large*, respectively.

Input

We condition our model on the sentence being classified and, if available, on the preceding context (although some datasets only contain the sentence). The model requires inputs in the following format:

<s>context</s><s>sentence</s>

Including the *{sentence}* part is mandatory as it represents the primary text for classification. Including the *{context}* is recommended, although optional, to clarify potential ambiguities in the *{sentence}*, improve its understanding and enable the model to process dialogues. Table 3.14 showcases how providing the context might influence the expected outputs (detected emotions).

Text	**Expected Output**
<s></s><s>You spend most of your time at home playing games.</s>	annoyance, anger, disgust
<s>I like chill people who enjoy playing games at home.</s><s>You spend most of the time at home playing games.</s>	approval, neutral

Table 3.14: Showing the importance of including context and the ability to process dialogues.

After being transformed into the described format, each sentence undergoes tokenization using a suitable tokenizer. Additionally, ground truth vectors are created to indicate which emotions apply to the sentence. Equation 3.1 describes how we create the ground truth vectors.

$$Y^{B\times C} = [y_0, y_1, \ldots, y_C] \tag{3.1}$$

where:

- $Y^{B\times C}$ is the ground truth vector, with B and C representing the batch size and the number of distinct labels.
- y_i corresponds to a specific label (emotion), indicating the label's applicability to a given sentence, where:
  - a value of -1 indicates that the label was not included in the dataset's original labelling.

- ∗ a value of 0 indicates that the label, although present in the dataset's original labelling, is not applicable to the given sentence.
- ∗ a value of 1 indicates that the label is present in the dataset's original labelling and is applicable to the given sentence.

Output

Equation 3.2 defines the output generated by the model for any given input. The model generates logits, representing the relevance of each label to these inputs.

$$Y_{\text{logits}}^{B \times C} = \text{model}(T^{B \times S}) \tag{3.2}$$

where:

- $Y_{\text{logits}}^{B \times C}$ represents the logits produced by the model, with B and C representing the batch size and the number of distinct classes. The logits are unnormalized scores used to determine each input sequence's label probabilities.
- model stands for the *RobertaForSequenceClassification* model that we use for the classification.
- $T^{B \times S}$ represents the tokenized input sentences, with B and S indicating the batch size and maximum sequence length.

Loss Function

Given the highly imbalanced nature of the *EmoMosaic-dataset*, we utilised the Focal loss function, as defined in Equation 3.3, that allows us to control the relative importance of positive and negative classes and assign different weights to well-classified examples. [23]

As a way to accommodate the categories of individual datasets, we propose the following modifications to the loss computation, which are shown in Equations 3.4 and 3.5. We start by creating a mask that masks any labels absent in the original dataset. Following this, a loss value is calculated for each logit the model produces using the Focal loss. Finally, we apply the mask to these values and reduce them using the mean reduction method.

$$\mathcal{L}^{B \times C} = -\alpha_t (1 - P_t)^{\gamma} \log(P_t) \tag{3.3}$$

where:

- $L^{B \times C}$ represents a matrix of unreduced loss function values. Here, B denotes the batch size, and C represents the number of distinct classes.
- $P^{B \times C} = \sigma(Y_{\text{logits}})$ is a tensor of predicted probabilities, with σ representing the sigmoid function.
- $P_t^{B \times C} = P \cdot Y + (1 - P) \cdot (1 - Y)$ is a tensor of the target class probabilities.
- $\alpha_t^{B \times C} = \alpha \cdot Y + (1 - \alpha) \cdot (1 - Y)$ is the balancing factor, with $\alpha \in < 0, 1 >$ and $\alpha > 0.5$ giving more importance to the positive class.
- γ is the focusing parameter that reduces the relative loss for well-classified samples. The expected range for this parameter is anywhere between 1 and 4.

$$M_{b,c}^{B \times C} = \begin{cases} 0 & \text{if } Y_{b,c} = -1, \\ 1 & \text{otherwise} \end{cases} \tag{3.4}$$

where:

- $M_{b,c}^{B\times C}$ is a mask applied to the loss function, designed to exclude labels that are absent in the original dataset labelling due to the uncertainty of their applicability. B and C represent the batch size and the number of distinct classes. The mask takes on the following values:
  - a value of 0 is assigned based on the ground truth to positions corresponding to the absent labels, omitting them from the calculation.
  - a value of 1 is assigned otherwise.

$$l = \frac{\sum_b^B \sum_c^C L_{b,c} \cdot M_{b,c}}{\sum_b^B \sum_c^C M_{b,c}} \tag{3.5}$$

where:

- l represents the mean reduced scalar loss function value, calculated by applying a mask to the unreduced loss values and then performing a mean reduction.
- $L_{b,c}$ represents the unreduced loss function values.
- $M_{b,c}$ represents the mask applied to exclude the absent labels from the loss calculation.

Architecture

We based our models, the *EmoMosiac-base* and *EmoMosiac-large*, on the *RobertaForSequenceClassification*. *RobertaForSequenceClassification* is a model from the Hugging Face's *transformers* library, available at: https://huggingface.co/FacebookAI/roberta-base and https://huggingface.co/FacebookAI/roberta-large. It utilizes pre-trained embeddings (*RobertaEmbeddings*) and pre-trained encoder backbone (*RobertaModel*) coupled with a newly initialized classification head (*RobertaClassificationHead*). We use the *base* and *large* variants of the model. The key distinction between these variants is the number of hidden layers and the number of features. Specifically, the base model comprises 12 hidden layers (*<NL> = 12*) and a hidden size of 768 (*<HS> = 768*), whereas the large model consists of 24 hidden layers (*<NL> = 24*) with a hidden size of 1024 (*<HS> = 1024*). Figure 3.2 illustrates the architecture of each variant, while Table 3.15 compares their computational requirements.

Specification	**RoBERTa-base**	**RoBERTa-large**
Number of Parameters	125M	355M
Number of Hidden Layers	12	24
Hidden Size	768	1024
Intermediate Size	3072	4096
Number of Attention Heads	12	16
Dropout Rate	0.1	0.1
Max Sequence Length	512 tokens	512 tokens
Training Data	160GB of text	160GB of text

Table 3.15: Compares the *RoBERTa-base* and *RoBERTa-large* models.

```
RobertaForSequenceClassification(
  (roberta): RobertaModel(
    (embeddings): RobertaEmbeddings(
      (word_embeddings): Embedding(50265, <HS>, padding_idx=1)
      (position_embeddings): Embedding(514, <HS>, padding_idx=1)
      (token_type_embeddings): Embedding(1, <HS>)
      (LayerNorm): LayerNorm((<HS>,), eps=1e-05, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False)
    )
    (encoder): RobertaEncoder(
      (layer): ModuleList(
        (0-<NL>-1): <NL> x RobertaLayer(
          (attention): RobertaAttention(
            (self): RobertaSelfAttention(
              (query): Linear(in_features=<HS>, out_features=<HS>, bias=True)
              (key): Linear(in_features=<HS>, out_features=<HS>, bias=True)
              (value): Linear(in_features=<HS>, out_features=<HS>, bias=True)
              (dropout): Dropout(p=0.1, inplace=False)
            )
            (output): RobertaSelfOutput(
              (dense): Linear(in_features=<HS>, out_features=<HS>, bias=True)
              (LayerNorm): LayerNorm((<HS>,), eps=1e-05, elementwise_affine=True)
              (dropout): Dropout(p=0.1, inplace=False)
            )
          )
          (intermediate): RobertaIntermediate(
            (dense): Linear(in_features=<HS>, out_features=<HS>*4, bias=True)
            (intermediate_act_fn): GELUActivation()
          )
          (output): RobertaOutput(
            (dense): Linear(in_features=<HS>*4, out_features=<HS>, bias=True)
            (LayerNorm): LayerNorm((<HS>,), eps=1e-05, elementwise_affine=True)
            (dropout): Dropout(p=0.1, inplace=False)
          )
        )
      )
    )
  )
  (classifier): RobertaClassificationHead(
    (dense): Linear(in_features=<HS>, out_features=<HS>, bias=True)
    (dropout): Dropout(p=0.1, inplace=False)
    (out_proj): Linear(in_features=<HS>, out_features=32, bias=True)
  )
)
```

Figure 3.2: Shows the architecture of the *RobertaForSequenceClassification* model. *<NL>* and *<HS>* represent number of layers and hidden size, respectively.

The embeddings module (*RobertaEmbeddings*) transforms input tokens into dense vectors. The word embeddings layer initially converts each token into a high-dimensional vector, capturing its semantics. Subsequently, each vector is combined with a corresponding position embedding, aiding the model in recognising the order of tokens in the input sequence. Lastly, token-type embeddings can be used to differentiate between types of sequences. Additionally, layer normalization and dropout layers are used to stabilise the training and prevent overfitting.

The encoder (*RobertaEncoder*) is composed of multiple identical encoder layers (*RobertaLayer*). Each encoder layer contains a multi-headed self-attention layer, several dense layers, layer normalization layers, dropout layers, and activation functions. The self-attention mechanism allows the model to recognize relationships between intermediate representations. Dense layers transform and extract features. Activation functions in these layers introduce a non-linearity, enabling the model to recognise complex patterns.

The classification head (*RobertaClassificationHead*) uses additional dense layers, coupled with layer normalization and dropout layers, to transform the extracted features into scores/logits corresponding to each class/label.

3.2.3 Model Training Methodology

Similar to previous studies, we started by selecting the main metric to be tracked during experimentation. We chose the macro-averaged F1 score because it gives the same importance to all classes, which is desirable for imbalanced datasets such as the *EmoMosaic-dataset*.

Considering the high impact of hyperparameters on the model's efficacy, we implemented a systematic hyperparameter selection strategy. We used the *Ray Tune* library coupled with *Optuna* searching algorithm and *Weights and Biases* platform to conduct and track our experiments. Since an exhaustive search (trying every combination) is not feasible due to computational requirements, we used the *Optuna* algorithm to find the best hyperparameters efficiently. We started by defining the hyperparameter space, which describes all possible values for each hyperparameter. Figure 3.3 shows an example of such a definition, while Figure introduces 3.4 a complete configuration of our *EmoMosiac-base* model with all the hyperparameters set to specific values.

```
from ray import tune
...
'steps': tune.choice([
   3000, 3500, 4000, 4500, 5500
]),
...
```

Figure 3.3: Shows how we specified all possible values for the number of training steps.

