People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

Yahia Fares University of Medea
Faculty of Sciences
Department of Mathematics and Computer Science

# Final Master Project

**Submitted for Obtaining the Master's Degree in Computer Science**
**Speciality: Applied Artificial Intelligence**

---

# Fine-Grained Emotion Recognition from Arabic and English Social Media Texts

---

**Prepared and Presented by:**
BENYEKHLEF HALIMA
BENABDI RANIA

**Supervisor :**
Dr. BOUDOUKHANI NESRINE

**Academic Year: 2025–2026**

# Rania's Dedication

I begin by expressing my gratitude and praise to **Allah**, the Almighty, for His countless blessings and for granting me the strength, determination, and patience throughout my academic journey, enabling me to complete this work, which I hope will be a blessed outcome of His grace and guidance.

I would like to extend my sincere appreciation and respect to my supervisor, **Mrs. Boudoukhani Nesrine**, for her invaluable guidance, continuous support, and insightful advice throughout this journey. Her dedication, patience, and attention to every detail, regardless of how small, greatly contributed to the completion of this work. I am truly grateful for her encouragement and for always being available when I needed guidance.

I would also like to dedicate this work to my beloved **father**, who is no longer living. I wish he could see this moment and be proud of my accomplishments. May God have mercy on his soul and grant him eternal peace. His memory has always been a source of motivation and strength for me.

I would like to thank **my mother**, who played a major role in my success. Her unconditional love, sacrifice, encouragement, and constant support helped me reach this rank. She has always believed in me, even during difficult times, and I will always be grateful for everything she has done for me.

I am also deeply thankful to my elder sister, **Fella**, who has been a second mother to me. She has always supported me, given me valuable advice, and stood by my side whenever I needed help or guidance.

My heartfelt appreciation also goes to my younger sister **Ikram** and my brothers **Oussama** and **Chems Eddine** for their encouragement, emotional support and motivation throughout my studies. Their presence in my life has always been a source of great strength.

Finally, I would like to thank all my family members and everyone who contributed, directly or indirectly, to my academic journey. Their support was essential in helping me to reach this stage.

# Halima's Dedication

In the Name of Allah, the Most Gracious, the Most Merciful. All praise and gratitude belong to Him alone, who illuminated my path with faith, granted me perseverance in times of hardship, and blessed me with the strength and patience to complete this work. I pray that this humble effort is accepted as a blessed deed and a source of His divine pleasure.

I would like to extend my sincere appreciation and respect to my supervisor, **Mrs. Boudoukhani Nesrine**, for her invaluable guidance, continuous support, and insightful advice throughout this journey. Her dedication, patience, and attention to every detail, regardless of how small, greatly contributed to the completion of this work. I am truly grateful for her encouragement and for always being available when I needed guidance.

I would also like to dedicate this work to the soul of my beloved **father**, who is no longer living. I wish he could see this moment and be proud of my accomplishments. May Allah have mercy on his soul and grant him eternal peace. His memory has always been a source of motivation and strength for me.

I would like to thank my dear **mother**, who has been and still is my true support in life. She played a major role in my success through her unconditional love, sacrifice, and constant encouragement. I pray to Allah Almighty to grant her a speedy recovery and to bless her with health and well-being. She has always believed in me, even during difficult times, and I will always be grateful for everything she has done for me.

I am also deeply thankful to my elder sister, **Khadidja**, who has been a great support and help in all circumstances. She has always supported me, given me valuable advice, and stood by my side whenever I needed help or guidance.

My heartfelt appreciation also goes to my brothers: **Mohamed** and **Taki Eddine**, for their encouragement, emotional support, and motivation throughout my studies. Their presence in my life has always been a source of great strength.

Finally, I would like to thank all my family members and everyone who contributed, directly or indirectly, to my academic journey. Their support was essential in helping me reach this stage.

# Abstract

Emotion recognition in social media using (NLP) is a critical step toward understanding human affect and mental states in digital communications. The present study aims to address fine-grained emotion recognition in English and Arabic user-generated contents, involving multi-class and multi-label emotion classification. This study introduces a comprehensive comparative analysis of multiple approaches based on (DL), Transfer Learning (TL), Transformers, and attention mechanism. Whereby, various combinations of transformer-based models with BiLSTM/BiGRU layers and self-attention mechanism are rigorously trained and tested to assess their performance on fine-grained emotion classification using 3 publicly available datasets. These include English-specific transformers such as BERT and RoBERTa and Arabic-specific transformers such as AraBERT, MarBERT, and CamelBERT. All of the proposed models are evaluated based on macro-average and per-class scores of accuracy, recall, precision, AUC, and F1-score metrics using ISEAR, CBET, and SemEval2018 datasets. Experimental results demonstrated the high effectiveness of fine-tuning hybrid transformer-based models for fine-grained emotion classification across Arabic and English social media texts by achieving 0.7308, 0.6482, and 0.6251 macro-average f1-scores on ISEAR, CBET, and SemEval2018 datasets, respectively.

**Keywords:** Fine-Grained Emotion Recognition, Transformers, Deep Learning, Social Media.

# Résumé

La reconnaissance des émotions dans les réseaux sociaux en utilisant le traitement du langage naturel (TALN) est une étape critique vers la compréhension de l'affect humain et des états mentaux dans les communications numériques. La présente étude vise à aborder la reconnaissance fine des émotions dans des contenus générés par les utilisateurs en anglais et en arabe, impliquant la classification multi-classe et multi-étiquette des émotions.Cette étude introduit une analyse comparative complète de multiples approches baséessur l'apprentissage profond (DL), l'apprentissage par transfert (TL), les transformeurs,et le mécanisme d'attention. Ainsi, diverses combinaisons de modèles basés sur des transformeurs avec des couches BiLSTM/BiGRU et un mécanisme d'auto-attention sont rigoureusement entraînées et testées pour évaluer leurs performances sur la classi- fication fine des émotions en utilisant 3 ensembles de données disponibles publiquement. Ceux-ci incluent des transformeurs spécifiques à l'anglais tels que BERT et RoBERTa et des transformeurs spécifiques à l'arabe tels que AraBERT, MarBERT, et Camel- BERT. Tous les modèles proposés sont évalués sur la base de la macro-moyenne et des scores par classe de l'exactitude, du rappel, de la précision, de l'AUC, et du score F1 en utilisant les ensembles de données ISEAR, CBET, et SemEval2018. Les résultats expérimentaux ont démontré la haute efficacité du réglage fin des modèles hybrides basés sur des transformeurs pour la classification fine des émotions à travers les textes des réseaux sociaux en arabe et en anglais en atteignant 0.7308, 0.6482 , et 0.6251 de scores F1 en macro-moyenne sur les ensembles de données ISEAR, CBET, et SemEval2018, respectivement.

**Mots-clés :** Reconnaissance fine des émotions,Transformeurs, Apprentissage profond, Réseaux sociaux.

# الملخص

تحليل المشاعر في وسائل التواصل الاجتماعي باستخدام معالجةالتعرّف على العواطف في وسائل التواصل الاجتماعي باستخدام معالجة اللغة الطبيعية (NLP) يُعدّ خطوة حاسمة نحو فهم المشاعر الإنسانية والحالات الذهنية في الاتصالات الرقمية. تهدف الدراسة الحالية إلى معالجة التعرّف الدقيق على العواطف (Fine-Grained Emotion Recognition) في المحتوى الذي ينشئه المستخدمون باللغتين الإنجليزية والعربية، بما يشمل تصنيف العواطف متعدد الفئات (Multi-Class)ومتعدد التسميات (Multi-Label). تقدّم هذه الدراسة تحليلاً مقارناً شاملاً لعدة مقاربات تعتمد على التعلّم العميق (DL)، والتعلّم بالنقل (TL)، والمحولات (Transformers)، وآلية الانتباه (Attention Mechanism). حيث يتم تدريب واختبار مجموعات مختلفة من النماذج المعتمدة على المحولات مع طبقات BiLSTM/BiGRU وآلية الانتباه الذاتي (Sefl-Attention) بشكل دقيق لتقييم أدائها في تصنيف العواطف الدقيق باستخدام 3 مجموعات بيانات متاحة للجمهور. وتشمل هذه النماذج محولات مخصّصة للغة الإنجليزية مثل BERT وRoBERTa، ومحولات مخصّصة للغة العربية مثل AraBERT وMarBERT وCamelBERT. تم تقييم جميع النماذج المقترحة استناداً إلى المتوسط الكلي (Macro-Average) ودرجات كل فئة على حدة لمقاييس الدقة (Accuracy)، والاسترجاع (Recall)، والإحكام (Precision)، ومنحنى المساحة تحت المنحنى (AUC)، ومقياس F1 باستخدام مجموعات البيانات ISEAR وCBET وSemEval2018. أظهرت النتائج التجريبية الفعالية العالية لعملية الضبط الدقيق (Fine-Tuning) للنماذج الهجينة المعتمدة على المحولات في تصنيف العواطف الدقيق عبر النصوص العربية والإنجليزية في وسائل التواصل الاجتماعي، حيث حققت متوسطات كلية لمقياس F1 بلغت 0.7308 و0.6482 و0.6251 على مجموعات البيانات ISEAR وCBET وSemEval2018 على التوالي.

**الكلمات المفتاحية:** التعرّف الدقيق على العواطف، المحوّلات، التعلم العميق، وسائل التواصل الاجتماعي.

# Contents

**Contents** i

**List of Tables** ii

**List of Figures** iii

**Acronyms** vi

**General Introduction** 1

**1 Emotion Recognition in Social Media** 3
- 1.1 Introduction . . . 3
- 1.2 What is the Relationship Between Sentiment Analysis, Opinion Mining, and Emotion Recognition in NLP? . . . 4
- 1.3 Definition of Emotion Recognition . . . 5
  - 1.3.1 Ekman's Six Basic Emotions . . . 5
  - 1.3.2 Plutchik's Eight Basic Emotions . . . 5
- 1.4 What Are the Applications and Importance of Emotion Recognition ? . 6
- 1.5 How is emotion recognition used in social media analysis? . . . 6
- 1.6 What are the main applications of emotion recognition in social media using NLP techniques? . . . 7
- 1.7 What are the challenges of NLP-based emotion recognition in social media? . . . 7
- 1.8 Conclusion . . . 9

**2 Literature Review** 11
- 2.1 Introduction . . . 11
- 2.2 Machine Learning . . . 11
  - 2.2.1 Support Vector Machine (SVM) . . . 11
  - 2.2.2 Logistic Regression . . . 12
  - 2.2.3 Random Forests . . . 12
  - 2.2.4 XGBoost . . . 12
- 2.3 Deep Learning . . . 12
  - 2.3.1 Multilayer Perceptron (MLP) . . . 12
  - 2.3.2 Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) . . . 13
  - 2.3.3 Gated Recurrent Unit (GRU) and Bidirectional Gated Recurrent Unit (BiGRU) . . . 14
- 2.4 FastText Static Word Embeddings . . . 15

2.5 Transformer -based Models . . . 15
2.5.1 Bidirectional Encoder Representations from Transformers(BERT) . . . 16
2.5.2 Robustly Optimized BERT Approach (RoBERTa) . . . 17
2.5.3 Arabic BERT (AraBERT) . . . 18
2.5.4 Arabic Pre-trained Language Models (CAMeLBERT) . . . 18
2.5.5 Cross-lingual Language Model:(XLM-RoBERTa) . . . 18
2.6 Related Work . . . 19
2.6.1 Textual Emotion Recognition Datasets . . . 19
2.6.2 Fine-Grained Emotion Recognition in English Texts . . . 19
2.6.3 Fine-Grained Emotion Recognition in Arabic Texts . . . 22
2.7 Conclusion . . . 23

**3 Research and Methodology 25**
3.1 Introduction . . . 25
3.2 Problem Statement . . . 25
3.3 Motivation . . . 26
3.4 Proposed Approaches . . . 27
3.4.1 Proposed Models for Emotion Recognition in English Texts . . . 27
3.4.2 Proposed Models for Emotion Recognition in Arabic Texts . . . 31
3.5 Performance Evaluation Metrics . . . 35
3.6 Conclusion . . . 36

**4 Results and discussion 38**
4.1 Introduction . . . 38
4.2 Experimental Study . . . 38
4.2.1 Overview . . . 38
4.2.2 Experimental Environment . . . 40
4.2.3 Training and Evaluation Datasets . . . 41
4.2.4 Emotion Label Distribution Across Three Datasets . . . 42
4.2.5 Training,validation,and test sets . . . 43
4.2.6 Hyperparameters . . . 44
4.3 Results . . . 45
4.3.1 Multi-Class Emotion Classification of English Texts:ISEAR Dataset 45
4.3.2 Global Discussion of Multi-Class Emotion Recognition Approaches on the ISEAR Dataset . . . 50
4.3.3 Multi-Label Emotion Classification of English Texts: CBET Dataset . . . 53
4.3.4 Global Discussion of Multi-Label Emotion Recognition Approaches on the CBET Dataset . . . 58
4.3.5 Multi-Label Emotion Classification of Arabic Texts : SemEval-2018 Dataset . . . 59
4.3.6 Global Discussion of Multi-Label Emotion Recognition Approaches on the SemEval-2018 Dataset . . . 63
4.3.7 Comparison with Literature . . . 65

**General Conclusion 67**

**Bibliography 69**

# List of Tables


2.1 Emotion datasets. . . . . 19
2.2 Summary Table of Related Work on English Datasets. . . . . 21
2.3 Summary Table of Related Work on Arabic Datasets. . . . . 22

4.1 Experimental environment specifications. . . . . 40
4.2 Detailed description of training and evaluation datasets. . . . . 41
4.3 Description of training , validation, and test sets. . . . . 43
4.4 Summarizesall the hyperparameters used for training the proposed models. 44
4.5 Macro-average scores obtained by Approach 1. . . . . 45
4.6 Per-class performance of RoBerta on ISEAR dataset. . . . . 45
4.7 Macro-average scores obtained with Approach 2 on ISEAR dataset. . . 46
4.8 Per-class scores of FastText + BiGRU + Attention on ISEAR dataset. 47
4.9 Macro-average scores obtained by Approach 3 on ISEAR dataset. . . . 47
4.10 Per-class scores of RoBERTa + BiGRU on ISEAR dataset. . . . . 48
4.11 Macro-average scores obtained by Approach 4 on ISEAR dataset. . . . 49
4.12 Performance of SVM using RoBERTa (GoEmotions) CLS embeddings. 49
4.13 Macro-average scores obtained by approach 1 on CBET dataset. . . . 53
4.14 Per-label performance of RoBERTa on CBET dataset. . . . . 53
4.15 Macro-average scores obtained with Approach 2 on CBET dataset. . . 54
4.16 Per-label scores of FastText + BiLSTM on CBET dataset. . . . . 54
4.17 Macro-average scores obtained by Approach 3 on CBET dataset. . . . 55
4.18 Per-label scores of RoBERTa + BiLSTM + MLP on CBET dataset. . . 56
4.19 Macro-average scores obtained by Approach 4 on CBET dataset. . . . 57
4.20 Performance of XGBoost using BERT (GoEmotions) CLS embeddings 57
4.21 Macro-average scores obtained by approach 1 on SemEval-2018 dataset. 60
4.22 Per-label performance of MarBERT on SemEal-2018 dataset. . . . . 60
4.23 Macro-average scores obtained with Approach 2 on SemEval-2018 dataset. 61
4.24 Per-label scores of AraBERT (Not Trainable) + BiLSTM + Self-Attention on SemEval-2018 dataset. . . . . 62
4.25 Macro-average scores obtained by Approach 3 on SemEval-2018 dataset. 63
4.26 Per-label scores MarBERT (Trainable) + BiGRU + Self-Attention on SemEval-2018 dataset. . . . . 63
4.27 Comparison with existing literature . . . . . 65

# List of Figures


1.1 Taxonomy of sentiment analysis tasks. . . . . 5

2.1 MLP architecture . . . . . 13
2.2 LSTM cell architecture (left) and Bidirectional LSTM (BiLSTM) architecture (right) . . . . . 14
2.3 Bidirectional GRU (BiGRU) architecture processing input sequences in both forward and backward directions . . . . . 15
2.4 Transformer architecture . . . . . 16
2.5 BERT general architecture . . . . . 17
2.6 RoBERTa architecture . . . . . 18
2.7 Arabic word embedding for AraBERT . . . . . 18

3.1 Fine-tuning Transformer-based Models. . . . . 27
3.2 Static Embedding with BiLSTM/BiGRU and Self Attention. . . . . 28
3.3 BERT/RoBERTa Contextual Word Embeddings with BiLSTM/BiGRU and Self Attention . . . . . 29
3.4 Fine-tuning Hybrid Transformer-based and BiLSTM/BiGRU and Self Attention. . . . . 30
3.5 Transfer Learning from GoEmotion + ML/DL Models. . . . . 31
3.6 Fine-tuning Transformer-based Models. . . . . 32
3.7 Static Embedding with BiLSTM/BiGRU and Self Attention . . . . . 33
3.8 AraBERT/MarBERT/CAMeLBERT Contextual Word Embeddings BiLSTM/BiGRU and Self Attention . . . . . 34
3.9 Fine-tuning Hybrid Transformer-based and BiLSTM/BiGRU and Self Attention. . . . . 35

4.1 Detailed Plan of the Proposed Approaches for Emotion Recognition on English Datasets. . . . . 39
4.2 Detailed Plan of the Proposed Approaches for Emotion Recognition on Arabic Dataset . . . . . 40
4.3 Distribution of emotion labels in the ISEAR dataset. . . . . 42
4.4 Distribution of emotion labels in the CBET dataset. . . . . 42
4.5 Distribution of emotion labels in the dataset. . . . . 43
4.6 Comparative performance of all four approaches: Macro F1 and AUC. . 50
4.7 Per-class F1 scores . . . . . 51
4.8 Confusion matrix of the best-performing model . . . . . 52
4.9 Comparative performance of all four approaches: Macro F1 and AUC. . 58
4.10 Per-class F1 scores across of approaches 1,2,3 and 4 on CBET dataset . 59

4.11 Macro F1 and AUC scores for Fine-tuned Transformers, Frozen Embeddings, and Hybrid Trainable approaches on the SemEval-2018 dataset. 64
4.12 Class-wise F1-scores for MARBERT FT (A1), AraBERT+BiLSTM+Att (A2), and MarBERT+BiGRU+Att (A3) on the SemEval dataset. . . . 65

# Acronyms

| | |
|---|---|
| AI | Artificial Intelligence |
| AraBERT | Arabic BERT |
| AUC | Area Under the ROC Curve |
| BERT | Bidirectional Encoder Representations from Transformers |
| BiGRU | Bidirectional Gated Recurrent Unit |
| BiLSTM | Bidirectional Long Short-Term Memory |
| CAMeLBERT | Arabic Pre-trained Language Models |
| CBET | Crowdsourced English Tweets Dataset |
| DL | Deep Learning |
| FN | False Negative |
| FP | False Positive |
| GRU | Gated Recurrent Unit |
| ISEAR | International Survey on Emotion Antecedents and Reactions |
| LSTM | Long Short-Term Memory |
| MARBERT | Multi-dialect Arabic BERT |
| MELD | Multimodal Emotion Lines Dataset |
| ML | Machine Learning |
| MLP | Multilayer Perceptron |
| NLP | Natural Language Processing |
| NSP | Next Sentence Prediction |
| OOV | Out-Of-Vocabulary |
| PLM | Pre-trained Language Model |
| RNN | Recurrent Neural Network |
| RoBERTa | Robustly Optimized BERT Approach |
| SemEval | Semantic Evaluation Workshop |
| SVM | Support Vector Machine |
| TN | True Negative |
| TP | True Positive |
| XGBoost | Extreme Gradient Boosting |
| XLM-RoBERTa | Cross-lingual Language Model - RoBERTa |

# General Introduction

The rapid growth of social media platforms has generated an unprecedented amount of textual data. such as X (formerly Twitter), Facebook and Instagram generate millions of user-generated posts every day, resulting in a wealth of social media text that is difficult to process. The textual data has an informal structure, linguistic creativity and contextual variability. These features distinguish social media text from formal writing and make it a unique and complex resource for analysis.