Then, we utilized the *Optuna* algorithm that utilizes a Gaussian process to model the relationship between individual hyperparameters and the objective function. We utilized the maximalization of the macro-averaged F1 score as the objective function. As *Optuna* conducts trials, its internal representation of the hyperparameter space improves, leading to better-performing models.

```
CONFIG = {
    # project-specific parameters
    ’project’: ’emo-mosaic’,

    # data-specific parameters
    ’max_seq_length’: 192,
    ’dataset’: {
        ’name’: ’vtlustos/emo-mosaic-v2-192’,
        ’subsets’: [
            ’go_emotions’, ’sem_eval_2018_task_1’,
            ’xed’, ’daily_dialog’
        ]
    },

    # architecture-specific parameters
    ’model_name’ : ’FacebookAI/roberta-base’,

    # training-specific parameters
    ’loss’: {
        ’name’: ’focal_loss’,
        ’args’: {
            ’alpha’: 0.75,
            ’gamma’: 1.75
        }
    },
    ’batch_size’: 128,
    ’accumulate_grad_batches’: 1,
    ’lr’: 1e-4,
    ’steps’: 3000,
    ’warmup_steps’: 1500,
    ’weight_decay’: 0.01,
    ’betas’ : (0.9, 0.98),
    ’val_check_interval’: 1000,

    # resources-specific parameters
    ’tuner’: ’optuna’,
    ’num_samples’: 100,
    ’resources’: {
        ’cpu’: 5,
        ’gpu’: 1
    },
    ’num_workers’: 10
}
```

Figure 3.4: Displays the configuration of the *EmoMosiac-base* model having all the hyperparameters set to specific values.

Our framework supports the following hyper-parameters:

- *project*: specifies the project name, used for grouping all the experiments governed by this configuration. Experiment names are generated automatically by the *Ray Tune*.
- *max_seq_length*: specifies the maximum sequence length. All items in the batch will be padded to this length. The limit is 512 tokens for the *RoBERTa* by default.
- *dataset*: specifies which and how the dataset will be used for training, validation and testing.
  - *name*: specifies the name of the dataset, either *vtlustos/EmoMosaic-dataset* or *vtlustos/EmoMosaic-dataset-192*.
  - *subsets*: specifies wich subsets will be used. Since these datasets are formed by combining multiple datasets together, there is an option to use specific portions.
- *model_name*: specifies the name of the model used for the classification. Must represent a *RobertaForSequenceClassification* model. We used *FacebookAI/roberta-base* and *FacebookAI/roberta-large*.
- *loss*: specifies settings of the loss function utilized during training. We used the Focal loss as described in Chapter 3.2.2.
  - *alpha*: is the balancing factor, with $\alpha \in< 0, 1 >$ and $\alpha > 0.5$ giving more importance to the positive class.
  - *gamma*: is the focusing parameter that reduces the relative loss for well-classified samples. The expected range for this parameter is anywhere between 1 and 4.
- *batch_size*: specifies how many samples will be in a batch.
- *accumulate_grad_batches*: specifies the number of steps before propagating gradients. This enables the use of larger batches (split across steps) on machines with limited VRAM.
- *lr*: specifies the peak learning rate.
- *steps*: specifies the number of steps before terminating the training.
- *warmup_steps*: specifies the number of steps over which the learning rate gradually increases up to the *lr*.
- *betas*: specifies the first and second statistical moments (β_1, β_2) utilized by the *AdamW* scheduler.
- *weight_decay*: specifies the regularization factor used to adjust the model's weights during training.
- *val_check_interval*: specifies the frequency of model validation and checkpoint creation.
- *tuner*: specifies the name of the *Ray Tune* search algorithm that will be used for searching the hyper-parameter space.

- *num_samples*: specifies the maximum number of varying configurations drawn from the hyper-parameter space.
- *resources*: specifies how many resources should be allocated for each trial. A trial will start only if the requested amount of resources are available.
- *num_workers*: specifies the number of data-loader workers.

3.2.4 Model Evaluation Methodology

The last step of our methodology involved a thorough evaluation of our models. We assessed how well the models perform on different datasets (dataset-level analysis). Consequently, we selected the best-performing *base* and *large* models and conducted an in-depth analysis at the level of individual emotions (label-level analysis), identifying their strengths and weaknesses. Additionally, we measured the calibration of those models to ensure prediction trustworthiness. Finally, we empirically evaluate our models (qualitative analysis) to demonstrate their real-world applicability under various scenarios.

Quantitative Analysis

We utilized a comprehensive set of 484 metrics to assess the model's effectiveness, evaluating performance at both the dataset and label levels.

Metric	Averaging	Explanation
Accuracy	-	Measures the overall correctness of the model's predictions.
Exact Match	-	Measures the proportion of samples with all labels predicted correctly.
Precision	macro	Measures the average correctness of the model's positive predictions for each class, treating all classes equally.
Recall	macro	Measures the average ability of the model to identify all relevant instances of a class, treating all classes equally.
Specificity	macro	Measures the model's average ability to correctly identify irrelevant instances (true negatives) of a class, treating all classes equally.
F1	macro	Measures, the average harmonic mean of precision and recall, treating all classes equally.
Precision	micro	Measures the overall correctness of the model's positive predictions, considering all positive predictions.
Recall	micro	Measures the model's overall ability to identify relevant instances, considering all relevant instances.
Specificity	micro	Measures the model's overall ability to identify irrelevant instances (true negatives), considering all non-relevant instances.
F1	micro	Measures the overall harmonic mean of precision and recall.

Table 3.16: Details the metrics used for the dataset-level analysis.

Evaluating metrics at the dataset level allows us to compare our results with other methods. Moreover, by averaging those scores, we can assess the cross-domain performance of our models. Table 3.16 shows all used dataset-level metrics. Due to the highly imbalanced nature of the *EmoMosiac-dataset*, it is necessary to use both micro-averaged and macro-averaged metrics. Macro-averaged scores give equal importance to all labels, even those that are underrepresented. On the other hand, micro-averaged scores assign the same importance to all instances. Hence, overrepresented categories have a greater impact on the micro-averaged measures. You can learn more about these metrics in Chapter 2.3.

Metric	**Explanation**
Accuracy	Measures the overall correctness of each label.
Precision	Measures the correctness of the positive predictions for each label.
Recall	Measures the model's ability to identify all relevant instances of each label.
F1	Measures the harmonic mean of precision and recall for each label.
ECE	Measures the expected calibration error, quantifying how closely predicted probabilities of positive instances match empirical probabilities.
Brier	Measures the mean squared error between predicted probabilities and actual outcomes, measuring how accurate the predictions are.

Table 3.17: Details the metrics used for the label-level analysis.

Evaluating metrics at the level of individual labels allows for a detailed analysis, helping us identify areas where our models excel and struggle. Additionally, we analyse the calibration of our models, assessing how well the predicted probabilities match the actual outcomes. Table 3.17 shows all label-level metrics (calculated for each label).

Qualitative Analysis

To demonstrate our model's versatility and adaptability, we designed four targeted test suites to evaluate its performance in diverse scenarios:

1. *Trivial Sentences*: This suite comprises simple, easy-to-classify sentences like „I can't believe you lied to me!". This suite is designed to assess the model's basic abilities. Refer to Table 4.12 for more details.

2. *Complex Sentences*: In this suite, we challenge the model by introducing sentences characterized by sarcasm, irony, and subtle emotional cues. This suite assesses the model's ability to classify challenging sentences like „Oh great, another day in paradise working with this ancient computer." correctly. Refer to Table 4.14 for more details.

3. *Cross-Domain Efficiancy*: This suite is designed to assess the effectiveness of the model in different communication styles and contexts (cross-domain efficacy), such as literature, social media platforms, news articles, and more. For instance, a sample sentence from legal documents could be, „The parties involved express their satisfaction with the resolved settlement terms." Refer to Table 4.16 for more details.

4. *Dialogues*: This suite assesses whether the model can efficiently process dialogues by considering the context of previous interactions. As shown in Table 4.18, we intentionally constructed pairs of examples with the same sentence to be classified but in different contexts, showcasing how the meaning changes based on context.

3.2.5 Usability and Reproducibility

To promote the usability and reproducibility of our research, we make our code and all models publicly available. Figure 3.5 provides an example of how to use the proposed models.

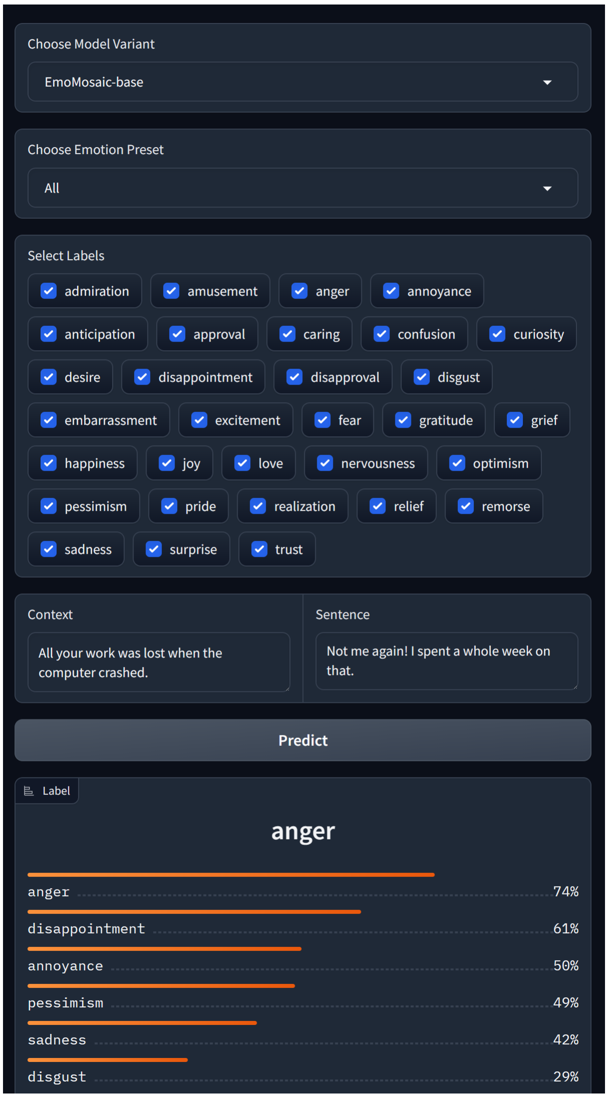
```
import torch
from transformers import RobertaTokenizer
from transformers import RobertaForSequenceClassification

# 1. initialize the model
tokenizer = RobertaTokenizer.from_pretrained(
    "vtlustos/EmoMosaic-base"
)
model = RobertaForSequenceClassification.from_pretrained(
    "vtlustos/EmoMosaic-base"
).to(’cuda:0’)

# 2. tokenize the sentences
tokens = tokenizer(
    [
        "All your work was lost when the computer crashed.</s><s>Oh
        my god. I spent a whole week on that."
    ],
    truncation=True,
    padding=True,
    return_tensors = "pt"
)

# 3. make the prediction
with torch.no_grad():
    logits = model(
        tokens["input_ids"].to(’cuda:0’),
        tokens["attention_mask"].to(’cuda:0’)
    ).logits

# 4. convert to probabilities
preds = torch.sigmoid(logits)

print(preds)
```

Figure 3.5: Shows an example of how to use the proposed models. Users should provide samples in the following format *context</s><s>sentence*. The *context* is optional and represents sentences preceding the sentence to be classified, while *sentence* refers to the actual sentence undergoing classification. This example demonstrates how to use *the EmoMosaic-base* model. If you prefer to use its larger counterpart, replace *vtlustos/EmoMosaic-base* with *vtlustos/EmoMosaic-large*.

The models can be downloaded from:

- *EmoMosaic-base*: https://huggingface.co/vtlustos/EmoMosaic-base,
- *EmoMosaic-large*: https://huggingface.co/vtlustos/EmoMosaic-large.

We have also developed a *Gradio* application, as shown in Figure 3.6, and deployed it on the *Hugging Face Spaces* platform. This allows anyone to experiment with the models easily without requiring any technical skills or setup. Note that the initial request may take up to 30 seconds because the application needs to download the necessary files. Additionally, the link represents a Git repository that houses all the code used in development, along with the *Gradio* application.



Choose Model Variant

EmoMosaic-base

Choose Emotion Preset

All

Select Labels

☑ admiration ☑ amusement ☑ anger ☑ annoyance
☑ anticipation ☑ approval ☑ caring ☑ confusion ☑ curiosity
☑ desire ☑ disappointment ☑ disapproval ☑ disgust
☑ embarrassment ☑ excitement ☑ fear ☑ gratitude ☑ grief
☑ happiness ☑ joy ☑ love ☑ nervousness ☑ optimism
☑ pessimism ☑ pride ☑ realization ☑ relief ☑ remorse
☑ sadness ☑ surprise ☑ trust

Context

All your work was lost when the computer crashed.

Sentence

Not me again! I spent a whole week on that.

anger

```
anger ........................................ 74%
disappointment ............................... 61%
annoyance .................................... 50%
pessimism .................................... 49%
sadness ...................................... 42%
disgust ...................................... 29%
```

Figure 3.6: shows the graphical user interface of the deployed *Gradio* application. The application allows users to select between two models: *EmoMosaic-base* and *EmoMosaic-large*. Additionally, users have the option to select emotions they want to assess, either by manual selection or by using presets. Users are prompted to input the sentence to be classified into the *Sentence* text area to use the application. The *Context* is optional. After clicking on the *Predict* button, the results will be displayed as soon as they are computed.

Chapter 4

Results and Discussion

This chapter summarises all our experiments, analyses the outcomes, and compares our models with recent state-of-the-art models in the field. The discussion begins with model training and hyperparameter optimization, followed by an in-depth examination of how the models performed on different datasets, comparing them to recent state-of-the-art models. Furthermore, we assess the performance at the level of individual emotions, providing a nuanced understanding. In addition, we assess the calibration of our models to ensure prediction trustworthiness. Finally, we empirically evaluate our models across various use cases to support our claims and demonstrate their real-world applicability.

4.1 Hyperparameter Tunning

To maximize the effectiveness of our model, we employed a meticulous process for selecting hyper-parameters, as detailed in Chapter 3.2.3.The table 4.1 presents the hyperparameters used for our final models. We started with a model with a slightly different architecture than our final model but also utilised the RoBERTa encoder. Therefore, we managed to derive several key observations from it that were applied to our final model. These include:

- Batch Size: We experimented with batch sizes of 32, 64, 128, and 256, among which a batch size of 128 achieved the highest macro-average F1 score. Thus, we used a batch size of 128 samples for further experiments.
- Weight Decay: We experimented with weight decay rates of 0.1, 0.01 and 0.001. We found that a rate of 0.1 led to severe underfitting (achieving macro-averaged F1 scores around 0.35), while 0.001 led to slightly worse results than 0.01. Therefore, we used a weight decay rate of 0.01 in all subsequent experiments.
- Moments: The AdamW optimizer's first and second moments set at ($\beta_1 = 0.9$,$\beta_1 = 0.98$) demonstrated superior performance compared to ($\beta_1 = 0.9$,$\beta_1 = 0.999$). Hence, we selected ($\beta_1 = 0.9$, $\beta_2 = 0.98$) for subsequent experiments.

Based on those observations, we trained an additional 67 models with varying configurations, initially utilizing a wide hyperparameter space. Due to the extensive training time required, we frequently halted the automated search to manually refine and narrow the hyper-parameter space. This iterative adjustment process enabled us to guide the optimization process more efficiently. To make this thesis concise, we decided only to present the configurations and, later in this chapter, the detailed analysis of our two top-performing models while highlighting the key observations from all our experiments. We made the following observations:

- Loss Function: We experimented with two loss functions: Binary Cross Entropy (BCE) and Focal Loss (FL). BCE achieved slightly lower F1 scores compared to Focal Loss. However, the high trade-off between precision and recall was a more significant issue with BCE. BCE typically resulted in a high average recall (around 0.65) and a much lower precision (around 0.35). Therefore, we opted for Focal Loss. Focal Loss offers more control via its α and γ parameters, with α allowing the adjustment of the importance of the underrepresented class and γ assigning more importance to hard-to-classify samples. We explored the following ranges of $\alpha \in (0.15 - 0.85)$ and $\gamma \in (1.5 - 3.0)$, selecting the $\alpha = 0.75$ and $\gamma = 1.75$. This approach resulted in more balanced outcomes, with precision reaching around 0.51 and recall reaching around 0.58.

- Learning Rate: The learning rate is another critical hyperparameter. We tested learning rates $\lambda \in < 5e^{-5}, 3e^{-4} >$ for the base models and $\lambda \in < 3e^{-5}, 7e^{-5} >$ for the large models. For the base model, the learning rate that led to the best results was $\lambda = 1e^{-4}$, while rates above $3e^{-4}$ caused divergence. The large model performed best when the rate was set at $\lambda = 5e^{-5}$.

- Steps: The number of steps is another hyperparameter significantly influencing the model's overall performance. As we employ a linear scheduler with a warmup period, training a single model for an extended period and then selecting a checkpoint with the best-reported results would lead to suboptimal performance. More specifically, the number of steps affects the learning rate dynamics, gradually increasing during the warmup period to the peak learning rate before linearly decreasing to zero by the end of training duration. We experimented with different training durations, starting from 1000 to 8000 steps, and surprisingly, we found that both the base and large variants of the model achieved their best results when trained for 3000 steps.

- Warmup Steps: Similarly, the number of warmup steps influences the learning rate dynamics. We experimented with 500, 1000, 1500, and 2000 steps, finding out that 1500 warmup steps worked best for our model's smaller and larger variants.

Name	**LR**	**Batch**	**Steps**	α	γ	λ	β_1	β_2
EmoMosaic-base	$1 * 10^{-4}$	128	1500/3000	0.75	1.75	0.01	0.9	0.98
EmoMosaic-large	$5 * 10^{-5}$	128	1500/3000	0.75	1.75	0.01	0.9	0.98

Table 4.1: Displays the configuration of our top-performing base and large model variants. „LR“ and „Batch“ indicate the learning rate and batch size. The „Steps“ column is formatted as *warm-up steps/total steps*. The parameters α and γ correspond to the settings of the focal loss function. The terms β_1 and β_2 represent the moments used by the optimizer. λ denotes the weight decay rate.

4.2 Quantitative Analysis

To quantitatively evaluate the performance of our models, we followed the methodology outlined in Chapter 3.2.4.

4.2.1 Cross-Domain Performance

The cross-domain performance of our models, *EmoMosaic-base* and *EmoMosaic-large*, is presented in Table 4.2. The *EmoMosaic-large* consistently outperforms *EmoMosaic-base* in all metrics, typically by less than 2 %. Considering that the EmoMosaic-large model outperforms existing state-of-the-art models on two datasets and delivers competitive results on others, we consider its performance to be quite satisfactory. Furthermore, the *EmoMosaic-base* model, despite being roughly a third the size of its larger counterpart, delivers a satisfactory performance, even though it does not surpass recent state-of-the-art models on any dataset.

		macro averaged			**micro averaged**		
Model	**Accuracy**	**P**	**R**	**F1**	**P**	**R**	**F1**
EmoMosaic-base	50.94	50.30	58.25	53.44	54.85	66.20	59.95
EmoMosaic-large	51.70	51.72	60.70	55.38	56.25	68.71	61.79

Table 4.2: Shows the cross-domain performance of our models. P and R denote precision and recall, respectively.