The analysis of emotions from text has gone through three main evolutionary stages over two decades. In the beginning, the analysis was limited to simple polarity classification (positive, negative, neutral), but it failed to capture the complex spectrum of human emotions. The field then evolved toward multi-class classification according to psychological models such as Ekman's basic emotions. Finally, research has advanced to multi-label classification, which acknowledges the reality that human communication often expresses multiple emotions simultaneously (such as anger and fear together). This evolution represents a paradigm shift toward systems that are more ecologically valid and computationally more challenging.

## Problem Statement

A variety of methodological approaches have been proposed for emotion recognition in text, each with inherent limitations. Lexical-based approaches rely on pre-compiled emotion lexicons and dictionaries, where words are associated with specific emotional categories. Although these methods are intuitive and require no training data, they suffer from critical drawbacks: they cannot process out-of-vocabulary words, fail to capture contextual and syntactic nuances, are insensitive to sarcasm and figurative language, and struggle with domain-specific terminology. Traditional machine learning approaches, employing feature extraction techniques such as Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), represent text as sparse, high-dimensional numerical vectors. While these representations enable the use of classifiers such as (SVM) [1] and XGBoost[2], they present significant limitations. The BoW model disregards word order, syntactic structure, and semantic relationships, treating each word as an independent feature, thereby producing sparse and poorly generalizable representations. TF-IDF, while offering improved feature weighting by down-weighting common terms, remains incapable of capturing semantic similarities between words, cannot resolve polysemy or synonymy, and requires extensive manual feature engineering. Consequently, both lexical and traditional machine learning approaches demonstrate limited effectiveness when applied to noisy, informal, and context-dependent social media text.

In recent years , (DL) has become a major paradigm for emotion recognition. Unlike traditional machine learning, which relies on manual feature extraction (e.g., BoW,

TF-IDF), deep learning automatically learns hierarchical representations from raw text using multi-layer neural networks. One of the key techniques in deep learning is word embedding, such as FastText [3], which transforms words into dense vector representations that capture semantic and syntactic relationships while handling out-of-vocabulary words through subword information. When these embeddings are integrated with traditional (ML) classifiers (e.g., SVM, XGBoost), they consistently outperform BoW and TF-IDF baselines. However, the most significant advancements have been achieved through transformer-based architectures. Unlike previous sequential models (RNN, LSTM [4]), transformers process all tokens in parallel and use the self-attention mechanism to weigh the importance of each word relative to others in a sentence, effectively capturing long-range contextual dependencies. Pre-trained transformer models such as BERT[5], RoBERTa[6], AraBERT[7], MARBERT, and CAMeLBERT[8] have established new state-of-the-art benchmarks for emotion recognition. Thanks to these advantages, deep learning, especially transformers and self-attention, is now the methodological standard for Arabic emotion recognition.
Nevertheless, other challenges remain, particularly the limited size of available training datasets and the high similarity between certain emotion categories, both of which continue to hinder model performance and generalization.

# Research Objectives

This study addresses fine-grained emotion recognition in both English and Arabic user-generated texts extracted from social media platforms. Four approaches are proposed and evaluated for multi-class and multi-label emotion recognition using three publicly available datasets: CBET, ISEAR, and SemEval 2018. The proposed approaches involve building DL models based on transformers, transfer learning, and attention mechanisms, including:

Approach 1: Fine-tuning pretrained transformer-based models on target emotion recognition datasets.

Approach 2: Building BiLSTM/BiGRU models with static and contextual word embeddings combined with self-attention.

Approach 3: Building hybrid transformer models combined with BiLSTM/BiGRU layers and self-attention.

Approach 4: Transfer learning from GoEmotions to target emotion recognition datasets.

The aim of this study is to assess the performance of the proposed approaches in order to identify the most robust method that demonstrates high performance across all three datasets and both languages.

# Chapter 1

# Emotion Recognition in Social Media

## 1.1 Introduction

Detection of emotions from social media has become significant within (NLP) research because there is massive growth in user-generated content on various platforms like Twitter, Facebook, Instagram, regular blogs, and online forums. These platforms are rich sources of information; as users express their opinions, feelings, and emotions in real-time on these sites, they become a good resource to analyze people's feelings.

Sentiment analysis is a broad NLP task that aims to identify and interpret opinions and emotions expressed in text. It includes several subtasks, among which sentiment classification is the most common, where texts are classified into predefined sentiment categories such as positive, negative, or neutral.

Emotion recognition goes further than sentiment classification. While sentiment classification identifies the overall sentiment polarity of a text, the goal of emotion recognition is to determine more specific types of emotions, such as happiness, anger, sadness, fear, or surprise.

Emotion recognition systems are being developed using newly developed machine learning and deep learning methods to analyze text data to extract and discover emotional patterns. Social media data is very diverse and unstructured. Typically, such data involves casual language, slang, emojis, sarcasm, and many mixed types of data, making it challenging to detect emotional content.

In the chapter, emotion recognition of social media is introduced from the aspect of NLP, providing a definition of emotion recognition; why it is important; the relationship it has to sentiment analysis, primary applications, and the challenges facing today's emotion recognition systems. The chapter provides enough detail to understand how and why emotions are being recognized and analyzed in a digital environment and why emotion recognition is a growing field in the realm of artificial intelligence.

## 1.2 What is the Relationship Between Sentiment Analysis, Opinion Mining, and Emotion Recognition in NLP?

NLP has two tasks that are similar but also very different: Emotion Recognition and Sentiment Analysis. Sentiment analysis examines the polarity of a body of text (i.e., whether it was written in a positive or negative way). Emotion recognition typically works with other NLP tasks to classify and investigate an opinion from a textual body to give the textual body a better understanding of its contents based on the body due to emotion-based context.

Regarding opinion mining, emotion detection enhances our understanding of user opinions by providing additional information about the emotions associated with each opinion expressed.
The recognition of emotions contributes to the advancement of text classification by providing an automated mechanism that can differentiate between pieces of text containing the same sentiment polarity.
However, these texts may differ according to the strength of the emotion expressed within them. The recognition and detection of emotions enhance applications based on opinion mining by yielding a more complete and thorough analysis. These two methods complement each other. The recognition of emotions conveys to analysts the many different emotional states that are expressed through textual media (for example, happiness, sadness, anger, fear, disgust, and surprise). As such, the two methods expand the use of sentiment analysis (in terms of which emotions are contained in the text) to yield a more thorough and precise examination of the emotions found within the text.
The general sentiment of a piece of text can be analyzed using sentiment analysis, and the emotions of the individuals who produced the text can be analyzed using emotion recognition, thereby enabling both forms of analysis to be combined to provide a comprehensive understanding of the sentiment of a text and sufficient detail about the emotions expressed by the producer of the text. To accomplish this, an analysis of the emotional cues contained in the words in the text, the linguistic style of the texts, and the general context of the texts is often conducted. The combined use of emotion recognition and detection can enhance the value of applications created using NLP by providing accurate, context-sensitive, and meaningful analyses of user-generated data , as illustrated in Figure 1.1 [9].

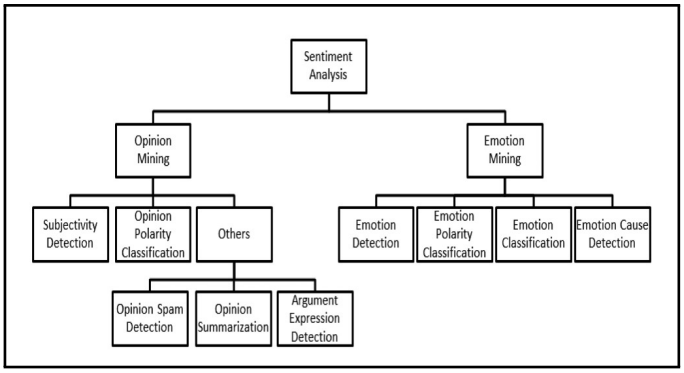



Figure 1.1: Taxonomy of sentiment analysis tasks. [10]

## 1.3 Definition of Emotion Recognition

Emotion recognition (also known as affect recognition) describes an automatic process used by AI techniques to identify and correctly interpret the emotions of people. In the context of text analysis it involves analyzing written materials (e.g. social media post,messages,blogs,chats) for the purpose of recognizing the emotions of users via their written expression. Emotion recognition could also take place via multiple forms such as facial expression, speech signals, physiological reactions, and textual semantics. Emotion recognition consists of detecting, classifying and analyzing emotional indicators to improve the ability of systems to comprehend and respond appropriately to the emotional states of humanity [9, 11].

The study of emotion recognition in NLP is closely connected to psychological theories that define and categorize human emotions. These emotion classification models provide a conceptual framework for identifying and analyzing emotions in textual data. Among the most influential and widely used models in emotion recognition research are Paul Ekman's six basic emotions model and Robert Plutchik's eight basic emotions model.

### 1.3.1 Ekman's Six Basic Emotions

The six emotions that Ekman describes are universal and can be identified from facial expressions and other nonverbal signals across different cultures. These are: Anger, Disgust, Fear, Happiness, Sadness, and Surprise. His research proved that emotions can be described as basic and that everyone expresses them in the same way regardless of the culture in which they live. Because of this, emotions are the basis for many studies and datasets that focus on emotion recognition [12].

### 1.3.2 Plutchik's Eight Basic Emotions

The eight basic emotions identified by Plutchick constitute the very essence of his wheel of emotion. Joy, trust, fear, surprise, sadness, disgust, anger and anticipation represent the eight primary feelings,affect states that we have as human beings; for example: Happiness, as the source of our Joy; Trust, as the basis of our Acceptance and Confidence; Fear, as from our Anxiety,Prejudice,Fear of Failure; Surprise, as a

result of the unexpected; Sadness, from Loss or an overall feeling of Burden; Disgust, from a feeling of Rejection,Distaste; Anger, because of an overwhelming sense of Frustration/Rage; and Anticipation, from a feeling of Curiosity/Quest for the Unknown. Plutchik describes these as Primary Emotions because they are also Universal across all human beings, they are Innate; meaning every human being is born having the ability to feel them, and that they are the Foundation for more complex emotions [13].

## 1.4 What Are the Applications and Importance of Emotion Recognition ?

Emotion recognition can apply to all aspects of life. Emotion recognition helps to understand how a person is feeling, therefore, it is essential. In the healthcare and mental health fields, Emotion Recognition can provide a means to identify emotional states helping to identify early signs of anxiety and depression, thus assisting with interventions and treatments.

In the area of automotive safety, Emotion Recognition can provide monitoring of a person's fatigue and stress and alertness levels while driving. Emotion Recognition defines physiological responses from wearables or in-vehicle sensors, allowing for timely and accurate warnings due to fatigue, stress, or a lack of alertness.

In a workplace, Emotion Recognition helps to improve employee happiness and productivity while providing a safer working environment because identifying emotional states and stress levels can reduce potential risks.

In educational environments, e-learning utilizes Emotion Recognition to assist the adaptive learning systems by providing feedback to students on their emotional states (e.g., boredom, stress, and engagement), providing improved learning opportunities.

In human-computer interactions and human-robot interactions, Emotion Recognition supports the systems (virtual assistants; robots) to change their behaviours in response to the user's emotions [14].

## 1.5 How is emotion recognition used in social media analysis?

Emotion recognition is primarily implemented for analyzing social media content to help identify and analyze the emotional states and opinions received from users of social media. By using (NLP) algorithms to analyze posts, comments, or any user-generated content, emotion recognition allows to identify emotions represented in the text and extract insights about how the user feels/reacts.

The traditional approach has typically included the use of keyword search or bag-of-words models to identify simple emotional cues expressed within text; however, many emotion recognition systems utilize more powerful tools based on semantic analysis and ontologies to provide insight into the context and underlying meaning behind user responses to various posts or comments. This methods allows for greater accuracy when identifying complex emotion patterns that may occur in an individual's social media activity.

By using these technologies, emotion recognition will be able to provide accurate monitoring and tracking of public sentiment, to measure emotional trends over time,

and to analyze users' emotional response to an event, product, or topic in real-time. Additionally, there are numerous benefits associated with this type of analysis across multiple sectors of business and society, including marketing (i.e., how to develop effective marketing strategies through customer engagement), public opinion analysis (i.e., how to determine societal attitudes), and mental health analysis (i.e., identifying potential issues with large-scale emotional patterns that can occur within a particular online community) [9].

## 1.6 What are the main applications of emotion recognition in social media using NLP techniques?

As emotion recognition and sentiment analysis methods are gaining acceptance across many application domains, particularly in social media, these approaches are typically used to determine users' emotional states through the analysis of textual data.

The number of social media messages analyzed through these techniques, which provide insights into how the public feels about events and issues, increases significantly each day. These methods of emotion detection and sentiment analysis can be applied to businesses for understanding customer feedback, reviews, and comments online so that they can evaluate consumer satisfaction and make improvements in the products and services offered to their customers. Additionally, these technologies are being used more widely in mental health-related applications, with emotion-extraction technology being increasingly applied to the analysis of user-generated content to identify potential indicators of stress, anxiety, or depression.

In addition to their application in business settings, emotion recognition and sentiment analysis are also being applied in recommendation systems to personalize the delivery of content based on the emotional state of users, which enhances user engagement and experience. Emotion recognition techniques serve as an important input into the interaction between chatbots and other human-computer systems to create more appropriate and context-sensitive response behavior. Emotion recognition and sentiment analysis techniques are also valuable for opinion mining and crisis detection by monitoring the sentiment trends within social media and identifying out-of-the-normal or emergency situations in real-time[15].

## 1.7 What are the challenges of NLP-based emotion recognition in social media?

Emotion recognition technologies present a number of technical and ethical hurdles in the social media realm. The principal hurdle with emotion recognition technologies is the quality of the data collected, as social media types of content are often highly unstructured, noisy and vaguely defined, which makes it challenging to accurately detect emotion present in a given amount of content. For example, an individual's post may use vague language, lack an adequate amount of previous context or be multimodal in form; all are factors that will present difficulties in determining the emotional state of the user.

Another major avenue of concerns in this area is the matter of privacy and ethics of their usage, as an individual's emotional state is highly personal and private in nature

and many users may not be willing to give their permission for their emotional state to be assessed and analyzed. This creates significant issues related to the protection of data and requires the use of methods to protect individuals' privacy.

The labeling process used in determining emotional expression for emotion recognition models is another significant challenge, as it is heavily based on cultural factors, language factors (both lexical and grammatical), and the individual factor of variance/bias. Therefore, it is difficult to produce reliable and consistent data sets for developers to use in creating models that recognize emotion.
Another obstacle we are facing is that when emotive states are communicated through social media there is frequently some ambiguity. This ambiguity can come from contextual issues or through sarcasm and irony. Thus, interpreting the true emotional connotation of any message is rendered more complicated than it would otherwise be. In addition, the vast majority of existing datasets can only be generalized to particular groups of people, to a single language, or to a single social media platform. Therefore, training an emotion recognition model based on data from one dataset will most likely lead to poor performance when the model is used on another dataset [14].

## 1.8 Conclusion

Emotions can be recognised on social media through (NLP). Emotion Recognition Technologies (ERT) will help us understand humans' emotions at a deeper level than simply classifying them as 'positive' or 'negative'. ERT can analyse text shared by users, enabling companies and researchers to accurately detect many different emotional states from people's writing, allowing them to better understand their actions and opinions.

The potential uses for ERT are many: healthcare, education, advertising, human-computer interaction and analysing human behaviour. However, ERT still faces challenges such as inaccurate data due to 'noise' in the data, privacy issues, subjective labelling of data, not being able to understand things from local contexts, and challenges with managing multilingual and multimodal data.

# Chapter 2

## Literature Review

# Chapter 2

# Literature Review

## 2.1 Introduction

Billions of people share their feelings on a daily basis on the growing number of social media platforms, making it more important than ever to understand human emotion expression through text. This chapter presents a comprehensive background and literature review on text-based emotion recognition with a special focus on English and Arabic social media content.

The chapter begins with the groundwork and context for emotion recognition in (NLP) and affective computing. It defines and clarifies key concepts such as emotion recognition, text classification and the important difference between sentiment analysis and emotion recognition.

Then, the chapter describes the exciting development of emotion recognition systems from classic rule-based systems to state-of-the-art deep learning models with a focus on the Arabic NLP challenges and advancements.

Finally, a thorough review of related work is presented including English emotion recognition, Arabic emotion recognition and multilingual approaches. The chapter concludes by identifying research gaps that motivate the present study and hopefully will inspire future research in this important area.

## 2.2 Machine Learning

Machine learning is the study of algorithms that automatically learn from data by generalizing beyond training examples to make predictions on unseen inputs [16].

### 2.2.1 Support Vector Machine (SVM)

SVM is a training algorithm introduced by Boser, Guyon, and Vapnik (1992) that constructs an optimal separating hyperplane by maximizing the margin between the decision boundary and the training patterns. The solution is expressed as a linear combination of support vectors, which are the training examples closest to the decision boundary. This approach automatically tunes the capacity of the classifier to match the complexity of the problem, and typically only a small subset of training patterns (the support vectors) are required to define the decision function [1].

### 2.2.2 Logistic Regression

Logistic regression is a regression method for binary outcome variables. It models the probability using the logistic function to ensure that the predicted values remain between 0 and 1. The model is linearized through the logit transformation, and parameter estimation is performed using maximum likelihood estimation [17].

### 2.2.3 Random Forests

Random Forests are an ensemble method introduced by Breiman (2001) that combines multiple decision trees. Each tree is grown using an independent random vector $\Theta_k$, and the forest votes for the most popular class. As the number of trees increases, the generalization error converges to a limit without overfitting. The accuracy of a random forest depends on the strength of individual trees and the correlation between them [18].

### 2.2.4 XGBoost

XGBoost is a scalable tree boosting system that implements gradient tree boosting with a regularized learning objective. It builds an ensemble of regression trees in an additive manner using second-order gradient statistics [2].

## 2.3 Deep Learning

Deep learning is a subfield of machine learning based on artificial neural networks with multiple hidden layers. These models learn hierarchical representations of data by passing inputs through successive non-linear transformations. Key advantages of DL include automatic feature extraction, ability to handle large-scale data, and state-of-the-art performance across various NLP tasks [19].

### 2.3.1 Multilayer Perceptron (MLP)

According to Du et al. (2022), a (MLP) is a feedforward neural network with one or more hidden layers between the input and output layers. Thanks to its hierarchical structure and nonlinear activation functions, the MLP can approximate any continuous function with high accuracy. This architecture (see Figure 2.1) is typically trained using the backpropagation algorithm, which iteratively adjusts the network weights to minimize prediction error [20].

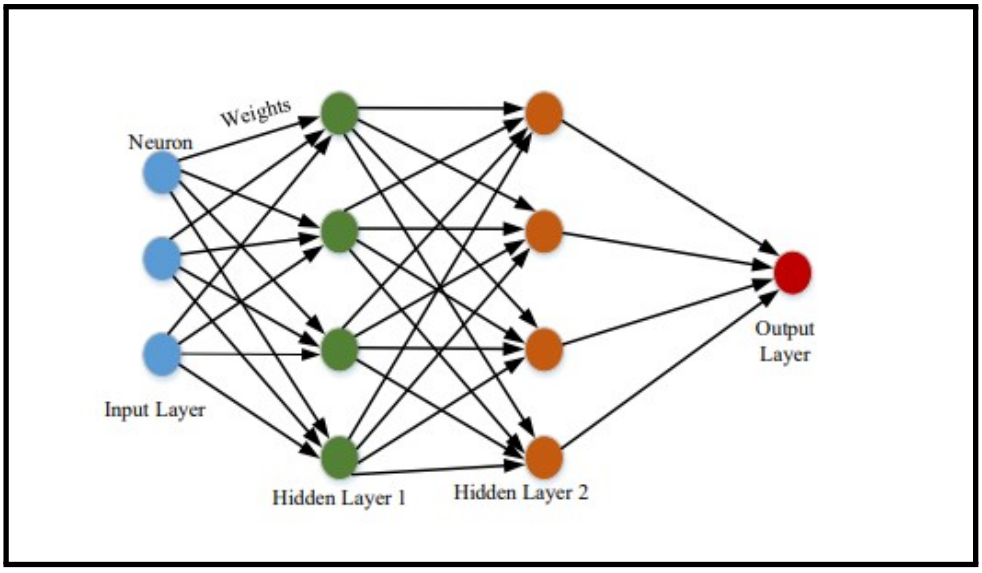



Figure 2.1: MLP architecture
[21].

### 2.3.2 Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM)

LSTM is a recurrent neural network architecture that addresses the vanishing gradient problem through a memory cell and three gating mechanisms (input, forget, and output gates) which regulate information flow to capture long-term dependencies in sequential data. Bidirectional LSTM (BiLSTM) extends this architecture by processing input sequences in both forward and backward directions using two separate LSTM layers, enabling the network to access past and future context simultaneously for improved performance on sequence labeling tasks such as speech and handwriting recognition, as illustrated in Figure 2.2 [4].

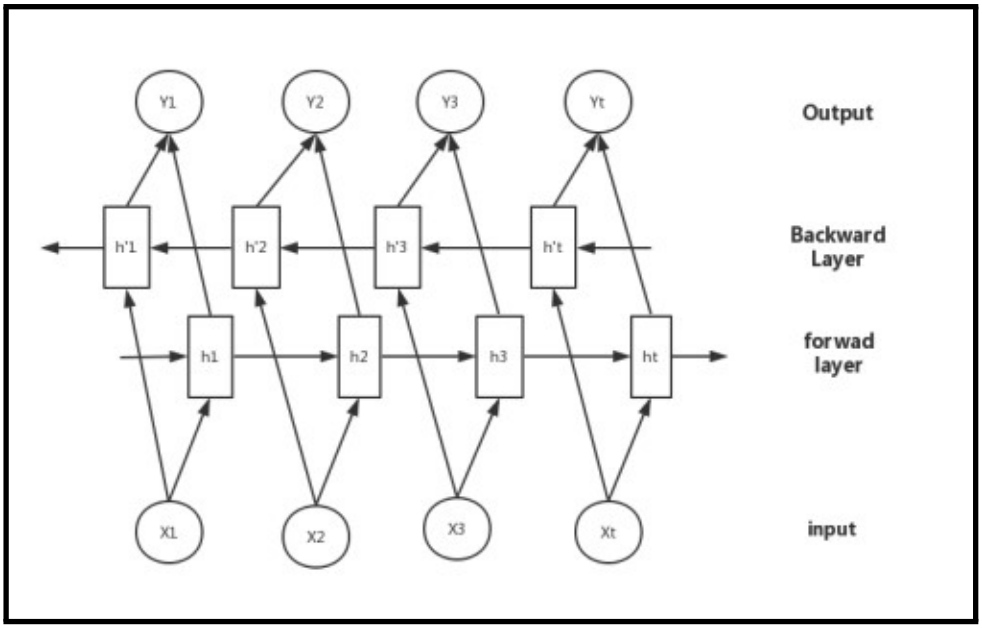



Figure 2.2: LSTM cell architecture (left) and Bidirectional LSTM (BiLSTM) architecture (right)
[22].

### 2.3.3 Gated Recurrent Unit (GRU) and Bidirectional Gated Recurrent Unit (BiGRU)

GRU is a computationally efficient recurrent neural network that uses an update gate and a reset gate to manage long-term dependencies with fewer parameters than LSTM. Extending this, the Bidirectional GRU (BiGRU) processes sequences in both forward and backward directions simultaneously, capturing information from past, current, and future time steps. This bidirectional capability (see Figure 2.3) makes BiGRU especially useful for sequential data such as audio, where context from both temporal directions is valuable [23].

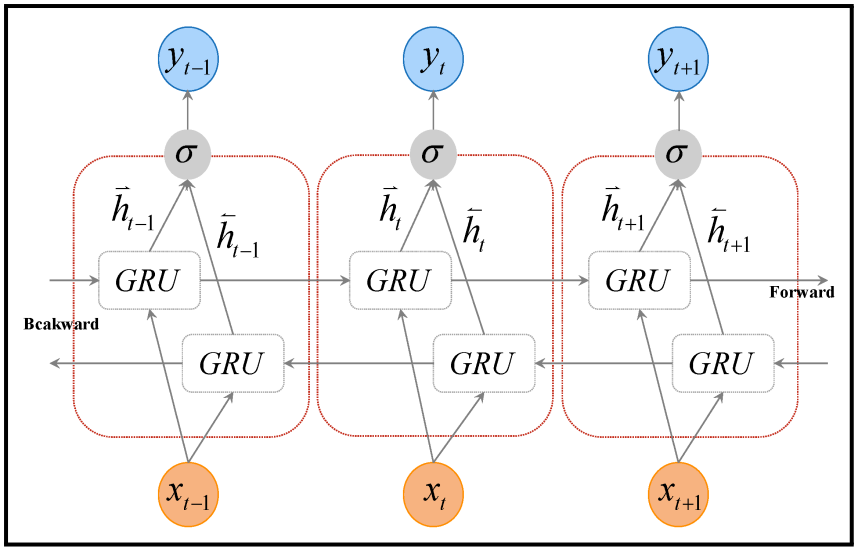



Figure 2.3: Bidirectional GRU (BiGRU) architecture processing input sequences in both forward and backward directions [3].

## 2.4 FastText Static Word Embeddings

fastText is an efficient and lightweight text classification model introduced by Facebook AI Research. The model represents words using distributed vector embeddings, which are combined to generate a representation of the entire text and then passed to a linear classifier for prediction. Designed for speed and scalability, fastText can process massive text corpora with low computational cost while achieving competitive performance compared to more complex deep learning approaches.
Furthermore, it leverages techniques such as word n-grams and feature hashing to effectively manage large output spaces, making it well suited for tasks including sentiment analysis, topic categorization, and tag prediction[24].

## 2.5 Transformer -based Models

The Transformer architecture introduced the self-attention mechanism, which allows models to weigh the importance of different tokens in a sequence regardless of their distance. Unlike RNNs, Transformers process all tokens in parallel, enabling efficient training on large datasets. This architecture (see Figure 2.4) forms the foundation of modern pre-trained language models [25].

In the Transformer, the encoder uses **multi-head self-attention** without masking, where each position can attend to all positions in the input sequence (past and future). The decoder uses **masked multi-head self-attention** which prevents attention to future positions.

The multi-head self-attention mechanism is defined as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

where each attention head is computed as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

and the scaled dot-product attention is given by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

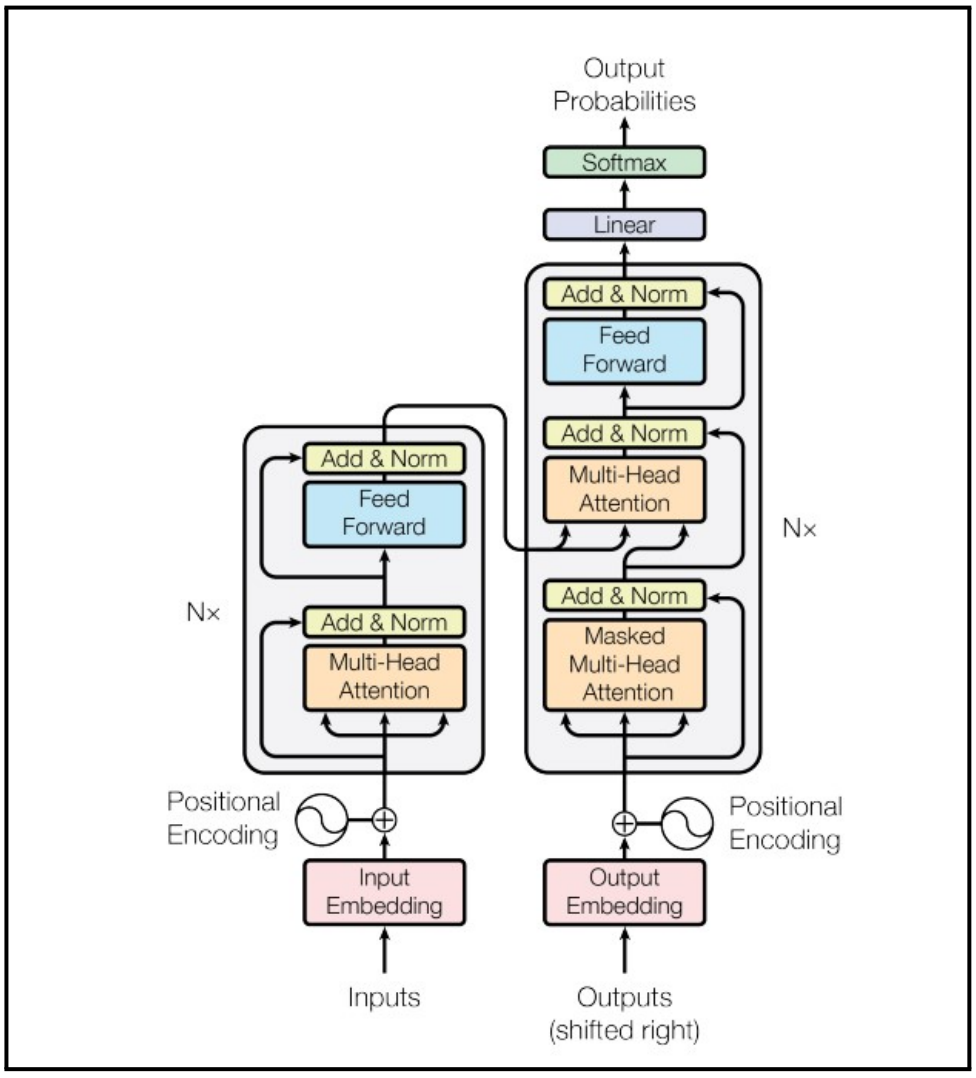



Figure 2.4: Transformer architecture [25].

### 2.5.1 Bidirectional Encoder Representations from Transformers(BERT)

BERT is a pre-trained transformer-based model that reads text bidirectionally to capture context from both left and right, enabling state-of-the-art performance on NLP tasks, as illustrated in Figure 2.5 [5].

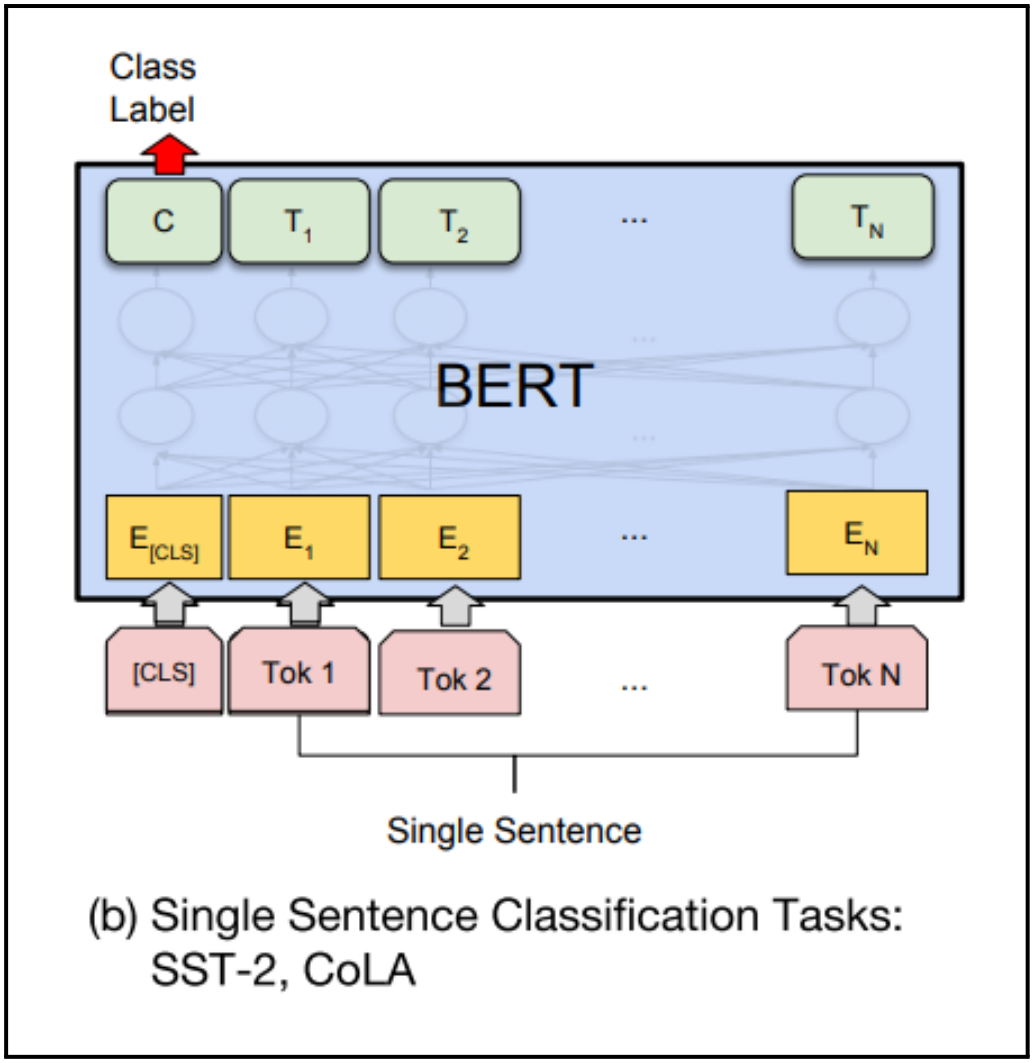



Figure 2.5: BERT general architecture
[5].