4.2.2 Per-Dataset Performance

GoEmotions

		macro averaged			**micro averaged**		
Model	**Accuracy**	**P**	**R**	**F1**	**P**	**R**	**F1**
Li:BART [19]	-	56.3	53.9	53.8	-	-	-
MIP-GAT [11]	-	56.4	51.7	53.8	-	-	-
REMTA [42]	-	52.12	54.08	52.27	-	-	-
EmoLit:RoBERTa [30]	-	-	-	52	-	-	-
Wang:BERT [33]	-	**57.27**	49.18	51.83	-	-	-
MADAAN:T5-11B [24]	-	-	-	50.9	-	-	-
Seq2Emo [10]	-	-	-	47.28	-	-	-
GoEmotions:BERT-base [8]	-	40	63	46	-	-	-
Li:GPT-3.5-Turbo [19]	-	53.1	40.9	42.1	-	-	-
MADAAN:BART [24]	-	-	-	30	-	-	-
EmoMosaic-base	46.47	51.41	57.81	53.72	52.70	62.53	57.19
EmoMosaic-large	**46.67**	51.35	**58.34**	**53.93**	**52.86**	**63.39**	**57.65**

Table 4.3: Shows the results of our two top-performing models measured on the test set of the *GoEmotions* dataset and compares them with the results of state-of-the-art models. P and R denote precision and recall, respectively.

Our *EmoMosaic-large* model, as shown in Table 4.3, achieves the highest macro-averaged F1 score, outperforming all recent state-of-the-art models. Additionally, the performance of our *EmoMosaic-base* model is closely behind that of the *EmoMosaic-large* model, placing it among the top three best-performing models.

SemEval-2018 Task 1: Affect in Tweets

Model	Accuracy	macro averaged			micro averaged		
		P	R	F1	P	R	F1
RoBERTa-MA [5]	**62.4**	-	-	60.3	-	-	**74.2**
UCCA-GAT [6]	61.2	-	-	60.0	-	-	66.1
DistilBERT-MA [5]	61.3	-	-	58.9	-	-	72.5
XLNet-MA [5]	60.5	-	-	58.4	-	-	70.4
Dep-GAT [6]	59.7	-	-	57.8	-	-	63.5
SpanEmo [2]	-	-	-	57.8	-	-	-
EmoGraph [36]	58.3	-	-	56.9	-	-	69.9
BERT-large+DK [37]	59.5	-	-	56.3	-	-	71.6
TCS Research [25]	58.2	-	-	53.0	-	-	69.3
NTUA-SLP [7]	58.8	-	-	52.8	-	-	70.1
Seq2Emo [10]	58.67	-	-	51.92	-	-	70.02
PlusEmo2Vec [28]	57.6	-	-	49.7	-	-	69.2
EmoMosaic-base	20.65	54.96	62.58	58.44	64.63	73.62	68.83
EmoMosaic-large	22.49	57.97	64.12	**60.72**	67.44	75.27	71.14

Table 4.4: Shows the results of our two top-performing models measured on the test set of the *SemEval-2018 Task 1: Affect in Tweets* dataset and compares them with the results of state-of-the-art models. P and R denote precision and recall, respectively.

Both our models, as shown in Table 4.4, demonstrated strong performance across several metrics except for the subset accuracy. Particularly, the *EmoMosaic-large* model excelled with its highest macro-averaged F1 score, thereby setting a new benchmark for this measure. Additionally, it displayed competitive performance in the micro-averaged F1 score, securing fourth place among compared models.

However, both models faced challenges with subset accuracy, which requires a perfect alignment with all labelled emotions. Upon closer examination, we identified the source of these lower-than-average results. This dataset, unlike others, frequently requires predicting sets of three (31.49 %) or more (11.47 %) co-occurring emotions, as shown in Table 2.4. In contrast, our training corpus mostly required predictions of zero (56.07 %), single emotions (33.09 %) and pairs (7.53 %), making it challenging for our models to work efficiently in cases involving three or more emotions. Therefore, our models predict sets with 3 or more emotions infrequently. This discrepancy led to the observed decrease in subset accuracy.

XED

Model	Accuracy	macro averaged			micro averaged		
		P	**R**	**F1**	**P**	**R**	**F1**
XED:BERT [27]	54.4	-	-	53.6	-	-	-
EmoMosaic-base	51.78	48.47	63.00	54.67	48.62	63.86	55.21
EmoMosaic-large	52.59	50.35	66.54	57.19	50.43	67.43	57.70

Table 4.5: Shows the results of our two top-performing models measured on the test set of the *XED* dataset and compares them with the results of state-of-the-art models.

Both our models, as shown in Table 4.5, seem to outperform the *BERT* model provided as a baseline in the dataset's paper. However, this comparison does not reflect equivalent experimental conditions because the *XED* dataset does not include predefined training, validation, and test splits. The baseline, as introduced in the dataset paper [27], was evaluated using a 5-fold cross-validation using a stratified splitting method of 70:20:10 for training, development, and test phases. Our method used an 80:10:10 split. Hence, comparisons should be interpreted cautiously.

DailyDialog

Model	Accuracy	macro averaged			micro averaged		
		P	**R**	**F1**	**P**	**R**	**F1**
S + PAGE [22]	-	-	-	-	-	-	**64.18**
CESTa [34]	-	-	-	-	-	-	63.12
TUCORE-GCN [13]	-	-	-	-	-	-	61.91
DualGATs [39]	-	-	-	-	-	-	61.84
RoBERTa + FSA [14]	-	-	-	**55.84**	-	-	61.67
CLED [12]	-	-	-	-	-	-	61.23
COSMIC + CKCL [9]	-	-	-	53.09	-	-	60.96
CoMPM [15]	-	-	-	53.15	-	-	60.34
Mtl-ERC-ES [32]	-	-	-	53.06	-	-	60.10
RoBERTa-large [18]	-	-	-	51.95	-	-	59.75
TODKAT [41]	-	-	-	52.56	-	-	58.47
COSMIC [9]	-	-	-	51.05	-	-	58.48
KI-Net [35]	-	-	-	-	-	-	57.3
CoG-BART [20]	-	-	-	-	-	-	56.29
KET [40]	-	-	-	-	-	-	53.37
EmoMosaic-base	84.85	46.34	49.60	46.94	53.44	64.81	58.57
EmoMosaic-large	85.05	47.20	53.80	49.65	54.24	68.77	60.65

Table 4.6: Shows the results of our two top-performing models measured on the test set of the *DailyDialog* dataset and compares them with the results of state-of-the-art models.

Our *EmoMosaic-large* model, as shown in Table 4.6, ranks in the top half among all compared models, featuring a micro-averaged F1 score of 60.65 %. This result showcases its ability to process dialogues. In contrast, the *EmoMosaic-base* model, securing a position fifth from the bottom with a micro-averaged F1 score of 58.57, indicates having difficulties with processing dialogues. Despite this, the results of the *EmoMosaic-large* model are approximately 3% below that of the top performers, still indicating a solid efficacy in processing dialogues. After analyzing the methods used by top performers, we discovered that they all employ Graph Attention Networks (GATs) often coupled with Transformer encoders such as *RoBERTa* or *BERT*. This implies that hybrid models are more suitable for processing conversations.

4.2.3 Per-Label Performance

We initially assumed that the *EmoMosaic-large* model would outperform the *EmoMosaic-base* for all emotions. However, as shown in Table 4.7, our analysis revealed that in 10 instances, the large model actually underperformed, typically by about 1.5% in the F1 score. Notably, in the case of *relief*, the discrepancy reached a striking 15.22%. The cause of this remains unclear. For other emotions, a typical improvement in F1 scores from the EmoMosaic-base to the EmoMosaic-large model ranged between 3-5%.

Both models, achieving F1 scores above 60%, excelled at recognizing the following 13 emotions: *admiration*, *amusement*, *gratitude*, *anger*, *disgust*, *fear*, *grief*, *happiness*, *joy*, *love*, *optimism*, *remorse*, and *sadness*. However, with F1 scores falling below 40%, we identified the following 5 weak spots: *annoyance*, *disappointment*, *nervousness*, *realization*, and *relief* (in the case of *EmoMosaic-large*).

Furthermore, we analysed how well the performance of our models aligned with individual emotional models discussed in Chapter 2.1. This analysis confirmed that both models are highly competent (when considering core emotions) in recognizing the emotions described by those models.

Please note that the results presented in Tables 4.7, 4.8, 4.9, 4.10 were calculated for all datasets combined, measuring the cross-domain performance of our models. Therefore, a direct comparison with other works is not possible.

	EmoMosaic-base			**EmoMosaic-large**		
Emotion	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
admiration	63.82	**80.16**	71.06	**65.25**	79.37	**71.62**
amusement	**74.11**	**94.32**	**83.00**	73.87	93.18	82.41
anger	63.46	74.08	68.36	**64.29**	**76.00**	**69.66**
annoyance	**35.15**	**44.37**	**39.23**	33.81	44.06	38.26
anticipation	39.09	55.15	45.75	**42.10**	**57.99**	**48.78**
approval	**43.40**	**45.87**	**44.60**	42.66	44.73	43.67
caring	**45.67**	**42.96**	**44.27**	40.26	45.93	42.91
confusion	36.10	**56.86**	44.16	**38.76**	52.94	**44.75**
curiosity	**48.48**	67.25	56.34	48.40	**74.65**	**58.73**
desire	53.09	**51.81**	52.44	**65.08**	49.40	**56.16**
disappointment	**35.57**	35.10	35.33	34.36	**37.09**	**35.67**
disapproval	**40.00**	**49.44**	**44.22**	39.14	47.94	43.10
disgust	62.05	71.31	66.36	**63.62**	**72.30**	**67.68**
embarrassment	57.69	**40.54**	**47.62**	**58.33**	37.84	45.90
excitement	37.40	**44.66**	40.71	**39.82**	43.69	**41.67**
fear	61.93	68.69	65.13	**64.22**	**71.24**	**67.55**
gratitude	**93.29**	90.91	**92.09**	91.01	**92.05**	91.53
grief	66.67	66.67	66.67	66.67	66.67	66.67
happiness	58.10	70.76	63.81	**58.21**	**75.23**	**65.63**
joy	73.43	81.18	77.11	**74.55**	**83.53**	**78.78**
love	**64.95**	73.74	69.07	64.13	**76.13**	**69.62**
nervousness	33.33	**43.48**	37.74	**42.86**	39.13	**40.91**
optimism	64.33	76.00	69.68	**66.98**	**79.38**	**72.66**
pessimism	42.31	**52.80**	**46.98**	**43.66**	47.73	45.61
pride	**66.67**	37.50	48.00	63.64	**43.75**	**51.85**
realization	32.71	24.14	27.78	**34.29**	**24.83**	**28.80**
relief	**55.56**	**45.45**	**50.00**	33.33	36.36	34.78
remorse	55.56	89.29	68.49	**57.78**	**92.86**	**71.23**
sadness	58.65	70.14	63.88	**61.08**	**72.67**	**66.37**
surprise	40.02	51.29	44.96	**44.02**	**55.67**	**49.16**
trust	35.33	47.01	40.34	**40.59**	**48.26**	**44.09**

Table 4.7: Shows the results of *EmoMosaic-base* and *EmoMosaic-large* models evaluated at the label-level.