### 2.5.2 Robustly Optimized BERT Approach (RoBERTa)

RoBERTa is an optimized variant of BERT proposed by Liu et al. that improves performance by training longer with larger batches and removing NSP objective, as illustrated in Figure 2.6 [6].

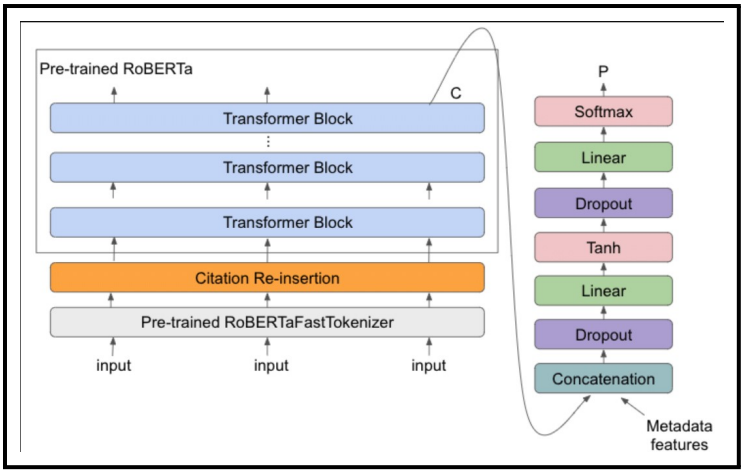



Figure 2.6: RoBERTa architecture [26].

### 2.5.3 Arabic BERT (AraBERT)

AraBERT is a BERT-based transformer model developed by Antoun et al. for Arabic natural language understanding. It is pre-trained on large-scale Arabic corpora and aims to improve performance on downstream NLP tasks by learning contextual representations of Arabic text, as illustrated in Figure 2.7 [7].

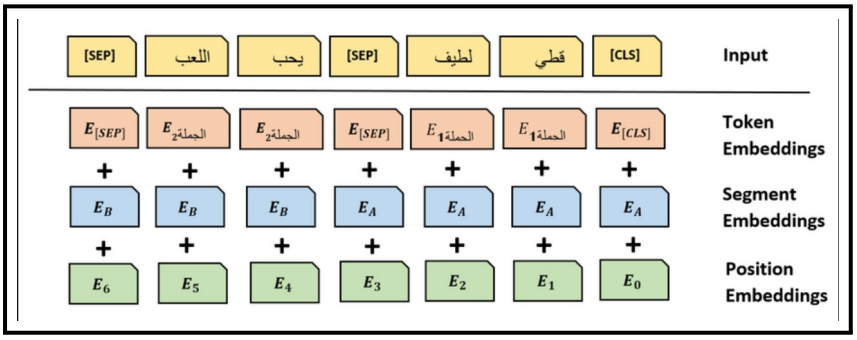



Figure 2.7: Arabic word embedding for AraBERT [27].

### 2.5.4 Arabic Pre-trained Language Models (CAMeLBERT)

CAMeLBERT is a collection of Arabic pre-trained language models developed by the CAMeL Lab at New York University Abu Dhabi. It is designed to systematically study the impact of training data size and language variety (Modern Standard Arabic, Dialectal Arabic, and Classical Arabic) on downstream NLP performance under controlled experimental settings with fixed hyperparameters [28].

### 2.5.5 Cross-lingual Language Model:(XLM-RoBERTa)

XLM-RoBERTa is a multilingual transformer model introduced by Conneau et al. that is pre-trained on 100 languages and demonstrates remarkable zero-shot cross-lingual transfer capabilities [29].

## 2.6 Related Work

### 2.6.1 Textual Emotion Recognition Datasets

In the literature on emotion recognition, a large number of studies rely on datasets collected from social media platforms, such as Twitter, Reddit, and Facebook, due to the abundance and diversity of emotional expressions found in user-generated content. These datasets are generally constructed using either a multi-class or a multi-label scheme. For example, MELD represents a multi-class dataset with 13,708 samples annotated into 7 emotion categories, whereas CBET, collected from Twitter, represents a multi-label dataset with 76,860 samples annotated into 9 emotion categories. Other notable resources include SemEval-2018 (Affect in Tweets), a multi-label Twitter dataset available in both English and Arabic .
Table 2.1presents a summary of some of the most commonly used emotion datasets in the literature, indicating their size, annotation type, and number of emotion classes.

Table 2.1: Emotion datasets.

| **Dataset** | **Size** | **Type** | **Classes** |
|---|---|---|---|
| ISEAR [30] | 7,652 | MultiClass | 7 |
| CBET [31] | 76,860 | MultiLabel | 9 |
| GoEmotions [32] | 58,000 comments | MultiLabel | 27 + Neutral |
| SemEval-2018 [33] | 2863 | MultiLabel | 11 |
| MELD [34] | 13,708 | MultiClass | 7 |

### 2.6.2 Fine-Grained Emotion Recognition in English Texts

Various researches targeted fine-grained emotion recognition by proposing a variety of approaches employing ML and DL models using different public and private datasets. Extensive experiments conducted by [35] aimed to address fine-grained emotion recognition and sentiment analysis in social media textual contents by comparing multiple transformer-based methods with previous DL models on several Twitter datasets. Their experimental results showcased the outperformance of the BERT model across datasets selected for evaluation.

The authors in [36] evaluated three CNN-based models for the task of emotion recognition from textual data using three public datasets ISEAR, CBET, and TEC. The experimental results indicate that the LMT-CNN provides the highest level of global performance on most datasets.

A comparative study between BERT, Bi-LSTM, CNN, and capsule networks is conducted by [37] using ISEAR, SemEval, GoEmotion, and CBET datasets. Empirical results highlighted the superior performance of BERT in terms of accuracy across all the datasets used for training and evaluation.

A novel framework proposed by [38] is composed by combining Hybrid Engineered Features (HEF) and DL Features (DLF) with Bi-directional Long BiLSTM and BERT models. Experiments revealed the outperformance of the hybrid models BiLSTM + HEF + DLF and BERT + HEF + DLF on SemEval2018 and GoEmotions datasets, respectively.

An investigation using the GoEmotion dataset was presented in[39] . The authors begin by reproducing baseline results using models such as BERT to validate their experimental setup. They then explore several enhancement techniques, including data augmentation, transfer learning from related datasets such as CARER, and fine-tuning of more advanced architectures like RoBERTa. The results show that transfer learning from the CARER dataset significantly improves performance, particularly when combined with data augmentation techniques. However, The authors reported limited performance due to sensitivity to class imbalance.

A rigorous comparison conducted by [40] aimed to study the impact of micro, macro, and weighted F1-scores on assessing the performance of fine-grained emotion recognition in a multi-label setting. Their findings revealed that all three variants of F1-score measure perform similarly in a balanced multi-label dataset (XED). However, when an imbalanced dataset (GoEmotion) is used the scores of macro, micro, and weighted averages of F1-score give different values, which indicates the necessity of using all of these variants for a reliable performance assessment.

Two frameworks referred by Hyper Graph Neural Network (HGNN) and Distil-RoBERTav2 were proposed by[41] to resolve multi-label emotion classification in user-generated contents. Experimental results using GoEmotion, Friends, and TEC datasets demonstrated the effectiveness of the proposed frameworks by high f1-scores.

Table 2.2: Summary Table of Related Work on English Datasets.

| Work | Dataset | Size | Model | Acc | Prec | Rec | F1 |
|---|---|---|---|---|---|---|---|
| [35] | CBET [31] | 76,860 tweets | BERT base model | 0.609 | 0.608 | 0.611 | 0.609 |
| | SemEval [33] | — | | 0.672 | 0.669 | 0.680 | 0.670 |
| [36] | CBET [31] | 51,240 (6emotion) | Lexicon-Enhanced Multi-Task CNN(LMT-CNN) | — | 0.6385 | 0.6331 | 0.6284 |
| | ISEAR [30] | 5431 (5emotion) | | — | 0.7300 | 0.7257 | 0.7350 |
| | TEC | 21,051 (6emotion) | | — | 0.5706 | 0.4997 | 0.5220 |
| [37] | CBET [31] | >30k | BERT | 0.56 | — | — | — |
| | GoEmotions [32] | >30k | | 0.60 | — | — | — |
| | ISEAR [30] | — | | 0.63 | — | — | — |
| | SemEval [33] | >30k | | 0.64 | — | — | — |
| | MELD [34] | — | | 0.49 | — | — | — |
| [38] | SemEval-2018 [33] | — | Bi-LSTM + DLF + HEF | 0.684 | — | — | Macro=0.6577 Micro=0.7855 |
| | GoEmotions [32] | — | BERT + DLF + HEF | 0.534 | Macro= 0.54 | Macro= 0.46 | Macro= 0.49 |
| [39] | GoEmotions [32] | 58k | BERT Fine tuning | — | 0.50 | 0.54 | 0.51 |
| | CARER [44] | — | BERT | 0.9375 | 0.9221 | 0.9101 | 0.9158 |
| | GoEmotions + Aug | 43,410 | BERT + ProtAug | — | 0.51 | 0.53 | 0. 51 |
| [40] | GoEmotions [32] | 38,237 | Unigram lexicon | — | — | — | Micro=0.75 Macro=0.66 |
| | XED | 13,682 | | — | — | — | Micro=0.59 Macro=0.58 |
| | GoEmotions Agg. | 38,237 | | — | — | — | Micro=0.78 Macro=0.71 |
| | XED Agg. | 13,682 | | — | — | — | Micro=0.765 Macro=0.668 |
| [41] | GoEmotions [32] | 58,000 | HypergraphNeural Network (HGNN-Min) | 0.737 | 0.752 | 0.737 | 0.738 |
| | TEC | 21,000 | | 0.436 | 0.454 | 0.354 | 0.443 |
| | Friends | 30,000 | DistilRoBERTa-v2 | 0.868 | 0.868 | 0.868 | 0.867 |

### 2.6.3 Fine-Grained Emotion Recognition in Arabic Texts

An in-depth study on fine-grained emotion detection in Arabic texts was presented in [42]. The authors explore several NLP approaches, including classical ML models, DL architectures, and Arabic-specific pre-trained language models (PLMs). The study uses a single dataset, AETD, consisting of 10,065 samples, for both training and evaluation.

Three main categories are examined: ML methods such as SVM and Random Forest, DL models such as BiLSTM and BiGRU, and PLMs such as AraBERT and MarBERT. After evaluation, the best-performing models from each category are integrated into hybrid approaches, including parallel (majority voting and probability averaging) and sequential (boosting-inspired) strategies. The results show that the sequential hybrid model achieves the best performance with an F1-score of 0.802, using a combination of MARBERTv2, Logistic Regression (LR), and BiGRU. This method focuses on hard-to-classify instances, improving overall performance and confirming the effectiveness of sequential hybrid approaches for Arabic emotion detection.

In the framework proposed by [43], an ensemble DL model for emotion detection in Arabic tweets is built by combining three modern models: MarBERT, BiLSTM, and BiGRU networks. The proposed framework processes Arabic texts through multiple stages including preprocessing, word vectorization, parallel processing by the three models, and final weighted ensemble combination to achieve emotion classification.

Evaluated on the SemEval 2018 dataset for multi-label classification involving 11 emotion labels, their proposed approach demonstrates that MARBERT outperforms other Arabic language models, and the ensemble model achieves better results than both individual models and state-of-the-art studies. The research highlights the effectiveness of transfer learning and model ensemble techniques for Arabic language processing, while also identifying dataset imbalance as a limitation that can be addressed in future work.

Table 2.3: Summary Table of Related Work on Arabic Datasets.

| **Work** | **Dataset** | **Size** | **Model** | **Acc** | **Prec** | **Rec** | **F1** |
|---|---|---|---|---|---|---|---|
| **[42]** | AETD [45] | 10,065 | LR | 0.678 | 0.675 | 0.678 | 0.668 |
| | | | BiGRU | 0.741 | 0.760 | 0.741 | 0.742 |
| | | | MARBERTv2 | 0.795 | 0.795 | 0.798 | 0.794 |
| | | | Sequential Hybrid | 0.801 | 0.805 | 0.800 | 0.802 |
| **[43]** | **SemEval -2018 [33]** | 4.381 | MarBERT | — | 0.542 | 0.559 | 0.692 |
| | | | Bi-LSTM | — | 0.522 | 0.559 | 0.653 |
| | | | Bi-GRU | — | 0.599 | 0.543 | 0.664 |
| | | | MARBERT, Bi- LSTM, Bi-GRU | — | 0.634 | 0.550 | 0.701 |

## 2.7 Conclusion

This chapter has provided a comprehensive background and literature review on emotion recognition from text. Key concepts were defined, including the distinction between sentiment analysis and emotion recognition. The evolution from rule-based systems to transformer-based models was traced, with particular attention to Arabic NLP developments.

The review revealed that while significant progress has been made in English emotion recognition, research on Arabic remains limited. Most existing Arabic datasets are small in scale, and few studies have systematically compared multiple transformer-based models across both languages. These gaps motivate the current study, which aims to develop and evaluate emotion recognition systems for both English and Arabic using state-of-the-art transformer models.

# Chapter 3

## Research and Methodology

# Chapter 3

# Research and Methodology

## 3.1 Introduction

This chapter presents the research design employed to study emotion recognition in text with the application of (NLP) techniques. Given the current trend of communication through digital channels, text data generated by users posting on social media, message boards, or review sites can provide insights into human emotions, opinions, and affective states. It is the intent of this research to utilize three datasets to complete emotion classification tasks (two in English and one in Arabic).

Throughout the methodology, text data were transformed from raw text into structured numerical representations suitable for classification. Emotion classification models were created to identify and classify emotions associated with the datasets using a Transformer-based model, along with traditional machine learning and deep learning approaches (e.g., emotions such as joy, sadness, anger, fear, surprise, and disgust).

## 3.2 Problem Statement

Despite recent advances in natural language processing, correctly categorizing emotional expressions in texts remains a major difficulty, especially when the emotional expression is subtle, contextual, or overlaps with other emotions.

In Arabic, this difficulty is worsened by the rich morphology of the language, the variety of dialects in use, and the extreme paucity of datasets with quality annotations on a large scale. These impediments limit the ability to develop robust emotion recognition models that can successfully generalize across different domains and languages.

In particular, the following key subproblems are addressed in this study:

- **Multiple and ambiguous emotions:** A single text expression often conveys more than one emotion simultaneously, and the boundaries between emotions are frequently fuzzy rather than clear-cut.
- **Small datasets:** Although emotion-annotated corpora exist, their size is often too small to support robust deep learning. This lack of sufficient scale, along with limited diversity and public accessibility, creates a major bottleneck in model training and evaluation.
- **Class imbalance:** Most emotive datasets reflect an abundance of certain emotions, such as joy and anger, in comparison to others, such as disgust and surprise.

Consequently, models based on these datasets produce predictions that are biased by the imbalanced nature of the dataset.

- **Similarity/overlapping emotions:** The emotions of sadness and disappointment overlap to a degree, as do anger and frustration, and fear and anxiety, making it difficult for human annotators to make subtle distinctions between these emotions because of their close proximity on the semantic differential scale.

## 3.3 Motivation

The challenges related to the classification of emotions (including ambiguous, overlapping or unbalanced expressions) cannot be solved by one model/technique alone; they require a combination of multiple complementary strategies which will include:

### – Text Representation Methods: FastText and Contextual Embeddings

FastText was designed to generate embeddings for all words within the vocabulary using both character-level n-grams and traditional word-based representations, ensuring robustness against the occurrence of (OOV) words. Consequently, the model encodes information about word roots and morphological patterns through subword modeling, which improves its ability to represent previously unseen words and enhances overall performance. In contrast, contextual word embeddings capture the meaning of words depending on their surrounding context, providing a more dynamic representation of semantic variations.

### – Machine and Deep Learning models

The combination of traditional machine learning classifiers and deep learning architectures allows for a comparative evaluation of performance. While deep learning excels at capturing complex patterns, ML models offer efficiency and interpretability, particularly with smaller datasets.

### – Attention mechanisms

Attention allows models to dynamically focus on the words that contribute most to predicting the emotion being expressed by an input text. The result of this capability will increase both the accuracy and interpretability of emotion recognition models.

### – Transformer-based architectures

Transformers use parallel processing of the entire sequential text rather than processing one word at a time, as in recurrent models. The ability of transformer models to recognize long-range dependencies and contextual relationships between words gives them an advantage over recurrent models when interpreting subtle emotions in a sentence.

## 3.4 Proposed Approaches

### 3.4.1 Proposed Models for Emotion Recognition in English Texts

#### Approach 1: Fine-tuning Transformer-based Models

The first approach relies on fine-tuning pre-trained Transformer models, namely BERT and RoBERTa. The [CLS] output of the BERT/RoBERTa model (see Figure 3.1) is forwarded to an output dense layer using a sigmoid activation for multi-label classification and a Softmax activation for multi-class classification.

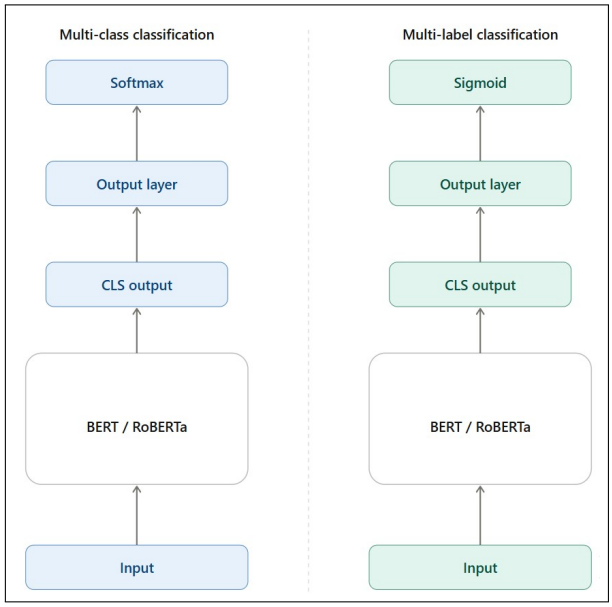



Figure 3.1: Fine-tuning Transformer-based Models.

#### Approach 2 : Static Vs.Contextual Word Embedding with BiGRU/BiLSTM and Self-attention

#### Part 1: FastText Embedding with BiLSTM/BiGRU and Self Attention

The aim of this approach is to evaluated the performance of static word embedding compared to contextual Word embedding for Fine-Grained Emotion Recognition when combined with BiLSTM/BiGRU and Self-attention attention.

In this study, fastext embedding is selected as static word representations due to their ability to handle OOV words generally found in user-generated contents like social Media platforms. On the other hand, Bert and Roberta word embedding are selected as contextual Word embedding as they can represent words' meaning depending on their context in the input sentence.

The first models based on Fastext embedding are illustrated in figure 3.2 below. Whereby, the output of the FastText layers is then passed through one of two deep learning word embedding architectures (BiGRU) or (BiLSTM) followed by final a dense layer with a sigmoid activation function for multi-label classification and a softmax activation function for multi-class classification.

***Models Without Self-Attention***

- **FastText + BiGRU without Self-Attention**
  - Global max pooling for multi-label Classification
  - Mean pooling for multi-class Classification
- **FastText + BiLSTM without Self-Attention**
  - Global max pooling for multi-Classification
  - Mean pooling for multi-class Classification

***Models With Self-Attention***

- **FastText + BiGRU + Self-Attention**
- **FastText + BiLSTM +Self-Attention**

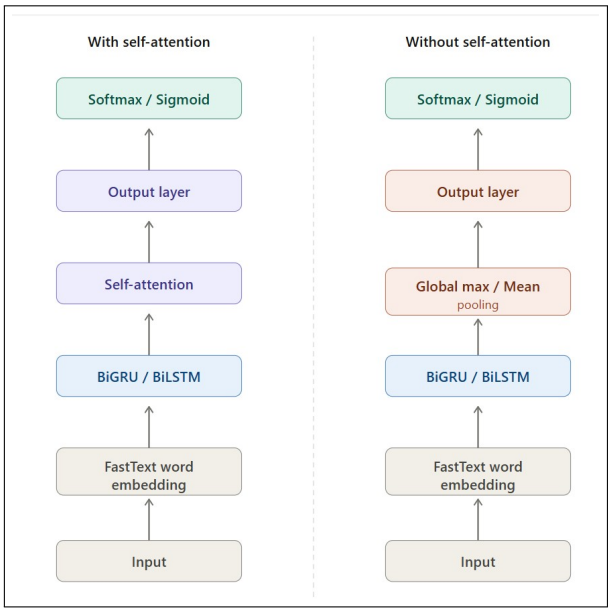



Figure 3.2: Static Embedding with BiLSTM/BiGRU and Self Attention.

### Part 2: BERT / RoBERTa Word Embedding with BiLSTM/BiGRU and Self Attention

In this models,frozen (not trainable) contextual embeddings from BERT and RoBERTa, are integrated with BiGRU/BiLSTM layers , with self-attention. as illustrated in figure 3.3,the output of this context-based word embedding layer serves as an input to either a (BiGRU) or (BiLSTM) followed by final a dense layer with a sigmoid activation function for multi-label classification and a softmax activation function for multi-class classification.

***Models Without Self-Attention***

- **BERT/Roberta (not trainable) + BiLSTM without Self-Attention**
  - Global max pooling for multi-label Classification
  - Mean pooling for multi-class Classification
- **BERT/Roberta (not trainable) + BiGRU without Self-Attention**
  - Global max pooling for multi-label Classification
  - Mean pooling for multi-class Classification

***Models With Self-Attention***

- **BERT/Roberta (not trainable) + BiGRU + Self-Attention**
- **BERT /Roberta(not trainable) + BiLSTM + Self-Attention**

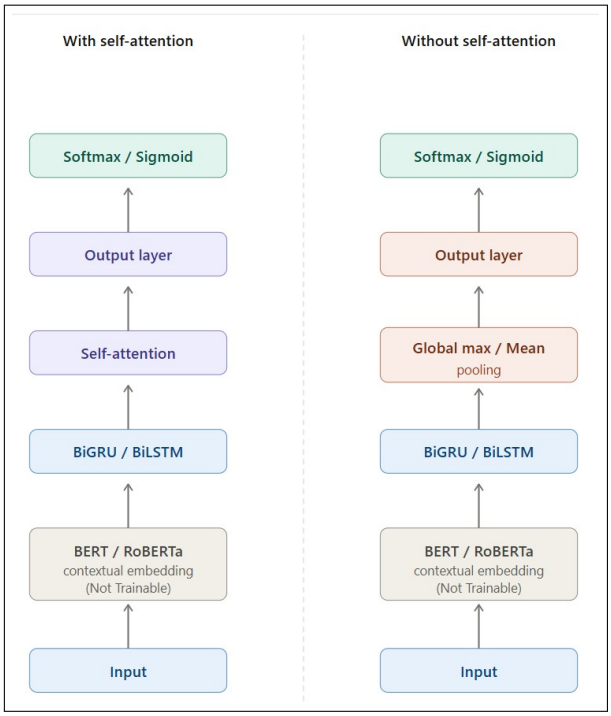



Figure 3.3: BERT/RoBERTa Contextual Word Embeddings with BiLSTM/BiGRU and Self Attention .

### Approach 3: Fine-tuning Hybrid Transformer-based Models with BiLSTM/BiGRU and Self Attention

In this approach , we fine-tune trainable BERT and RoBERTa models jointly with Bi-GRU/BiLSTM layers and self-attention.As illustrated in the figure 3.4), This transformer-based model layer's output [CLS] becomes input for a (BIGRU) or (BILSTM) followed by final a dense layer. This final layer will consists of either a sigmoid activation function for multi-label classification, or a softmax activation function for multi-class classification.

### BERT/Roberta-based Models

#### *Models Without Self-Attention*

- **BERT/Roberta (trainable) + BiLSTM without Self-Attention**
  - Global max pooling for multi-label Classification
  - Mean pooling for multi-class Classification
  - **MLP for multi-label Classification only**
- **BERT/Roberta(trainable) + BiGRU without Self-Attention**
  - Global max pooling for multi-label dataset
  - Mean pooling for multi-class dataset
  - **MLP for multi-label dataset only**

#### *Models With Self-Attention*

- **BERT/Roberta (trainable) + BiGRU +Self-Attention**
- **BERT/Roberta (trainable) + BiLSTM + Self-Attention**

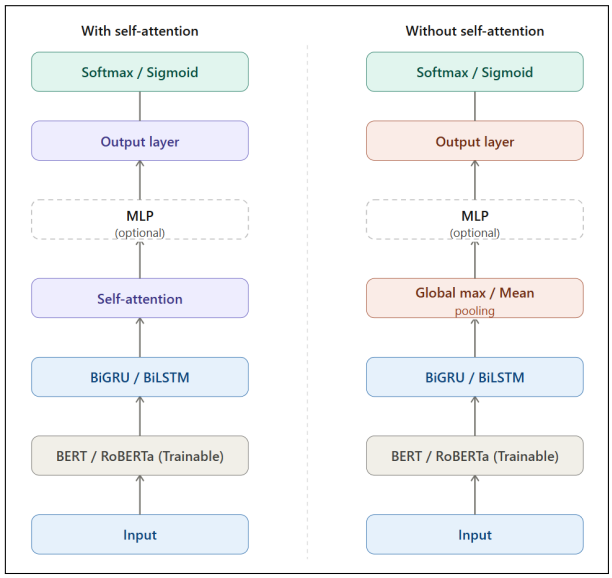



Figure 3.4: Fine-tuning Hybrid Transformer-based and BiLSTM/BiGRU and Self Attention.

### Approach 4: Transfer Learning from GoEmotion + ML/DL Models

In this approach, BERT and Roberta models pretrained on GoEmotion for assigning textual inputs 27 emotions labels, are used to convert textual inputs into feature vectors. Whereby, the [CLS] output of these pretrained models is employed as a feature representation of the input text and is then fed into ML and DL models for classification (see figure 3.5). A set of ML and DL models is selected in this study to be used

as classifiers on top of the feature representation extracted by the transformer models pretrained on GoEmotion. These include ML models like SVM, Logistic Regression, Random Forest, and XGBoost, in addition to DL models such as BiLSTM, BiGRU, and MLP.

**BERT/RoBERTa pretrained on GoEmotion for feature extraction + ML Models**

- **BERT/RoBERTa GoEmotion CLS output + SVM**
- **BERT/RoBERTa GoEmotion CLS output + Logistic Regression**
- **BERT/RoBERTa GoEmotion CLS output + Random Forest**
- **BERT/RoBERTa GoEmotion CLS output + XGBoost**

**+ DL Models**

- **BERT/RoBERTa GoEmotion CLS output + BiLSTM**
- **BERT/RoBERTa GoEmotion CLS output + BiGRU**
- **BERT/RoBERTa GoEmotion CLS output + MLP**

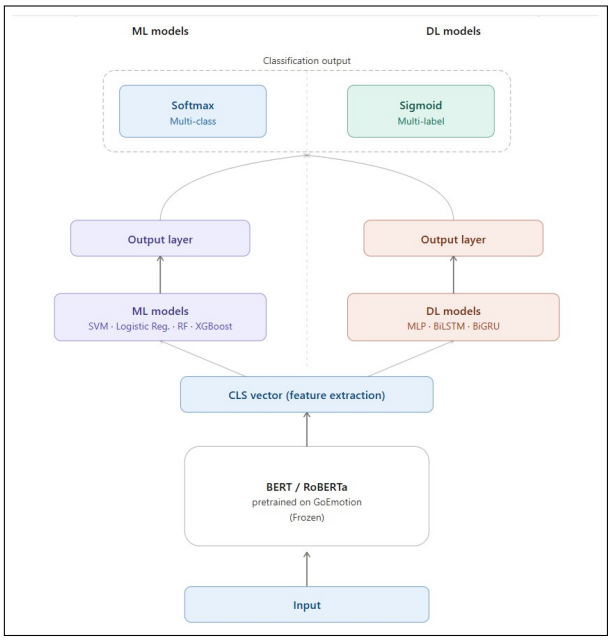



Figure 3.5: Transfer Learning from GoEmotion + ML/DL Models.

### 3.4.2 Proposed Models for Emotion Recognition in Arabic Texts

**Approach 1: Fine-tuning Transformer-based Models**

The first approach relies on fine-tuning pre-trained Transformer models, namely AraBERT, MarBERT, CAMeLBERT, and XLM-RoBERTa. The [CLS] output of the AraBERT,MarBERT,

CAMeLBERT, and XLM-RoBERTa models (see Figure 3.6) is forwarded to an output dense layer using a sigmoid activation for multi-label classification.

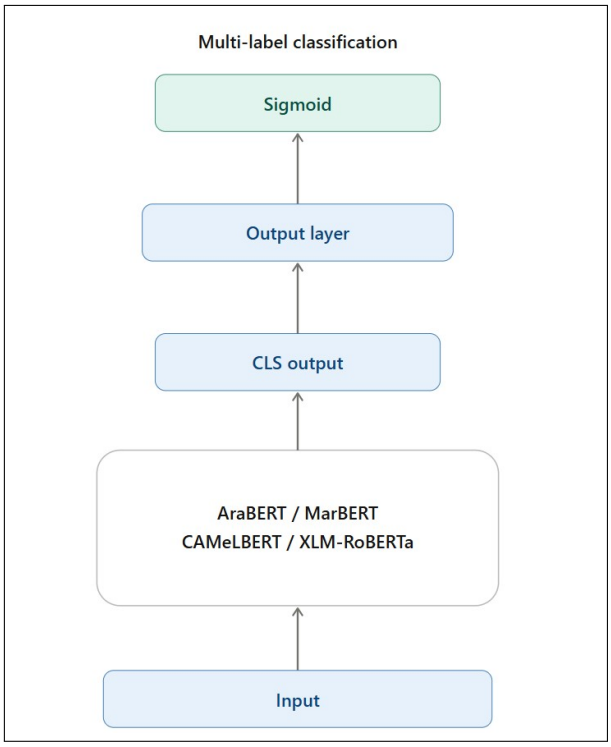



Figure 3.6: Fine-tuning Transformer-based Models.

**Approach 2 : Static Vs. Contextual Word Embedding with BiGRU/BiLSTM and Self-attention**

**Part 1: FastText Embedding with BiLSTM/BiGRU and Self-attention**

The aim of this approach is to evaluate static versus contextual word embeddings for fine-grained emotion recognition, combined with BiLSTM/BiGRU and self-attention. FastText is selected as the static representation using the pre-trained cc.ar.300.vec model to handle OOV words in social media text. AraBERT, MarBERT, CAMeLBERT, and XLM-RoBERTa are chosen as contextual embeddings for their context-aware representations. The FastText-based models, shown in figure 3.7, pass the embedding output through either BiGRU or BiLSTM, followed by a dense layer with sigmoid activation for multi-label classification.

***Models without Self-Attention***

- **FastText + BiGRU without Self-Attention**
  - Global max pooling for multi-label classification
- **FastText + BiLSTM without Self-Attention**
  - Global max pooling for multi-label Classification

***Models with Self-Attention***

- **FastText + BiGRU + Self-Attention**
- **FastText + BiLSTM + Self-Attention**

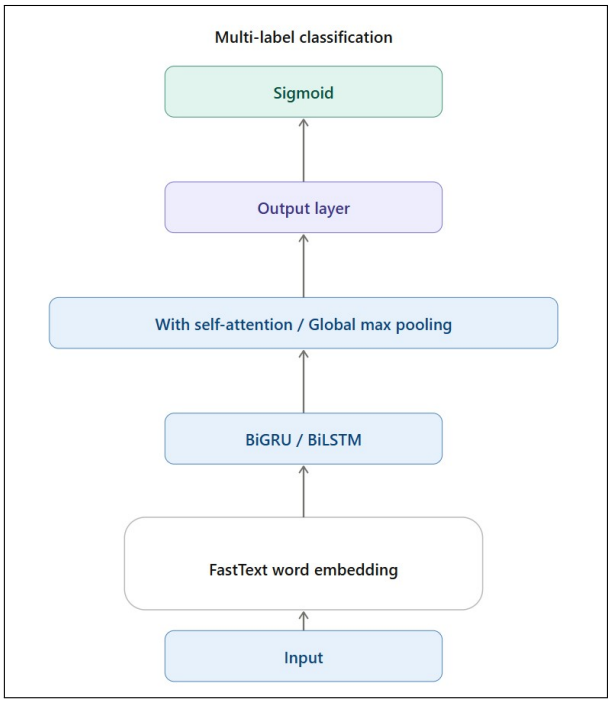



Figure 3.7: Static Embedding with BiLSTM/BiGRU and Self Attention .

**Part 2: AraBERT/MarBERT and CAMeLBERT Word Embedding with BILSTM/BIGRU and Self-attention**

In this models,frozen (not trainable) contextual embeddings from AraBERT /MarBERT and CAMeLBERT , are integrated with BiGRU/BiLSTM layers ,with self-attention. As illustrated in Figure 3.8,the output of this context-based word embedding layer serves as an input to either a (BiGRU) or (BiLSTM) followed by final a dense layer with a sigmoid activation function for multi-label classification.

***Models Without Self-Attention***

- **AraBERT/MarBERT and CAMeLBERT (not trainable) + BiLSTM without Self-Attention**
  - Global max pooling for multi-label Classification
- **AraBERT/MarBERT and CAMeLBERT (not trainable) + BiGRU without Self-Attention**
  - Global max pooling for multi-label Classification

***Models With Self-Attention***

- **AraBERT/MarBERT and CAMeLBERT (not trainable) + BiGRU + Self-Attention**
- **AraBERT/MarBERT and CAMeLBERT(not trainable) + BiLSTM + Self-Attention**

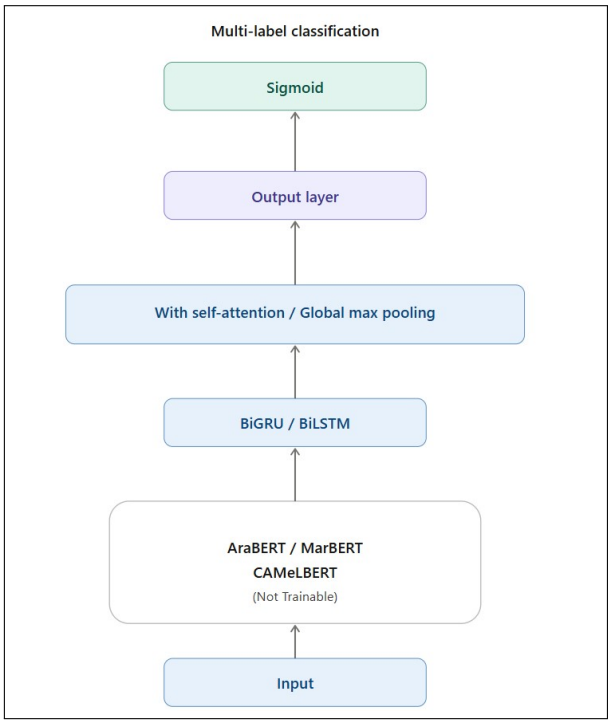



Figure 3.8: AraBERT/MarBERT/CAMeLBERT Contextual Word Embeddings BiLSTM/BiGRU and Self Attention .