Paul Ekman Model

Emotion	EmoMosaic-base Precision	 Recall	 F1	EmoMosaic-large Precision	 Recall	 F1
anger	63.46	74.08	68.36	**64.29**	**76.00**	**69.66**
disgust	62.05	71.31	66.36	**63.62**	**72.30**	**67.68**
fear	61.93	68.69	65.13	**64.22**	**71.24**	**67.55**
happiness	58.10	70.76	63.81	**58.21**	**75.23**	**65.63**
sadness	58.65	70.14	63.88	**61.08**	**72.67**	**66.37**
surprise	40.02	51.29	44.96	**44.02**	**55.67**	**49.16**

Table 4.8: Shows the results of *EmoMosaic-base* and *EmoMosaic-large* models evaluated at the label-level. This table exclusively shows results for the emotion categories in *Paul Ekman's* model.

In Table 4.8, which outlines performance for the six basic emotions identified by Paul Ekman, both models demonstrated strong overall capabilities, with a slight exception of *surprise*. Additionally, the *EmoMosaic-large* model consistently outperformed its smaller counterpart.

Parrot Model

Emotion	EmoMosaic-base Precision	 Recall	 F1	EmoMosaic-large Precision	 Recall	 F1
anger	63.46	74.08	68.36	**64.29**	**76.00**	**69.66**
fear	61.93	68.69	65.13	**64.22**	**71.24**	**67.55**
sadness	58.65	70.14	63.88	**61.08**	**72.67**	**66.37**
joy	73.43	81.18	77.11	**74.55**	**83.53**	**78.78**
love	**64.95**	73.74	69.07	64.13	**76.13**	**69.62**
surprise	40.02	51.29	44.96	**44.02**	**55.67**	**49.16**

Table 4.9: Shows the results of *EmoMosaic-base* and *EmoMosaic-large* models evaluated at the label-level. This table exclusively shows results corresponding to primary emotions as defined in *Parrot's model.*

In Table 4.9, which evaluates primary emotions from Parrot's model, both our models demonstrated strong overall performance, with the *EmoMosaic-large* consistently outperforming the *EmoMosaic-base* across all metrics. However, it is worth noting that both models, particularly the *EmoMosaic-base*, struggled with the emotion of *surprise.*

Wheel of Emotions Model

	EmoMosaic-base			EmoMosaic-large		
Emotion	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
anger	63.46	74.08	68.36	**64.29**	**76.00**	**69.66**
anticipation	39.09	55.15	45.75	**42.10**	**57.99**	**48.78**
disgust	62.05	71.31	66.36	**63.62**	**72.30**	**67.68**
fear	61.93	68.69	65.13	**64.22**	**71.24**	**67.55**
joy	73.43	81.18	77.11	**74.55**	**83.53**	**78.78**
sadness	58.65	70.14	63.88	**61.08**	**72.67**	**66.37**
surprise	40.02	51.29	44.96	**44.02**	**55.67**	**49.16**
trust	35.33	47.01	40.34	**40.59**	**48.26**	**44.09**

Table 4.10: Shows the results of *EmoMosaic-base* and *EmoMosaic-large* models evaluated at the label level. This table exclusively lists results for the primary emotions as defined in the *Wheel of Emotions* model.

In Table 4.10, which evaluates primary emotions from the Wheel of Emotions model, the *EmoMosaic-large* generally demonstrated good performance, even though it displayed mediocre results for the emotion of *trust* and *surprise*.

4.2.4 Calibration

We evaluated the calibration of our models using two metrics: Expected Calibration Error (ECE) and Brier score, which are presented in Table 4.11. Since no other studies have assessed calibration, we have not compared our results with any other methods. In emotion classification, the data's somewhat subjective and ambiguous nature makes precise probabilistic calibration less critical than in fields like medical diagnostics, making a slightly higher ECE acceptable. The *EmoMosaic-large* model typically exhibited slightly lower values for both metrics than *EmoMosaic-base*, indicating its slightly superior calibration.

To assess the overall calibration of our models, we averaged all the measured ECE and Brier scores. The average ECE values, while on the higher side for some applications, are still perfectly acceptable for emotion classification. The average Brier score for both models is notably low, indicating highly reliable predictions. However, for emotions such as pessimism, optimism, anticipation, approval, and realization, both models display significantly higher ECE and Brier scores than the average, indicating the model's miscalibration of these emotions.

Emotion	**EmoMosaic-base**		**EmoMosaic-large**	
	ECE	**Brier**	**ECE**	**Brier**
admiration	0.1040	0.0530	**0.1005**	**0.0519**
amusement	0.0896	0.0237	**0.0778**	**0.0211**
anger	0.0802	0.0535	**0.0784**	**0.0508**
annoyance	0.1516	0.0742	**0.1493**	**0.0740**
anticipation	0.2012	0.1358	**0.1944**	**0.1301**
approval	**0.1756**	**0.0816**	0.1758	0.0818
caring	**0.0846**	**0.0272**	0.0907	0.0291
confusion	0.1056	0.0370	**0.1048**	**0.0361**
curiosity	**0.0821**	**0.0392**	0.0911	0.0407
desire	0.0809	0.0185	**0.0731**	**0.0163**
disappointment	**0.1284**	**0.0449**	0.1355	0.0474
disapproval	0.1332	0.0585	**0.1234**	**0.0557**
disgust	0.0785	0.0479	**0.0704**	**0.0455**
embarrassment	0.0726	0.0124	**0.0650**	**0.0108**
excitement	**0.0990**	**0.0286**	0.1021	0.0287
fear	0.0740	0.0312	**0.0707**	**0.0290**
gratitude	0.0850	0.0175	**0.0789**	**0.0165**
grief	**0.0428**	0.0040	0.0457	**0.0038**
happiness	**0.0928**	0.0731	0.0943	**0.0719**
joy	**0.1086**	0.0673	0.1098	**0.0632**
love	0.0796	0.0461	**0.0793**	**0.0452**
nervousness	0.0616	0.0094	**0.0564**	**0.0079**
optimism	**0.1105**	0.0815	0.1113	**0.0788**
pessimism	0.1918	0.1230	**0.1733**	**0.1139**
pride	**0.0533**	0.0060	0.0543	**0.0059**
realization	0.1619	0.0534	**0.1561**	**0.0516**
relief	**0.0605**	**0.0067**	0.0643	0.0070
remorse	0.0542	0.0095	**0.0533**	**0.0087**
sadness	0.0936	0.0514	**0.0842**	**0.0480**
surprise	0.1037	0.0425	**0.1017**	**0.0408**
trust	0.1760	0.0895	**0.1634**	**0.0821**
average	0.1038	0.0467	**0.1010**	**0.0450**

Table 4.11: Shows the calibration results of our two top-performing models evaluated at the label level.

4.3 Qualitative Analysis

We conducted a qualitative analysis to supplement our findings, distinguishing us from other studies. To empirically validate the effectiveness of our models, we followed the methodology outlined in chapter 3.2.4. We conducted experiments, each designed to evaluate a specific capability. First, we evaluated how well they handle sentences of varying complexity ranging from easy-to-classify to challenging cases of irony. Second, we examined their ability to perform consistently across different domains (legal documents, everyday conversations, etc.). Finally, we assessed the ability of the models to process dialogues.

4.3.1 Trivial Sentences Suite

The Trivial Suit's examples and responses are provided in Table 4.12 and Table 4.13, respectively. Both models produced nearly identical responses, except for sentence *si-se-1*. In this case, the *EmoMosaic-base* model provided a more favourable response because it recognized annoyance. Nevertheless, we consider all responses to be valid.

ID	**Prompt**
si-se-1	I can't believe you lied to me!
si-se-2	I don't understand these instructions.
si-se-3	How does this machine work?
si-se-4	I'm disappointed by the movie's ending.
si-se-5	I'm so excited for the holiday!
si-se-6	Thank you so much for your help.
si-se-7	I'm very happy with the results.
si-se-8	I love you more than words can express.
si-se-9	I doubt that will ever happen.
si-se-10	Wow, that was unexpected!

Table 4.12: Shows the examples used in Trivial Sentences Suite. Note these examples were generated using ChatGPT 4.

ID	**EmoMosaic-base**	**EmoMosaic-large**
si-se-1	anger, annoyance	anger
si-se-2	confusion	confusion
si-se-3	curiosity	curiosity
si-se-4	disappointment, pessimism	disappointment, pessimism
si-se-5	excitement, happiness, joy	excitement, happiness, joy
si-se-6	gratitude, happiness	gratitude, happiness
si-se-7	happiness, joy	happiness, joy
si-se-8	happiness, love	happiness, love
si-se-9	confusion, pessimism	confusion, pessimism
si-se-10	surprise	surprise

Table 4.13: Shows the responses that the *EmoMosaic-base* and *EmoMosaic-large* models created in response to the Trivial Sentences Suite's prompts.

4.3.2 Complex Sentences Suite

The Complex Sentences Suite's examples and responses are provided in Table 4.14 and Table 4.15, respectively. Upon reviewing responses to sentences *co-se-1* and *co-se-3*, we uncovered that both models struggled to detect intended emotions in complex cases of irony. Both models successfully detected *approval* for sentence *co-se-2*, although *confusion* was also expected. For sentence *co-se-5*, we lean towards the response of the *EmoMosaic-large* model because it recognized *admiration*, though the response of its smaller counterpart is also correct. For sentence *co-se-7*, both responses appear valid: the *EmoMosaic-large* model didn't detect any specific emotions, which seems appropriate given the lack of strong emotional expression, whereas the *EmoMosaic-base* model identified *love* and *happiness*. While *happiness* is questionable, *love* could be inferred based on the context. Generally, both models correctly identified emotions, with the exception of instances involving irony.

ID	**Prompt**
co-se-1	Oh great, another day in paradise working with this ancient computer.
co-se-2	Your explanation is as clear as mud, but please, go on.
co-se-3	I'm totally looking forward to giving a speech in front of hundreds, said no one ever.
co-se-4	Of course, the elevator would break when I'm late and on the top floor.
co-se-5	Somehow, you always manage to read between the lines.
co-se-6	Your energy is contagious! I find myself planning more ambitious projects after our meetings.
co-se-7	You know you are like a brother to me.

Table 4.14: Shows examples used in Complex Sentences Suite. Note these examples were generated using ChatGPT 4.

ID	**EmoMosaic-base**	**EmoMosaic-large**
co-se-1	admiration, happiness, joy	admiration, happiness, joy
co-se-2	approval	approval
co-se-3	anticipation, excitement, happiness	anticipation, excitement, happiness
co-se-4	pessimism	pessimism
co-se-5	realization	admiration
co-se-6	admiration, anticipation, happiness, optimism	admiration, happiness, joy, optimism
co-se-7	happiness, love	

Table 4.15: Shows the responses that the *EmoMosaic-base* and *EmoMosaic-large* models created in response to the Complex Sentences Suite's prompts.

4.3.3 Cross-Domain Efficiency Suite

The Cross-Domain Efficiency Suite's examples and responses are provided in Table 4.16 and Table 4.17, respectively. The results reveal that both models identified similar emotions for most prompts. The *EmoMosaic-base* model identified *approval* for the legal document prompt *cr-do-2*, while the *EmoMosaic-large* model also recognised *happiness* and *joy*. We prefer the response the *EmoMosaic-large* provided in this case. Responses to sentences *cr-do-4*, *cr-do-5*, and *cr-do-8* slightly differ, but we couldn't determine which ones we like more since all seem perfectly valid. Further differences were found in response to the sports commentary *cr-do-6*, where the *EmoMosaic-base* model identified only *surprise*, whereas its larger counterpart also identified *admiration*, favouring the *EmoMosaic-large* model. We've empirically verified that the model can respond effectively to texts from various domains. While not every response was perfect, all met our criteria.

ID	Domain	Prompt
cr-do-1	literature	Under the endless sky, she rediscovered a sense of boundless possibility.
cr-do-2	legal document	The parties hereby express their satisfaction with the resolved settlement terms.
cr-do-3	news	Local hero saves a family from fire, community praises his quick action.
cr-do-4	blog post	Embracing challenges in the workplace can lead to substantial growth.
cr-do-5	advertisement	Discover freedom like never before with our latest range of electric cars.
cr-do-6	sports commentary	With seconds left, he scores! Unbelievable performance!
cr-do-7	conversation	I finally passed my driving test. Let's hit the road this weekend!
cr-do-8	social media	Just welcomed our new baby into the world. #newborn #family

Table 4.16: Shows examples used in Cross-Domain Efficiency Suite. Note these examples were generated using ChatGPT 4.

ID	EmoMosaic-base	EmoMosaic-large
cr-do-1	happiness, joy, optimism	happiness, joy, optimism
cr-do-2	approval	approval, happiness, joy
cr-do-3	admiration, happiness, trust	admiration, happiness, trust
cr-do-4	approval, happiness, optimism	anticipation, approval, optimism, trust
cr-do-5	anticipation, happiness, joy	happiness, joy
cr-do-6	surprise	admiration, surprise
cr-do-7	excitement, happiness, joy	anticipation, excitement, happiness, joy
cr-do-8	excitement, happiness, joy, love	happiness, joy, love, optimism

Table 4.17: Shows the responses that the *EmoMosaic-base* and *EmoMosaic-large* models created in response to the Cross-Domain Efficiency Suite's prompts.

4.3.4 Dialogues Suite

The Dialogues Suite's examples and responses are provided in Table 4.18 and Table 4.19, respectively. Overall, the *EmoMosaic-large* model performed satisfactorily and generally outperformed the *EmoMosaic-base* in handling dialogues. However, when it comes to ironical sentences such as *di-2-2* and *di-3-2*, both models failed to recognise intended emotions, an issue we have already uncovered in the Complex Suit. In instances like *di-5-2*, the *EmoMosaic-large* model accurately identified *pessimism*, unlike its smaller counterpart. Additionally, while both models correctly recognized surprise in sentence *di-3-1*, the *EmoMosaic-large* also successfully identified *excitement*. Nevertheless, in the case of *di-4-1*, neither model identified *happiness* or *relief*, which was expected. This suit posed a great challenge for both models.

ID	Context	Sentence
di-1-1	Good job! We will promote you to regional manager.	That came out of nowhere.
di-1-2	We have to let you go. Your performance was not to our standards!	
di-2-1	They found your lost luggage and will deliver it tonight.	Oh, that's wonderful!
di-2-2	They've permanently lost your luggage, no compensation.	
di-3-1	The article you have been working on was selected as the feature for next month's issue. Congrats!	Really? That's unbelievable!
di-3-2	The article you have been working on was permanently rejected. Cite properly next time!	
di-4-1	I managed to save all your work before the computer crashed.	Oh my god. I spent a whole week on that.
di-4-2	All your work was lost when the computer crashed.	
di-5-1	They approved your leave request for the holidays.	Just like last year.
di-5-2	They denied your leave request for the holidays. You will have to work even on Christmas Eve.	

Table 4.18: Show examples used in Dialogues Suite. Note these examples were generated using ChatGPT 4.

ID	EmoMosaic-base	EmoMosaic-large
di-1-1	admiration,happiness	admiration,happiness
di-1-2	disappointment,disapproval,pessimism	disappointment,pessimism
di-2-1	admiration,happiness	admiration,excitement,happiness,joy
di-2-2	admiration,happiness	admiration,happiness
di-3-1	surprise	excitement,surprise
di-3-2	surprise	surprise
di-4-1	surprise	surprise
di-4-2	disappointment,pessimism,sadness,surprise	disappointment,sadness,surprise
di-5-1	approval	approval
di-5-2		pessimism

Table 4.19: Shows the responses that the *EmoMosaic-base* and *EmoMosaic-large* models created in response to the Dialogues Suite prompts.

4.4 Computational Resources

Computational resources for this project were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. All experiments were carried out on multiple nodes of the *zia* cluster. Each node of the cluster featured 4x NVIDIA A100 SXM4 40GB GPUs. In total, these experiments ran for approximately 124.5 CPU days. As we utilized 4 CPUs for 1 GPU, we estimate that the total GPU usage was around 31.125 GPU days.

4.4.1 Experiments

Additionally, we measured the amount of floating-point operations (GFLOPS), the time, and the peak graphical memory (VRAM) usage required to process N sentences, each having L tokens. We utilized a setup equipped with an RTX 4070 12GB GPU and a Ryzen 9 7900 CPU, complemented by 32GB of DDR5 RAM. The recorded values were calculated as an average across 100 tensors.

Table 4.20 shows the memory and computational requirements for our models. We measured the average time to process a tensor of shape *(1,128)*, corresponding to a single sentence with 128 tokens. Table 4.21 shows the memory and computational requirements for our models. We measured the average time to process a tensor of shape *(32,128)*, corresponding to 32 sentences with 128 tokens.

Model	$GFLOPS_{(1,128)}$	$VRAM_{(1,128)}$	$Time_{(1,128)}$
EmoMosiac-base	11.19	0.51 GB	8 ms
EmoMosiac-large	39.49	1.44 GB	15 ms

Table 4.20: Compares the computational complexity of the proposed models using tensors of shape *(1,128)*.

Model	$GFLOPS_{(32,128)}$	$VRAM_{(32,128)}$	$Time_{(32,128)}$
EmoMosiac-base	357.96	0.65 GB	52 ms
EmoMosiac-large	1263.78	1.61 GB	182 ms

Table 4.21: Compares the computational complexity of the proposed models using tensors of shape *(32,128)*.

Chapter 5

Conclusion

Most human interactions are either text-based or can be converted to text using *speech-to-text* technologies. This thesis is dedicated to recognizing emotions from these texts. We began with a thorough and extensive literature review. This analysis highlighted the need for diverse, high-quality datasets to develop robust, well-generalizable methods. We have identified *SemEval-2018 Task 1: Affect in Tweets*, *GoEmotions*, *XED* and *DailyDialog* as datasets that, when combined, meet the established criteria. See Chapter 2.2 for further details. The analysis also revealed that models based on Transformers such as *BERT*, *RoBERTa* and *BART* excel on the following datasets *SemEval-2018 Task 1: Affect in Tweets* and *GoEmotions*. Additionally, from the results reported on the *DailyDialog* dataset, we concluded that hybrid models incorporating Transformer encoders like *BERT* alongside graph attention networks (GANs) deliver the best performance in conversational processing. Subsequently, as detailed in Chapter 2.5, we identified the following key areas for advancement in emotion recognition from text:

1. Cross-domain Efficacy: Most current methods focus on one specific domain, often Twitter or Reddit, and are typically not tested on data from other domains. This significantly limits their applicability in real-world situations, which are usually diverse. Studies that addressed cross-domain effectiveness often oversimplified the issue. To unify generally differing categories (across datasets), they remap many categories into a much smaller set. For instance, they map the 27 categories from the *GoEmotions* dataset to the six basic emotions according to *Ekman's model*. A similar type of remapping is applied to other datasets as well. Consequently, this approach greatly loses depth and makes comparisons with other methods impossible.