**Approach 3: Fine-tuning Hybrid Transformer-based Models and BiLSTM/BiGRU and Self Attention**

In this approach, we fine-tune trainable AraBERT,MarBERT and CAMeLBERT models jointly with BiGRU/BiLSTM layers and Self-attention As illustrated in the figure 3.9, This transformer- based model layer's output [CLS] becomes input for a (BIGRU) or (BILSTM) followed by final a dense layer. This final layer will consists of either a sigmoid activation function for multi-label classification.

***Models Without Self-Attention***

- **AraBERT/MarBERT and CAMeLBERT(trainable) + BiLSTM without Self-Attention**
  - Global max pooling for multi-label Classification
- **AraBERT/MarBERT and CAMeLBERT(trainable) + BiGRU without Self-Attention**
  - Global max pooling for multi-label Classification

***Models With Self-Attention***

- **AraBERT/MarBERT and CAMeLBERT (trainable) + BiGRU + Self-Attention**
- **AraBERT/MarBERT and CAMeLBERT (trainable) + BiLSTM + Self-Attention**

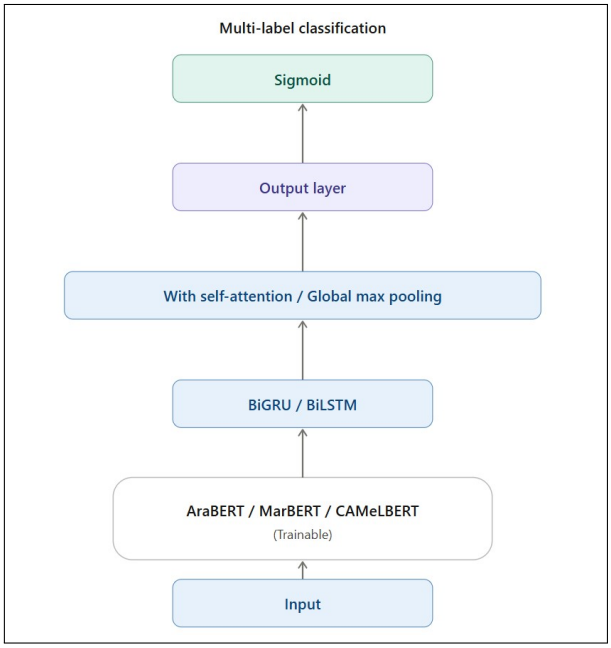



Figure 3.9: Fine-tuning Hybrid Transformer-based and BiLSTM/BiGRU and Self Attention.

## 3.5 Performance Evaluation Metrics

To assess the performance of the proposed approaches on multi-class and multi-label settings, we relied on five evaluation metrics in addition to confusion matrices, including: accuracy, recall, precision, f1-score, and AUC. These metrics are reported as macro-average and per-class/per-label scores.

**Accuracy**: Measures the overall correctness of the model.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

**Precision**: Indicates how many of the predicted positive instances are actually positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

**Recall**: Indicates how many of the actual positive instances were correctly predicted.

$$\text{Recall} = \frac{TP}{TP + FN}$$

**F1 Score**: The harmonic mean of Precision and Recall.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

**Confusion Matrix**: Summarizes the performance of a classification model.

| | **Predicted** | |
|---|---|---|
| | **Positive** | **Negative** |
| **Actual Positive** | $TP$ | $FN$ |
| **Actual Negative** | $FP$ | $TN$ |

## 3.6 Conclusion

This chapter has presented the research approach adopted for fine-grained emotion recognition from Arabic and English social media texts. Key components of the methodology were defined, including the problem statement, proposed approaches, and evaluation metrics. Four approaches were proposed for English texts: fine-tuning transformers, word embeddings with BiGRU/BiLSTM, hybrid transformer-recurrent architectures, and transfer learning from GoEmotions. For Arabic texts, three analogous approaches were proposed using AraBERT, MarBERT, CAMeLBERT, and XLM-RoBERTa.

The proposed approach enables a comprehensive comparison between fine-tuning, frozen embeddings, hybrid trainable architectures, and cross-corpus transfer learning. These foundations motivate the experimental study presented in the following chapter, which aims to evaluate and identify the most effective approach for emotion recognition across both languages.

# Chapter 4

## Results and discussion

# Chapter 4

# Results and discussion

## 4.1 Introduction

This chapter presents the experimental results and discussion of our emotion recognition system. We first provide an overview of the experimental study, followed by a detailed description of the experimental environment including the hardware and software tools used. Next, we present the datasets employed for training and evaluation, along with their characteristics and partition statistics. Finally, we discuss the results obtained from our four approaches and compare them with existing literature and state-of-the-art models.

## 4.2 Experimental Study

### 4.2.1 Overview

This study presents the experimental design for evaluating fine-grained emotion recognition models on English and Arabic texts. Three datasets were used: ISEAR (English, multi-class), CBET (English, multi-label), and SemEval-2018 (Arabic, multi-label), with a unified split of 70% training, 10% validation, and 20% testing.
It is important to note that no text preprocessing (such as stemming, stop-word removal, or normalization) was applied to any of the datasets, as preliminary experiments showed that such preprocessing steps consistently degraded model performance across all architectures. This is likely due to the reliance of transformer-based and contextual embedding models on the original surface forms, word order, and punctuation to capture nuanced emotional cues.

For English, four approaches were applied:

- **Approach 1:** Fine-tuning pretrained transformer-based models on target emotion recognition datasets.
- **Approach 2:** Building BiLSTM/BiGRU models with static and contextual word embeddings combined with self-attention.
- **Approach 3:** Building hybrid transformer models combined with BiLSTM/BiGRU layers and self-attention.

- **Approach 4:** Transfer learning from GoEmotions to target emotion recognition datasets.

For Arabic, three analogous approaches were applied using specialized transformers such as AraBERT, MarBERT, CAMeLBERT, and XLM-RoBERTa. All models were evaluated using precision, recall, F1-score, accuracy, and AUC metrics(see figure 4.2).

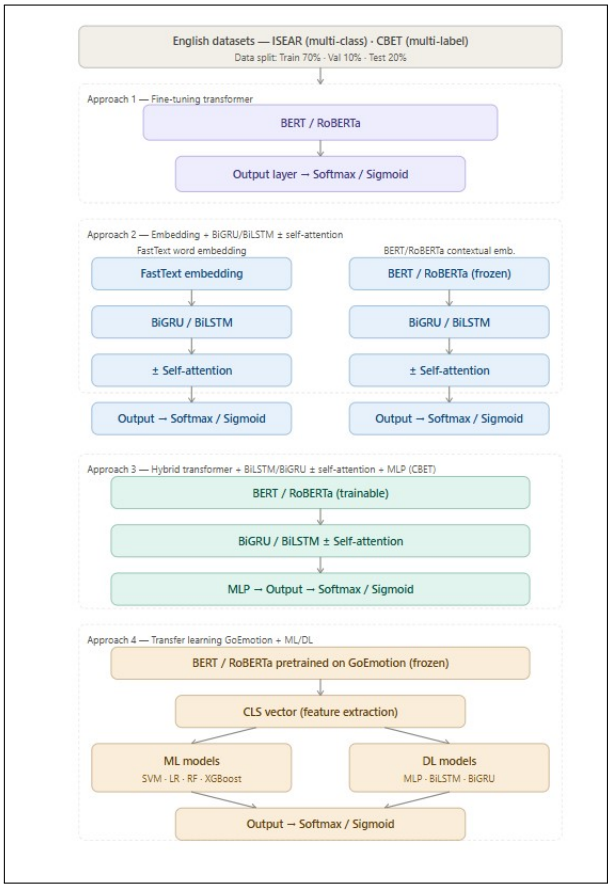



Figure 4.1: Detailed Plan of the Proposed Approaches for Emotion Recognition on English Datasets.

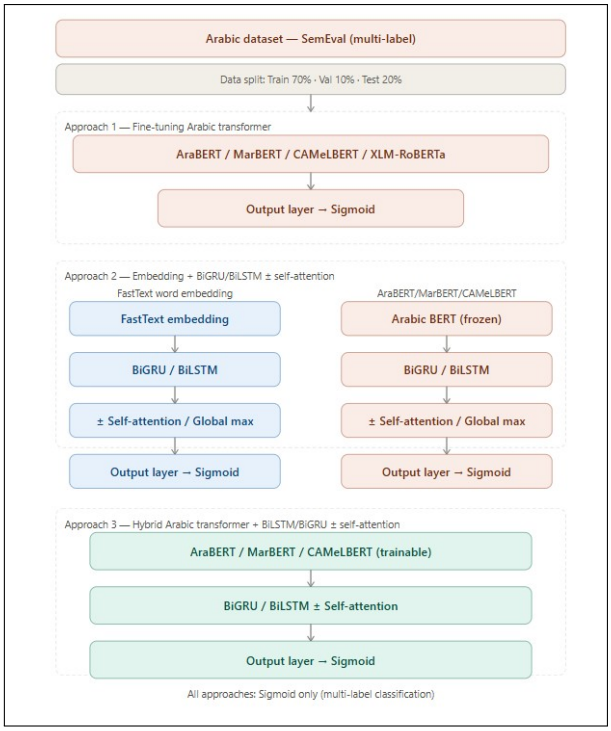



Figure 4.2: Detailed Plan of the Proposed Approaches for Emotion Recognition on Arabic Dataset

### 4.2.2 Experimental Environment

All experiments were conducted using Google Colab which provides free access to NVIDIA T4 GPUs, 16-32 GB of RAM, persistent storage integration with Google Drive, pre-installed libraries for deep learning, and support for PyTorch and Transformers libraries.Table 4.1 below details the hardware configuration and software environment used during our experiments in Google Colab.

Table 4.1: Experimental environment specifications.

| **Component** | **Version / Specification** |
|---|---|
| Python | 3.12.13 |
| PyTorch | 2.11.0+cu128 |
| Transformers | 5.0.0 |
| Scikit-learn | 1.6.1 |
| Pandas | 2.2.2 |
| NumPy | 2.0.2 |
| CUDA | 12.8 |
| GPU | Tesla T4 |
| GPU Memory | 15.64 GB |
| RAM | 13.3 GB |

### 4.2.3 Training and Evaluation Datasets

In this study, three public datasets are used for our experiments: ISEAR, CBET, and SemEval 2018. These datasets were obtained from their original sources:ISEAR from Hugging Face, CBET from a public GitHub repository, and SemEval from the Codalab competition platform.table 4.2 describes the English and Arabic datasets used in our study for training and evaluating the proposed models .

Table 4.2: Detailed description of training and evaluation datasets.

| **Dataset** | **Characteristics** |
|---|---|
| **ISEAR** | Size: 7,516 samples<br>Source: International Survey on Emotion Antecedents and Reactions<br>Number of Classes: 7<br>Classes: joy, fear, anger, sadness, disgust, shame, guilt<br>Type: Multi-class<br>Link: https://huggingface.co/datasets/gsri-18/ISEAR-dataset-complete |
| **CBET** | Size: 81,164 samples<br>Source: Crowdsourced English tweets<br>Number of Classes: 9<br>Classes: anger, surprise, disgust, fear, joy, love, guilt, thankfulness, sadness<br>Type: Multi-label<br>Link: https://github.com/chenyangh/CBET-dataset/tree/master |
| **SemEval-2018** | Size: 2,866 samples<br>Source: SemEval-2018 Task 1 (E-c)<br>Number of Classes: 11<br>Classes: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust<br>Type: Multi-label<br>Link: https://competitions.codalab.org/competitions/17751 |

### 4.2.4 Emotion Label Distribution Across Three Datasets

**- ISEAR dataset:**

To visualize the distribution of emotion classes/labels in the English and Arabic datasets used during experiments, figures 4.2, 4.3, and 4.4 below present bar plots indicating the number of occurrences for each class/label in the datasets ISEAR, CBET, and SemEval2018, respectively.
As shown in figure 4.3, ISEAR dataset showcases a balanced distribution of its seven emotion classes, namely : joy, sadness, anger, fear, shame, disgust, and guilt.

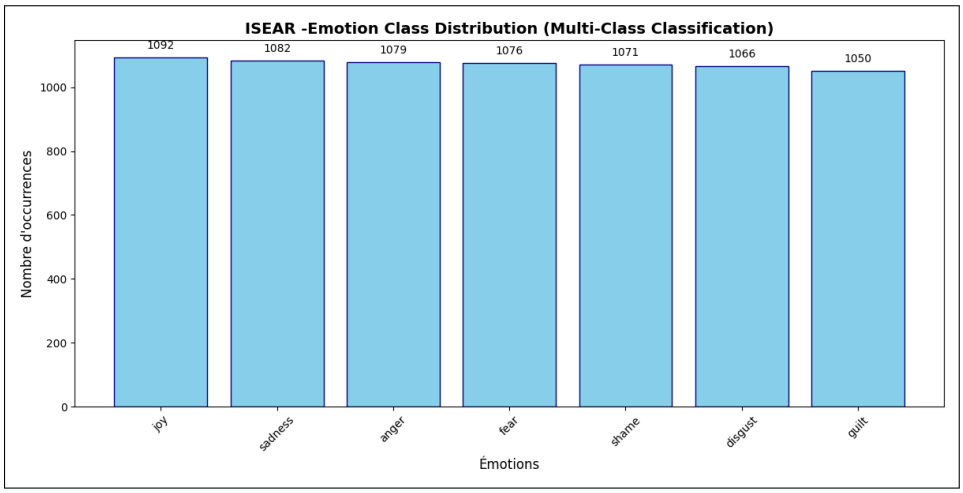



Figure 4.3: Distribution of emotion labels in the ISEAR dataset.

**- CBET dataset:**

As shown in figure 4.4, CBET presents a light label imbalance, especially for thankfulness, disgust, and guilt emotion labels.

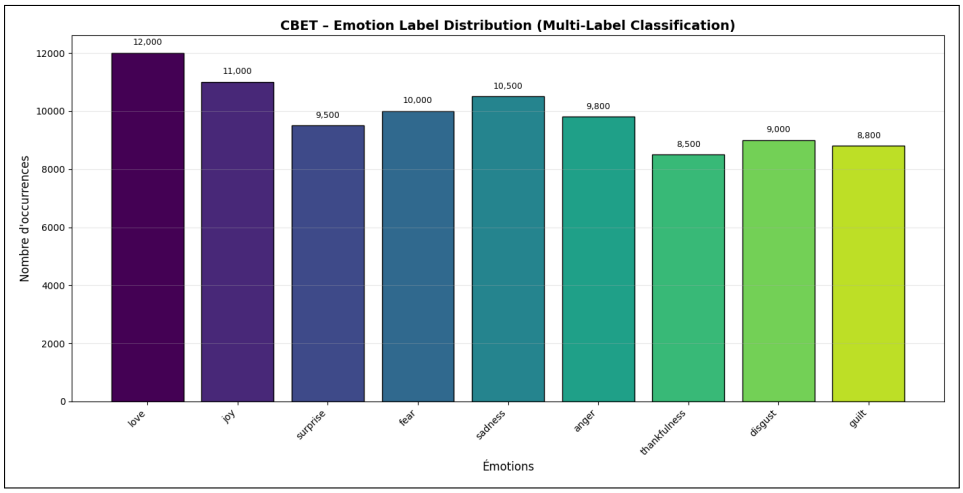



Figure 4.4: Distribution of emotion labels in the CBET dataset.

**- Sem Eval dataset:**

As shown in figure 4.5, SemEval-2018 dataset presents a severe label imbalance , especially for labels anticipation,trust, and surprise being represented by few samples compared to the other emotion labels.

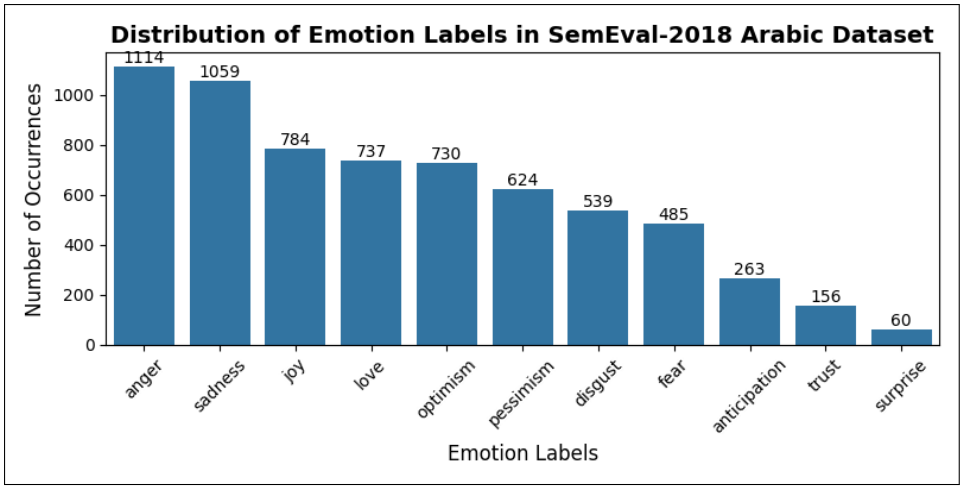



Figure 4.5: Distribution of emotion labels in the dataset.

### 4.2.5 Training,validation,and test sets

As it is aforementioned, all of the English and Arabic datasets were divided into 70% training , 10% validation, and 20% test (see table 4.3 ).