2. Results Analysis: To perform a thorough analysis of the method's performance, it is important to evaluate multiple metrics at both the individual dataset and category/emotion levels. However, most studies only report 1-3 global metrics, which is insufficient. Additionally, none of the studies have assessed the calibration of their method.

3. Usability: Researchers often publish only papers without accompanying the code or models used in their research. This hinders the reproducibility and practical applications of their findings.

In this study, we propose a method that tackles all the outlined challenges. We started by selecting appropriate datasets and continued by establishing a consistent data format.

Specifically, we selected *SemEval-2018 Task 1: Affect in Tweets*, *GoEmotions*, *XED* and *DailyDialog*. By combining these datasets, we created a diverse, high-quality, *multi-label* dataset, which we call the *EmoMosaic-dataset*. It is important to mention that we preserved the original categories for all datasets, unlike other studies. See Chapter 3.2.1 for more details.

In this study, we propose two models, *EmoMosaic-base* and *EmoMosaic-large*, the former based on the *RoBERTa-base* model and the latter on the *RoBERTa-large* model. In contrast to other studies, we trained our models on all the datasets simultaneously while preserving their original categories. To achieve this, our models process sentences stripped of any information regarding categories or datasets. This forces the models to predict the entire spectrum of categories, which was created by unifying emotions from individual datasets. Since we do not remap categories, each sentence in our unified dataset, the *EmoMosaic-dataset*, contains several categories lacking proper annotations. To avoid mistakes, we mask these categories during training. Consequently, our models perform well across different domains and are directly comparable to other methods. Additionally, our models retain the original level of detail, enabling understanding of complex emotions. Further details can be found in Chapter 3.2.

After training the models, we proceeded to testing. We followed the procedures outlined in Chapter 3.2.4. We first evaluated the performance of our models at the level of individual datasets and then at the level of individual categories/emotions. Our best-performing model *EmoMosaic-large* demonstrated excellent results across multiple domains and outperformed current *state-of-the-art* (SOTA) models on the following datasets: *SemEval-2018 Task 1: Affect in Tweets* and *GoEmotions*. *EmoMosaic-large* achieves a macro-averaged F1 score of 60.72 % (an increase of 0.42 % over SOTA) on the *SemEval-2018 Task 1: Affect in Tweets* dataset and on the *GoEmotions* dataset 53.93 % (an increase of 0.13 % over SOTA). Its smaller counterpart, *EmoMosaic-base*, falls short of the SOTA models and, on average (calculated across all datasets), lags behind it by 1.94 % in the macro-averaged F1 score. Despite being roughly one-third of its size, *EmoMosaic-base* provides a good balance between performance and computational efficiency. Subsequently, we tested the ability of our models to process conversations using the *DailyDialog* dataset. The *EmoMosaic-large* model achieved a micro-averaged F1 score of 60.65% (3.56% decrease from SOTA). Although it fell short of the *state-of-the-art* methods, it can be concluded that it performed reasonably well given the value. The top-performing methods were based on hybrid models combining Transformer encoders and Graph Neural Networks (GATs) in all cases. A detailed comparison and discussion of the results can be found in Chapter 4. Next, we evaluated the performance of our models at the level of individual categories, revealing their respective strengths and weaknesses. This analysis can be found in Chapter 4.2. Following this, we evaluated the calibration of our models and found that both proposed models are relatively well-calibrated. Therefore, we consider their predictions to be credible. Finally, we substantiate our claims by empirically testing our models under various scenarios. Although both of our models performed well in general, we found that they responded poorly to ironic sentences. In addition, we concluded that the *EmoMosaic-base* model is not very suitable for processing conversations. See Chapter 4.3 for more details. However, none of the models showed systematic errors, and both models proved themselves for processing texts from different domains, which was the main goal of this study.

By making all of our models and code publicly available, we increase the reproducibility and applicability of our research and provide valuable resources for further research and practical use. All models and code are freely available at https://huggingface.co/vtlustos.

In conclusion, by proposing and thoroughly testing models that exhibit strong cross-domain performance, we have advanced the field of emotion recognition from text. Further research could focus on enhancing the accuracy of these systems, though it is speculated that achieving macro-averaged F1 scores over 75 % may be unattainable due to the inherent ambiguity of emotional expressions. Another area of research could involve developing a human baseline that would enable comparisons between artificial models and human-level performance. Moreover, while large language models (LLMs) are not yet suitable for emotion recognition, their continuous development suggests they might be in the future. Adapting an LLM for emotion recognition could increase the flexibility of such systems, as they do not require hard-wiring the labels into their architecture.

Bibliography

[1] *Putting some emotion into your design – Plutchik's Wheel of emotions* [online]. Interaction Design Foundation. Accessed: 2024-5-11. Available at: https://www.interaction-design.org/literature/article/putting-some-emotion-into-your-design-plutchik-s-wheel-of-emotions.

[2] Alhuzali, H. and Ananiadou, S. SpanEmo: Casting Multi-label Emotion Classification as Span-prediction. In: Merlo, P., Tiedemann, J. and Tsarfaty, R., ed. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, April 2021, p. 1573–1584. DOI: 10.18653/v1/2021.eacl-main.135. Available at: https://aclanthology.org/2021.eacl-main.135.

[3] Alon, D. and Ko, J. *Goemotions: A dataset for fine-grained emotion classification*. Available at: https://blog.research.google/2021/10/goemotions-dataset-for-fine-grained.html.

[4] Alswaidan, N. and Menai, M. E. B. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*. Springer. 2020, vol. 62, p. 2937–2987. Available at: https://link.springer.com/article/10.1007/s10115-020-01449-0.

[5] Ameer, I., Bölücü, N., Siddiqui, M. H. F., Can, B., Sidorov, G. et al. Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*. Elsevier. 2023, vol. 213, p. 118534. Available at: https://www.sciencedirect.com/science/article/pii/S0957417422016098.

[6] Ameer, I., Bolucu, N., Sidorov, G. and Can, B. Emotion Classification in Texts Over Graph Neural Networks: Semantic Representation is Better Than Syntactic. *IEEE Access*. Institute of Electrical and Electronics Engineers Inc. 2023, vol. 11, p. 56921–56934. DOI: 10.1109/ACCESS.2023.3281544. ISSN 2169-3536. Publisher Copyright: © 2013 IEEE.

[7] Baziotis, C., Nikolaos, A., Chronopoulou, A., Kolovou, A., Paraskevopoulos, G. et al. NTUA-SLP at SemEval-2018 Task 1: Predicting Affective Content in Tweets with Deep Attentive RNNs and Transfer Learning. In: Apidianaki, M., Mohammad, S. M., May, J., Shutova, E., Bethard, S. et al., ed. *Proceedings of the 12th International Workshop on Semantic Evaluation*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, p. 245–255. DOI: 10.18653/v1/S18-1037. Available at: https://aclanthology.org/S18-1037.

[8] Demszky, D., Movshovitz Attias, D., Ko, J., Cowen, A., Nemade, G. et al. GoEmotions: A Dataset of Fine-Grained Emotions. In: Jurafsky, D., Chai, J., Schluter, N. and Tetreault, J., ed. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, p. 4040–4054. DOI: 10.18653/v1/2020.acl-main.372. Available at: https://aclanthology.org/2020.acl-main.372.

[9] Ghosal, D., Majumder, N., Gelbukh, A., Mihalcea, R. and Poria, S. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In: Cohn, T., He, Y. and Liu, Y., ed. *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, November 2020, p. 2470–2481. DOI: 10.18653/v1/2020.findings-emnlp.224. Available at: https://aclanthology.org/2020.findings-emnlp.224.

[10] Huang, C., Trabelsi, A., Qin, X., Farruque, N., Mou, L. et al. Seq2Emo: A Sequence to Multi-Label Emotion Classification Model. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani Tur, D., Beltagy, I. et al., ed. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, p. 4717–4724. DOI: 10.18653/v1/2021.naacl-main.375. Available at: https://aclanthology.org/2021.naacl-main.375.

[11] Jia, A., Zhang, Y., Uprety, S. and Song, D. Learning interactions across sentiment and emotion with graph attention network and position encodings. *Pattern Recognition Letters*. 2024, vol. 180, p. 33–40. DOI: https://doi.org/10.1016/j.patrec.2024.02.013. ISSN 0167-8655. Available at: https://www.sciencedirect.com/science/article/pii/S0167865524000461.

[12] Kang, Y. and Cho, Y.-S. Improving Contrastive Learning in Emotion Recognition in Conversation via Data Augmentation and Decoupled Neutral Emotion. In: Graham, Y. and Purver, M., ed. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. St. Julian's, Malta: Association for Computational Linguistics, March 2024, p. 2194–2208. Available at: https://aclanthology.org/2024.eacl-long.134.

[13] Lee, B. and Choi, Y. S. Graph Based Network with Contextualized Representations of Turns in Dialogue. In: Moens, M.-F., Huang, X., Specia, L. and Yih, S. W.-t., ed. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, November 2021, p. 443–455. DOI: 10.18653/v1/2021.emnlp-main.36. Available at: https://aclanthology.org/2021.emnlp-main.36.

[14] Lee, J. The Emotion is Not One-hot Encoding: Learning with Grayscale Label for Emotion Recognition in Conversation. In: *Proc. Interspeech 2022*. 2022, p. 141–145. DOI: 10.21437/Interspeech.2022-551. Available at: https://api.semanticscholar.org/CorpusID:249674470.

[15] Lee, J. and Lee, W. CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation. In: Carpuat, M., Marneffe,

M.-C. de and Meza Ruiz, I. V., ed. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, July 2022, p. 5669–5679. DOI: 10.18653/v1/2022.naacl-main.416. Available at: https://aclanthology.org/2022.naacl-main.416.