Table 4.3: Description of training , validation, and test sets.

| **Dataset** | **Train** | **Validation** | **Test** | **Total** |
|---|---|---|---|---|
| ISEAR | 5,261 | 751 | 1,504 | 7,516 |
| CBET | 56,814 | 8,116 | 16,233 | 81,164 |
| SemEval 2018 | 2,005 | 286 | 574 | 2,866 |

### 4.2.6 Hyperparameters

During our experiments, all the proposed models were trained for 5 epochs using Adam optimizer by applying early stopping (see figure 4.4).

Table 4.4: Summarizesall the hyperparameters used for training the proposed models.

| **Parameter** | **Value** |
|---|---|
| Max Length | 128 |
| Learning Rate (LR) | (2e-5)Transformer models , (1e-3) Fast text |
| Epochs | 5 |
| Batch Size | 16 |
| Dropout | 0.3 |
| Early Stopping | 3 epochs (patience) |
| Optimizer | AdamW |
| Loss Function | |
| Multi-label datasets | BCEWithLogitsLoss |
| Multi-class datasets | CrossEntropyLoss |

## 4.3 Results

### 4.3.1 Multi-Class Emotion Classification of English Texts:ISEAR Dataset

#### *Approach 1: Fine-tuning Transformer-based Models*

The following table presents the results obtained by Fine-Tuning BERT and RoBERTa models on ISEAR dataset ( Approach 1) evaluated across multiple macro-averaged performance metrics.These results indicate the outperformance of RoBERTa over BERT across all the macro-average metrics used for evaluation, by achieving an f1-score of 0.7115 and an AUC of 0.9327.

Table 4.5: Macro-average scores obtained by Approach 1.

| Approach | Model | Accuracy | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|---|
| Fine-Tuning Transformer-Based Models | BERT | 0.7061 | 0.7019 | 0.7051 | 0.7024 | 0.9290 |
| | **RoBERTa** | **0.7168** | **0.7115** | **0.7160** | **0.7115** | **0.9327** |

Table 4.6 presents per-class scores of precision, recall, f1-score, and AUC metrics achieved by RoBerta on ISEAR dataset.Based on the results above,Roberta model achieved the highest values for recall and F1- score for the "Joy" class by obtaining 0.9315 and 0.8718, respectively. Whereas, this model recorded the lowest scores for "anger" class (0.5463 for recall and 0.5990 for f1-score).
The strong overall AUC 0.9327 indicates that the model provides an excellent level of discrimination; however, the mean F1 score of 0.7115 indicates that the model struggles with certain negative emotions.Such as anger 0.5990,shame 0.6100,and guilt 0.6748.

Table 4.6: Per-class performance of RoBerta on ISEAR dataset.

| Metric | anger | disgust | fear | guilt | joy | sadness | shame | Macro Avg |
|---|---|---|---|---|---|---|---|---|
| Precision | 0.6629 | 0.7070 | 0.7479 | 0.6881 | 0.8193 | 0.6992 | 0.6559 | 0.7115 |
| Recall | 0.5463 | 0.7136 | 0.8279 | 0.6619 | 0.9315 | 0.7604 | 0.5701 | 0.7160 |
| F1-score | 0.5990 | 0.7103 | 0.7859 | 0.6748 | 0.8718 | 0.7285 | 0.6100 | 0.7115 |
| AUC | 0.8981 | 0.9313 | 0.9642 | 0.9078 | 0.9806 | 0.9448 | 0.9022 | 0.9327 |

### *Approach 2:Static and Contextual Word Embeddings with BiGRU/BiLSTM and Self-attention*

The following table presents the results obtained by Static Vs. Contextual Word Embedding with BiGRU/BiLSTM and Self-attention for Multi-class Emotion Classification on the ISEAR dataset (Approach 2) evaluated across multiple macro-averaged performance metrics.These results demonstrate the outperformance of Fasttext static embedding based on BERT and RoBERTa , especially when combined with BIGRU and self-attention (Fasttext + BiGRU +Att) by achieving the highest f1-score 0.5571 and the largest AUC 0.8692.

Table 4.7: Macro-average scores obtained with Approach 2 on ISEAR dataset.

| Approach | Model | Accuracy | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|---|
| Fast Text Embedding | FastText+BiGRU | 0.5419 | 0.5528 | 0.5410 | 0.5334 | 0.8522 |
| | **FastText+BiGRU+A** | **0.5631** | **0.5735** | **0.5623** | **0.5571** | **0.8692** |
| | FastText+BiLSTM | 0.5166 | 0.5245 | 0.5156 | 0.5094 | 0.8264 |
| | FastText+BiLSTM+Att | 0.5047 | 0.5542 | 0.5046 | 0.4906 | 0.8472 |
| BERT Contextual Word Embedding | BERT(Embedding)+ BiGRU | 0.4880 | 0.4806 | 0.4868 | 0.4664 | 0.8290 |
| | BERT(Embedding)+ BiGRU+Att | 0.4980 | 0.4932 | 0.4961 | 0.4788 | 0.8338 |
| | BERT(Embedding)+ BiLSTM | 0.5020 | 0.5074 | 0.5010 | 0.4840 | 0.8350 |
| | BERT(Embedding)+ BiLSTM+Att | 0.5266 | 0.5222 | 0.5255 | 0.5128 | 0.8413 |
| RoBERTa Contextual Word Embedding | RoBERTa(Embedding)+ BiGRU | 0.2999 | 0.3465 | 0.2986 | 0.2673 | 0.6908 |
| | RoBERTa(Embedding)+ BiGRU+Att | 0.3424 | 0.3297 | 0.3414 | 0.3198 | 0.7106 |
| | RoBERTa(Embedding)+ BiLSTM | 0.3012 | 0.3449 | 0.3005 | 0.2796 | 0.7019 |
| | RoBERTa(Embedding)+ BiLSTM+Att | 0.2965 | 0.3178 | 0.2961 | 0.2788 | 0.6981 |

Table 4.8 Below shows per-class scores of precision, recall, f1-score, and AUC metrics achieved by RoBerta on ISEAR dataset.Based on the results of Table 4.8 above, the highest score for precision 0.6791 is obtained for the ”joy” class, while achieving the highest recall 0.8000 for the ”fear” class. On the other hand, the ”guilt” class exhibited the lowest recall with a score of 0.4429 and the ”shame” class recorded the lowest precision score (0.4262).

Table 4.8: Per-class scores of FastText + BiGRU + Attention on ISEAR dataset.

| Metric | Anger | Disgust | Fear | Guilt | Joy | Sadness | Shame | Macro Avg |
|---|---|---|---|---|---|---|---|---|
| Precision | 0.6029 | 0.6243 | 0.5029 | 0.5671 | 0.6791 | 0.6118 | 0.4262 | 0.5735 |
| Recall | 0.3796 | 0.5070 | 0.8000 | 0.4429 | 0.6667 | 0.6682 | 0.4720 | 0.5623 |
| F1-score | 0.4659 | 0.5596 | 0.6176 | 0.4973 | 0.6728 | 0.6388 | 0.4479 | 0.5571 |
| AUC | 0.8330 | 0.8502 | 0.9175 | 0.8420 | 0.9365 | 0.8883 | 0.8170 | 0.8692 |

### *Approach 3: Hybrid Transformer-Based models with BiLSTM/BiGRU and self attention*

The following table presents the results obtained by Hybrid Transformer- based and BiLSTM/BiGRU and self attention for Multi-class Emotion Classification on the ISEAR dataset (Approach 3) evaluated across multiple macro-averaged performance metrics.These results showcase the superior performance of Roberta-based models over bert-based models across all the performance metrics. Whereby, the best-performing model is Roberta combined with BiGRU (Roberta + BiGRU) achieving the highest values of f1-score 0.7308 and AUC 0.9461.

Table 4.9: Macro-average scores obtained by Approach 3 on ISEAR dataset.

| Approach | Model | Accuracy | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|---|
| HybridTransformer-based [BERT] and BiLSTM/BiGRU models | BERT+ BiGRU | 0.7088 | 0.7112 | 0.7083 | 0.7083 | 0.9427 |
| | BERT+BiGRU+Att | 0.7061 | 0.7005 | 0.7053 | 0.7016 | 0.9374 |
| | BERT+ BiLSTM | 0.7088 | 0.7126 | 0.7078 | 0.7084 | 0.9334 |
| | BERT+ BiLSTM+Att | 0.7061 | 0.7067 | 0.7052 | 0.7023 | 0.9306 |
| HybridTransformer-based [RoBERTa] and BiLSTM/BiGRU models | **RoBERTa+BiGRU** | **0.7307** | **0.7347** | **0.7301** | **0.7308** | **0.9461** |
| | RoBERTa+BiGRU+Att | 0.7207 | 0.7244 | 0.7201 | 0.7194 | 0.9404 |
| | RoBERTa + BiLSTM | 0.7281 | 0.7307 | 0.7275 | 0.7274 | 0.9387 |
| | RoBERTa+BiLSTM+At | 0.7214 | 0.7230 | 0.7207 | 0.7202 | 0.9380 |

Table 4.10 Below shows per-class scores of precision, recall, f1-score, and AUC metrics achieved by f RoBERTa + BiGRU on ISEAR dataset

Table 4.10: Per-class scores of RoBERTa + BiGRU on ISEAR dataset.

| Metric | Anger | Disgust | Fear | Guilt | Joy | Sadness | Shame | Macro Avg |
|---|---|---|---|---|---|---|---|---|
| Precision | 0.6754 | 0.6723 | 0.8750 | 0.6682 | 0.9120 | 0.7265 | 0.6137 | 0.7347 |
| Recall | 0.5972 | 0.7418 | 0.7488 | 0.6714 | 0.8995 | 0.7834 | 0.6682 | 0.7301 |
| F1-score | 0.6339 | 0.7054 | 0.8070 | 0.6698 | 0.9057 | 0.7539 | 0.6398 | 0.7308 |
| AUC | 0.9197 | 0.9407 | 0.9660 | 0.9336 | 0.9882 | 0.9509 | 0.9235 | 0.9461 |

The table above demonstrates that the label with the best performance in terms of precision (0.9120) and recall (0.8995) is joy, because both metrics indicate that the predictions made by the model for this emotion are highly accurate and complete. Conversely, two labels exhibit poor performance: anger and shame. In terms of precision, anger shows a low score (0.6137). Overall, shame is the weakest label since it has the lowest precision (0.6137) and F1 score (0.6398). However, it retains a slightly better recall than anger at 0.6682.

### *Approach 4: Transfer Learning from GoEmotion + ML/DL models BERT RoBERTa based models*

The experimental results of Approach 4 on ISEAR dataset are presented in Table 4.11 below. Across all metrics used for evaluation, The presented results highlight the higher performance of feature representations derived by the RoBERTa model pretrained on GoEmotion compared to representations extracted by BERT. Besides, The SVM model using transferred GoEmotion' RoBERTa representations produced the best overall performance for this approach by achieving the highest F1-score (0.6347) and the largest AUC (0.8957).

Table 4.11: Macro-average scores obtained by Approach 4 on ISEAR dataset.

| Approach | Model | Accuracy | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|---|
| Transfer Learning from GoEmotion [BERT] + ML/DL models | Logistic Regression | 0.5392 | 0.5365 | 0.5385 | 0.5367 | 0.8329 |
| | SVM | 0.5858 | 0.5824 | 0.5846 | 0.5807 | 0.8730 |
| | Random Forest | 0.5632 | 0.5626 | 0.5620 | 0.5578 | 0.8603 |
| | XGBoost | 0.5751 | 0.5709 | 0.5741 | 0.5715 | 0.8662 |
| | MLP | 0.5459 | 0.5500 | 0.5400 | 0.5400 | 0.8551 |
| | BiLSTM | 0.5392 | 0.5400 | 0.5400 | 0.5300 | 0.8503 |
| | BiGRU | 0.5359 | 0.5300 | 0.5300 | 0.5300 | 0.8529 |
| Transfer Learning from GoEmotion [RoBERTa] + ML/DL models | Logistic Regression | 0.6104 | 0.6062 | 0.6092 | 0.6067 | 0.8804 |
| | **SVM** | **0.6396** | **0.6370** | **0.6383** | **0.6347** | **0.8957** |
| | Random Forest | 0.6203 | 0.6100 | 0.6136 | 0.6089 | 0.8867 |
| | XGBoost | 0.6250 | 0.6195 | 0.6238 | 0.6201 | 0.8933 |
| | MLP | 0.6117 | 0.6100 | 0.6100 | 0.6000 | 0.8919 |
| | BiLSTM | 0.6217 | 0.6200 | 0.6200 | 0.6100 | 0.8925 |
| | BiGRU | 0.6336 | 0.6300 | 0.6300 | 0.6300 | 0.8927 |

Table 4.12 Below shows per-class scores of precision, recall, f1-score, and AUC metrics achieved by SVM with RoBERTa Representations Transferred from GoEmotion.on ISEAR dataset

Table 4.12: Performance of SVM using RoBERTa (GoEmotions) CLS embeddings.

| Metric | Anger | Disgust | Fear | Guilt | Joy | Sadness | Shame | Macro Avg |
|---|---|---|---|---|---|---|---|---|
| Precision | 0.4883 | 0.6186 | 0.7466 | 0.6074 | 0.7649 | 0.6804 | 0.5531 | 0.6370 |
| Recall | 0.5787 | 0.6244 | 0.7674 | 0.4714 | 0.8767 | 0.6866 | 0.4626 | 0.6383 |
| F1-score | 0.5297 | 0.6215 | 0.7569 | 0.5308 | 0.8170 | 0.6835 | 0.5038 | 0.6347 |
| AUC | 0.8513 | 0.8870 | 0.9433 | 0.8538 | 0.9717 | 0.9198 | 0.8432 | 0.8957 |

The results of Table 4.12 presented above demonstrate the highest performance of SVM model with RoBERTa GoEmotion representations in distinguishing the ”Joy” samples by achieving the highest recall value (0.8767) and the highest F1-score (0.8170). On the other hand, Tthe ”Shame” class is observed to be the weakest label, with a low recall value of 0.4626 and limited F1-score of 0.5038. The ”Anger” and ”Guilt” classes also displayed difficulties, with the lowest precision measure of 0.4883 and the lowest recall value of 0.4714.

### 4.3.2 Global Discussion of Multi-Class Emotion Recognition Approaches on the ISEAR Dataset

The experimental framework of this study was applied to the ISEAR multi-class dataset, which contains seven balanced emotion classes, in which four different approaches to emotion classification were evaluated. The results demonstrated that Approach 3, based on fine-tuning hybrid Transformer models combined with BiLSTM/BiGRU architectures, achieved the best overall performance. Specifically, the trainable RoBERTa + BiGRU model obtained the highest F1-score of 0.7308 and an AUC of 0.9461.(see Figure 4.6).

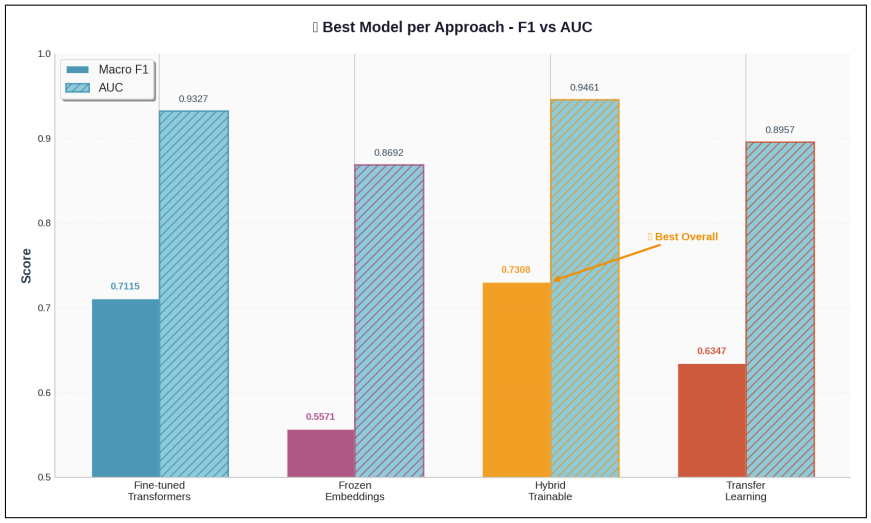



Figure 4.6: Comparative performance of all four approaches: Macro F1 and AUC.

Within Approach 2, a clear performance difference was observed between Fast Text and contextual embedding models (BERT and RoBERTa). FastText uses static word representations, where each word has a single fixed vector. In contrast, BERT and RoBERTa produce contextual embeddings that vary depending on surrounding words. However, the results on the ISEAR dataset showed that FastText outperformed both BERT and RoBERTa in this approach, suggesting that simpler static embeddings were more effective in this setting. The relatively lower performance of contextual word embeddings can be attributed to the increased model complexity induced by high-dimensional token representations. Besides, BERT and RoBERTa word embedding encode each token as a 768-dimensional vector, whereas FastText represents tokens using 300-dimensional embeddings. This increase in input dimensionality results in a larger number of trainable parameters, making the learning process more complex and prone to overfitting. Such an effect becomes particularly evident when training on small datasets, such as the ISEAR dataset, where the limited number of training samples may be insufficient to support the effective learning of high-dimensional representations and achieve robust generalization.

Regarding the impact of Self-Attention, the experimental results indicate that adding explicit Self-Attention layers does not improve performance on the ISEAR dataset and may even degrade it. The strongest evidence is that the best-performing model across all approaches was the trainable RoBERTa + BiGRU model without any

additional attention mechanism. This can be explained by the relatively small size of the ISEAR dataset and the fact that RoBERTa already incorporates Multi-Head Self-Attention within its architecture. Therefore, adding extra attention layers increases model complexity without providing significant benefits, which may negatively affect performance.