[16] Lee, S. J., Lim, J., Paas, L. and Ahn, H. S. Transformer transfer learning emotion detection model: synchronizing socially agreed and self-reported emotions in big data. *Neural Computing and Applications*. may 2023, vol. 35, no. 15, p. 10945–10956. Available at: https://doi.org/10.1007/s00521-023-08276-8.

[17] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A. et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: Jurafsky, D., Chai, J., Schluter, N. and Tetreault, J., ed. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, p. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. Available at: https://aclanthology.org/2020.acl-main.703.

[18] Li, J., Lin, Z., Fu, P. and Wang, W. Past, Present, and Future: Conversational Emotion Recognition through Structural Modeling of Psychological Knowledge. In: Moens, M.-F., Huang, X., Specia, L. and Yih, S. W.-t., ed. *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, November 2021, p. 1204–1214. DOI: 10.18653/v1/2021.findings-emnlp.104. Available at: https://aclanthology.org/2021.findings-emnlp.104.

[19] Li, J., Zhang, Y., Chen, S. and Xu, R. Enhancing Multi-Label Classification via Dynamic Label-Order Learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2024, vol. 38, no. 17, p. 18527–18535. Available at: https://doi.org/10.1609/aaai.v38i17.29814.

[20] Li, S., Yan, H. and Qiu, X. Contrast and generation make bart a good dialogue emotion recognizer. In: *Proceedings of the AAAI conference on artificial intelligence*. 2022, vol. 36, no. 10, p. 11002–11010.

[21] Li, Y., Su, H., Shen, X., Li, W., Cao, Z. et al. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In: Kondrak, G. and Watanabe, T., ed. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, November 2017, p. 986–995. Available at: https://aclanthology.org/I17-1099.

[22] Liang, C., Xu, J., Lin, Y., Yang, C. and Wang, Y. S+PAGE: A Speaker and Position-Aware Graph Neural Network Model for Emotion Recognition in Conversation. In: He, Y., Ji, H., Li, S., Liu, Y. and Chang, C.-H., ed. *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online only: Association for Computational Linguistics, November 2022, p. 148–157. Available at: https://aclanthology.org/2022.aacl-main.12.

[23] Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Dollár, P. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2020, vol. 42, no. 2, p. 318–327. DOI: 10.1109/TPAMI.2018.2858826.

[24] Madaan, A., Rajagopal, D., Tandon, N., Yang, Y. and Bosselut, A. Conditional set generation using Seq2seq models. In: Goldberg, Y., Kozareva, Z. and Zhang, Y., ed. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing.* Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, December 2022, p. 4874–4896. DOI: 10.18653/v1/2022.emnlp-main.324. Available at: https://aclanthology.org/2022.emnlp-main.324.

[25] Meisheri, H. and Dey, L. TCS Research at SemEval-2018 Task 1: Learning Robust Representations using Multi-Attention Architecture. In: Apidianaki, M., Mohammad, S. M., May, J., Shutova, E., Bethard, S. et al., ed. *Proceedings of the 12th International Workshop on Semantic Evaluation.* New Orleans, Louisiana: Association for Computational Linguistics, June 2018, p. 291–299. DOI: 10.18653/v1/S18-1043. Available at: https://aclanthology.org/S18-1043.

[26] Mohammad, S., Bravo Marquez, F., Salameh, M. and Kiritchenko, S. SemEval-2018 Task 1: Affect in Tweets. In: Apidianaki, M., Mohammad, S. M., May, J., Shutova, E., Bethard, S. et al., ed. *Proceedings of the 12th International Workshop on Semantic Evaluation.* New Orleans, Louisiana: Association for Computational Linguistics, June 2018, p. 1–17. DOI: 10.18653/v1/S18-1001. Available at: https://aclanthology.org/S18-1001.

[27] Öhman, E., Pàmies, M., Kajava, K. and Tiedemann, J. XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection. In: Scott, D., Bel, N. and Zong, C., ed. *Proceedings of the 28th International Conference on Computational Linguistics.* Barcelona, Spain (Online): International Committee on Computational Linguistics, December 2020, p. 6542–6552. DOI: 10.18653/v1/2020.coling-main.575. Available at: https://aclanthology.org/2020.coling-main.575.

[28] Park, J. H., Xu, P. and Fung, P. PlusEmo2Vec at SemEval-2018 Task 1: Exploiting emotion knowledge from emoji and #hashtags. In: Apidianaki, M., Mohammad, S. M., May, J., Shutova, E., Bethard, S. et al., ed. *Proceedings of the 12th International Workshop on Semantic Evaluation.* New Orleans, Louisiana: Association for Computational Linguistics, June 2018, p. 264–272. DOI: 10.18653/v1/S18-1039. Available at: https://aclanthology.org/S18-1039.

[29] Peng, H., Wang, X., Chen, J., Li, W., Qi, Y. et al. *When does In-context Learning Fall Short and Why? A Study on Specification-Heavy Tasks.* 2024. Available at: https://openreview.net/forum?id=Cw6lk56w6z.

[30] Rei, L. and Mladenić, D. Detecting Fine-Grained Emotions in Literature. *Applied Sciences.* 2023, vol. 13, no. 13. DOI: 10.3390/app13137502. ISSN 2076-3417. Available at: https://www.mdpi.com/2076-3417/13/13/7502.

[31] Semeraro, A., Vilella, S. and Ruffo, G. PyPlutchik: Visualising and comparing emotion-annotated corpora. *PLoS One.* Public Library of Science (PLoS). september 2021, vol. 16, no. 9, p. e0256503. Available at: https://doi.org/10.1371/journal.pone.0256503.

[32] Wang, J. and Mine, T. Multi-Task Learning for Emotion Recognition in Conversation with Emotion Shift. In: Huang, C.-R., Harada, Y., Kim, J.-B., Chen, S., Hsu, Y.-Y. et al., ed. *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*. Hong Kong, China: Association for Computational Linguistics, December 2023, p. 257–266. Available at: https://aclanthology.org/2023.paclic-1.26.

[33] Wang, K., Jing, Z., Su, Y. and Han, Y. Large Language Models on Fine-grained Emotion Detection Dataset with Data Augmentation and Transfer Learning. *ArXiv preprint arXiv:2403.06108*. 2024. Available at: https://arxiv.org/pdf/2403.06108.

[34] Wang, Y., Zhang, J., Ma, J., Wang, S. and Xiao, J. Contextualized Emotion Recognition in Conversation as Sequence Tagging. In: Pietquin, O., Muresan, S., Chen, V., Kennington, C., Vandyke, D. et al., ed. *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 1st virtual meeting: Association for Computational Linguistics, July 2020, p. 186–195. DOI: 10.18653/v1/2020.sigdial-1.23. Available at: https://aclanthology.org/2020.sigdial-1.23.

[35] Xie, Y., Yang, K., Sun, C., Liu, B. and Ji, Z. Knowledge-Interactive Network with Sentiment Polarity Intensity-Aware Multi-Task Learning for Emotion Recognition in Conversations. In: Moens, M.-F., Huang, X., Specia, L. and Yih, S. W.-t., ed. *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, November 2021, p. 2879–2889. DOI: 10.18653/v1/2021.findings-emnlp.245. Available at: https://aclanthology.org/2021.findings-emnlp.245.

[36] Xu, P., Liu, Z., Winata, G. I., Lin, Z. and Fung, P. Emograph: Capturing emotion correlations using graph networks. *ArXiv preprint arXiv:2008.09378*. 2020. Available at: https://arxiv.org/abs/2008.09378.

[37] Ying, W., Xiang, R. and Lu, Q. Improving Multi-label Emotion Classification by Integrating both General and Domain-specific Knowledge. In: Xu, W., Ritter, A., Baldwin, T. and Rahimi, A., ed. *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Hong Kong, China: Association for Computational Linguistics, November 2019, p. 316–321. DOI: 10.18653/v1/D19-5541. Available at: https://aclanthology.org/D19-5541.

[38] Zad, S., Heidari, M., Jones, J. H. J. and Uzuner, O. Emotion Detection of Textual Data: An Interdisciplinary Survey. In: *2021 IEEE World AI IoT Congress (AIIoT)*. 2021, p. 0255–0261. DOI: 10.1109/AIIoT52608.2021.9454192. Available at: https://ieeexplore.ieee.org/document/9454192.

[39] Zhang, D., Chen, F. and Chen, X. DualGATs: Dual Graph Attention Networks for Emotion Recognition in Conversations. In: Rogers, A., Boyd Graber, J. and Okazaki, N., ed. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, July 2023, p. 7395–7408. DOI: 10.18653/v1/2023.acl-long.408. Available at: https://aclanthology.org/2023.acl-long.408.

[40] Zhong, P., Wang, D. and Miao, C. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. In: Inui, K., Jiang, J., Ng, V. and Wan, X., ed. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, November 2019, p. 165–176. DOI: 10.18653/v1/D19-1016. Available at: https://aclanthology.org/D19-1016.

[41] Zhu, L., Pergola, G., Gui, L., Zhou, D. and He, Y. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In: Zong, C., Xia, F., Li, W. and Navigli, R., ed. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, August 2021, p. 1571–1582. DOI: 10.18653/v1/2021.acl-long.125. Available at: https://aclanthology.org/2021.acl-long.125.

[42] Zou, J., Zhang, Y., Yang, J., Wu, S., Jiang, M. et al. Emotion Recognition in Social Network Texts Based on A Multilingual Architecture. In: *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*. 2023, p. 809–815. DOI: 10.1109/ICDMW60847.2023.00109. Available at: https://ieeexplore.ieee.org/abstract/document/10411662.

Appendix A

Contents of the DVD

The attached DVD contains the following items:

- `code/` folder: contains implementation source codes.
- `EmoMosaic-base/` folder: contains the best checkpoint and evaluation results of the *EmoMosaic-base* model.
- `EmoMosaic-large/` folder: contains the best checkpoint and evaluation results of the *EmoMosaic-large* model.
- `thesis_source/` folder: contains the LaTeXsource code of the thesis.
- `thesis.pdf`: contains the text of the thesis in PDF format.