The per-class analysis revealed that the Joy class achieved the best performance. This can be explained by the fact that joy is typically expressed through clear and distinctive linguistic patterns with minimal overlap with other emotions, making it easier for the model to classify correctly.

In contrast, the anger class was the worst-performing category. This is mainly due to the strong semantic and psychological overlap of the Anger class with Disgust, Sadness, and Shame, which leads to frequent misclassification between these emotions. (see Figure 4.7and Figure 4.8 below).

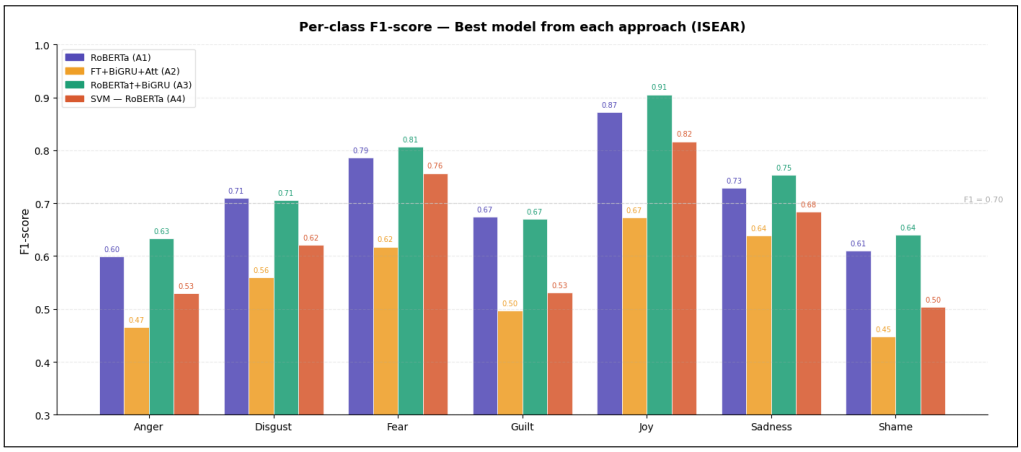



Figure 4.7: Per-class F1 scores across of approaches 1,2,3 and 4 on ISEAR dataset.

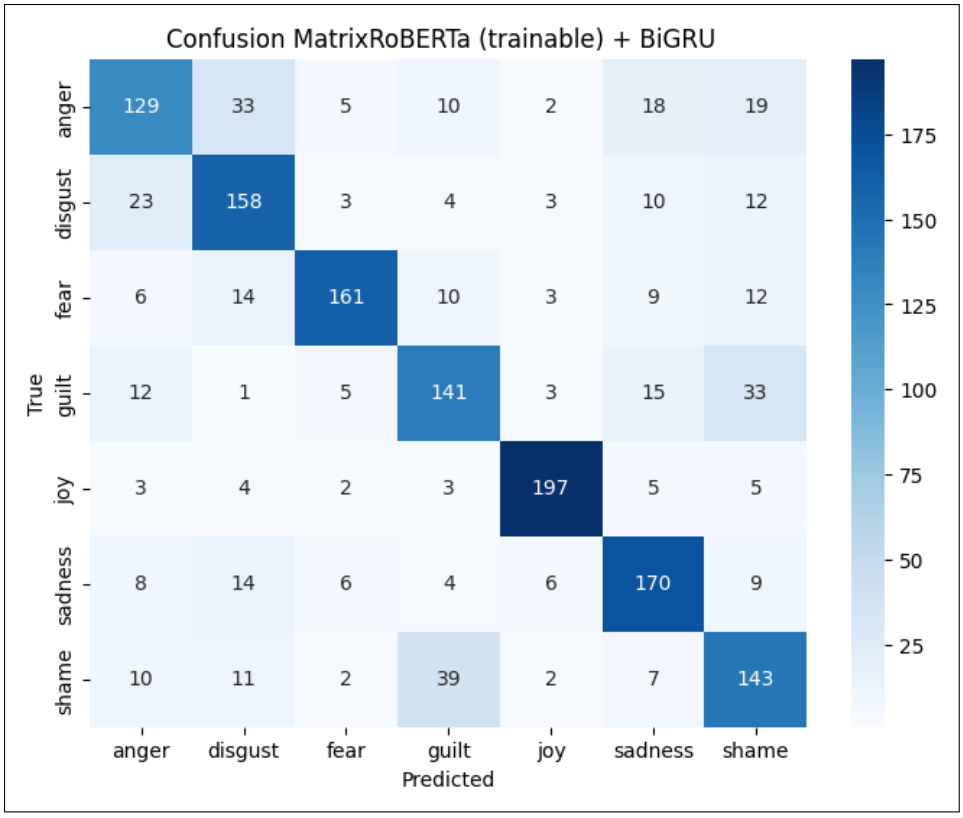



Figure 4.8: Confusion matrix of the best-performing model (RoBERTa + BiGRU / Approach 3) on the ISEAR set.

### 4.3.3 Multi-Label Emotion Classification of English Texts: CBET Dataset

#### *Approach 1: Fine-tuning Transformer-based Models*

The following table presents the results obtained with Approach 1 on the CBET dataset evaluated across the macro-averaged performance metrics. These results highlight the superior performance of RoBERTa over BERT across all the performance metrics when fine-tuned on CBET dataset for fine-grained multi-label emotion classification. RoBERTa model slightly outperformed BERT in terms of f1-score and AUC by obtaining 0.6482 and 0.9203, respectively.

Table 4.13: Macro-average scores obtained by approach 1 on CBET dataset.

| Approach | Model | Accuracy | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|---|
| Fine-Tuning Transformer-Based Models | BERT | 0.5785 | 0.6634 | 0.6171 | 0.6376 | 0.9080 |
| | **RoBERTa** | **0.5749** | **0.6994** | **0.6059** | **0.6482** | **0.9203** |

Per-label scores for the best-performing model RoBERTa in measure of precision, recall, f1-score and AUC are presented in Table 4.14 below.

Table 4.14: Per-label performance of RoBERTa on CBET dataset.

| Metrics | Anger | Fear | Joy | Love | Sadness | Surprise | Thankfulness | Disgust | Guilt | Macro Avg |
|---|---|---|---|---|---|---|---|---|---|---|
| Precision | 0.6181 | 0.8636 | 0.6535 | 0.7435 | 0.5942 | 0.6753 | 0.8223 | 0.6892 | 0.6381 | 0.6994 |
| Recall | 0.4961 | 0.6956 | 0.6032 | 0.5692 | 0.4644 | 0.6643 | 0.7472 | 0.6391 | 0.5771 | 0.6059 |
| F1-score | 0.5501 | 0.7701 | 0.6276 | 0.6453 | 0.5212 | 0.6693 | 0.7833 | 0.6632 | 0.6063 | 0.6482 |
| AUC | 0.8822 | 0.9545 | 0.9173 | 0.9151 | 0.8891 | 0.9204 | 0.9573 | 0.9406 | 0.9092 | 0.9203 |

The per-label scores of Table 4.14 above obtained by fine-tuning RoBERTa on CBET dataset demonstrates the high performance of this model in recognizing ”Fear” and ”Thankfulness” labels. Whereby, the ”Fear” label recorded an F1-score of (0.7701) and a precision of (0.8636), while ”Thankfulness” reached a high recall of (0.7472) and a high F1-score of (0.7833). In contrast, the ”Sadness” and ”Anger” showcased to be the weakest labels by obtaining the lowest recall (0.4644) and F1-score (0.5212) for ”Sadness” and the lowest precision (0.6181) with a low recall (0.4961) for ”anger”. Guilt also received moderate scores across all metrics.

### *Approach 2:Static and Contextual Word Embeddings with BiGRU/BiLSTM and Self-attention*

The following table presents the results obtained with Approach 2 on CBET dataset evaluated across the macro-averaged values of all performance evaluation metrics. Compared to contextual embeddings derived by BERT and RoBERTa models, the static word embeddings based on FastText higlighted to produce superior performance in terms of all performance metrics. The FastText+BiLSTM demostrated the highest performance by obtaining high scores for F1-Score (0.5507) and AUC (0.8658).

Table 4.15: Macro-average scores obtained with Approach 2 on CBET dataset.

| Approach | Model | Accuracy | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|---|
| FastText Embedding | FastText+BiGRU | 0.3714 | 0.4578 | 0.6088 | 0.5216 | 0.8483 |
| | FastText+BiGRU+Att | 0.3901 | 0.4806 | 0.6051 | 0.5347 | 0.8584 |
| | **FastText+BiLSTM** | **0.4009** | **0.4878** | **0.6287** | **0.5507** | **0.8658** |
| | FastText+BiLSTM+A | 0.3800 | 0.4640 | 0.6368 | 0.5405 | 0.8630 |
| BERT Contextual Embedding | BERT(Embedding)+ BiGRU | 0.2760 | 0.7123 | 0.2963 | 0.4056 | 0.8607 |
| | BERT(Embedding)+ BiGRU+ Att | 0.3196 | 0.7350 | 0.3196 | 0.4267 | 0.8696 |
| | BERT (Embedding)+ BiLSTM | 0.2982 | 0.6959 | 0.3168 | 0.4229 | 0.8643 |
| | BERT(Embedding)+ BiLSTM+ Att | 0.3082 | 0.6939 | 0.3168 | 0.4229 | 0.8643 |
| RoBERTa Contextual Embedding | RoBERTa (Embedding)+BiGRU | 0.2764 | 0.7217 | 0.2977 | 0.3892 | 0.8637 |
| | RoBERTa(Embedding) BiGRU+ Att | 0.2913 | 0.7228 | 0.3148 | 0.4146 | 0.8704 |
| | RoBERTa (Embedding)+BiLSTM | 0.2493 | 0.6687 | 0.2644 | 0.3440 | 0.8583 |
| | RoBERTa(Embedding) BiLSTM+ Att | 0.2913 | 0.7228 | 0.3148 | 0.4146 | 0.8704 |

Table 4.16 below presents the per-label scores of the best performing model FastText+BiLSTM on the CBET dataset in measure of recall, precision, f1-score, and AUC.

Table 4.16: Per-label scores of FastText + BiLSTM on CBET dataset.

| Metrics | Anger | Fear | Joy | Love | Sadness | Surprise | Thankfulness | Disgust | Guilt | Macro Avg |
|---|---|---|---|---|---|---|---|---|---|---|
| Precision | 0.4350 | 0.5435 | 0.4326 | 0.6001 | 0.3511 | 0.5448 | 0.5734 | 0.5241 | 0.4691 | 0.4878 |
| Recall | 0.5477 | 0.7608 | 0.7140 | 0.5772 | 0.5853 | 0.5529 | 0.7566 | 0.6043 | 0.5562 | 0.6287 |
| F1-score | 0.4849 | 0.6340 | 0.5388 | 0.5884 | 0.4389 | 0.5488 | 0.6524 | 0.5613 | 0.5090 | 0.5507 |
| AUC | 0.8326 | 0.9126 | 0.8688 | 0.8478 | 0.8230 | 0.8485 | 0.9157 | 0.8883 | 0.8525 | 0.8658 |

The per-label scores of the FastText + BiLSTM model presented in Table 4.16 above showcase a high sensitivity in recognizing "Fear", "Joy", and "Thankfulness" labels by recording high recall values of 0.7608, 0.7140, and 0.7566, respectively. Conversely, the "Anger" and "Sadness" labels scored the worst in terms of F1-score and recall metrics, shocasing limited recognition ability of this model in classifying these emotions.

### *Approach 3: Hybrid Transformer-based and BiLSTM/BiGRU and Self-attention*

The experimental results of Approach 3 when applied on the CBET dataset are presented in Table 4.17 below. Competitive scores are depicted between hybrid models using BERT and RoBERTa models as backbone. The RoBERTa model combined with BiLSTM and MLP yielded the best f1-score (0.6425) while maintaining a high AUC (0.9182). The RoBERTa + BiGRU-MLP obtained a similar F1-score of 0.6422, whereas RoBERTa + BiLSTM+Att+MLP achieved a slightly higher AUC (0.9198) but at lower F1-score (0.6392).

Table 4.17: Macro-average scores obtained by Approach 3 on CBET dataset.

| Approach | Model | Accuracy | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|---|
| Hybrid Transformer-based [BERT] and BiLSTM/BiGRU models | BERT+BiGRU +MLP | 0.5676 | 0.6890 | 0.6039 | 0.6395 | 0.9160 |
| | BERT+BiGRU + Att+MLP | 0.5745 | 0.6575 | 0.6154 | 0.6329 | 0.9085 |
| | BERT+BiLSTM+ MLP | 0.5754 | 0.6752 | 0.6089 | 0.6391 | 0.9095 |
| | BERT+BiLSTM+ Att+MLP | 0.5573 | 0.6976 | 0.5867 | 0.6350 | 0.9172 |
| Hybrid Transformer-based [RoBERTa] and BiLSTM/BiGRU models | RoBERTa+BiGRU+ MLP | 0.5735 | 0.6891 | 0.6062 | 0.6422 | 0.9166 |
| | RoBERTa+BiGRU+ Att+MLP | 0.5754 | 0.6869 | 0.6088 | 0.6417 | 0.9114 |
| | **RoBERTa+BiLSTM +MLP** | **0.5697** | **0.6944** | **0.6036** | **0.6425** | **0.9182** |
| | RoBERTa+BiLSTM+ Att+MLP | 0.5600 | 0.7036 | 0.5906 | 0.6392 | 0.9198 |

Table 4.18 presents per-label scores obtained by the best-performing model RoBERTa + BiLSTM + MLP on the CBET dataset. Similar to the aformentioned approaches, the best hybrid transformer-based model combining BiLSTM and MLP revealed stronger sensitivity and precision in recognizing the "Fear" and "Thankfulness" labels by obtaining the higher values of f1-score (0.769 for "fear" and 0.766 for "Thankfulness"). The same for "Sadness", "Anger", and "Guilt" labels where the model performed the worst by yielding the lowest f1 scores of 0.495, 0.552, and 0.600, respectively.

Table 4.18: Per-label scores of RoBERTa + BiLSTM + MLP on CBET dataset.

| Metrics | Anger | Fear | Joy | Love | Sadness | Surprise | Thankfulness | Disgust | Guilt | Macro Avg |
|---|---|---|---|---|---|---|---|---|---|---|
| Precision | 0.633 | 0.846 | 0.702 | 0.751 | 0.616 | 0.661 | 0.775 | 0.656 | 0.693 | 0.7036 |
| Recall | 0.490 | 0.704 | 0.555 | 0.548 | 0.414 | 0.647 | 0.757 | 0.671 | 0.530 | 0.5906 |
| F1-score | 0.552 | 0.769 | 0.620 | 0.633 | 0.495 | 0.654 | 0.766 | 0.663 | 0.600 | 0.6392 |
| AUC | 0.882 | 0.951 | 0.923 | 0.915 | 0.887 | 0.919 | 0.956 | 0.939 | 0.906 | 0.9198 |

### *Approach 4: Transfer Learning from GoEmotion + ML/DL models BERT RoBERTa based models*

The following table presents the results obtained with Approach 4 on the CBET dataset. Compared to previous approaches based on fine-tuning transformer models (approaches 1 and 2), transferring knowledge from GoEmotion dataset enabling the classification of user-generated texts derived from social media platforms into 27 emotion labels through feature extraction using BERT and RoBERTa models revealed to be less effective when applied on the CBET dataset. The best performing model XGBoost trained using GoEmotion BERT feature representations demonstrated lower performance scores, especially for Recall and f1-score. Whereby, this model obtained 0.3658 for f1-score as the best overall score achieved by Approach 4 is considered limited when compared to 0.6482 obtained by RoBERTa in Approach 1 and 0.6435 reached by hybrid RoBERTa + BiLSTM + MLP in Approach 3. Moreover, the results recorded by all models in Approach 4 are still limited even when compared to models based on FastText embeddings in Approach 2 (best f1-score is 0.5507).

Table 4.19: Macro-average scores obtained by Approach 4 on CBET dataset.

| Approach | Model | Accuracy | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|---|
| Transfer Learning from GoEmotion [BERT] + ML/DL Models | Logistic Regression | 0.2655 | 0.6808 | 0.2937 | 0.4000 | 0.8390 |
| | SVM | 0.1427 | 0.8290 | 0.1461 | 0.2196 | 0.7831 |
| | Random Forest | 0.1332 | 0.8601 | 0.1410 | 0.2306 | 0.8167 |
| | **XGBoost** | **0.2340** | **0.7329** | **0.2548** | **0.3658** | **0.8430** |
| | MLP | 0.1259 | 0.6932 | 0.1349 | 0.2156 | 0.7990 |
| | BiLSTM | 0.1348 | 0.6916 | 0.1445 | 0.2289 | 0.7980 |
| | BiGRU | 0.1413 | 0.6935 | 0.1526 | 0.2369 | 0.7993 |
| Transfer Learning from GoEmotion [RoBERTa] + ML/DL Models | Logistic Regression | 0.2509 | 0.6707 | 0.2784 | 0.3825 | 0.8342 |
| | SVM | 0.2185 | 0.7268 | 0.2350 | 0.3431 | 0.8306 |
| | Random Forest | 0.1769 | 0.8009 | 0.1853 | 0.2904 | 0.8272 |
| | XGBoost | 0.1769 | 0.8009 | 0.1853 | 0.2904 | 0.7336 |
| | MLP | 0.1676 | 0.7140 | 0.1763 | 0.2752 | 0.8184 |
| | BiLSTM | 0.1782 | 0.6996 | 0.1909 | 0.2887 | 0.8197 |
| | BiGRU | 0.1762 | 0.7116 | 0.1880 | 0.2887 | 0.8197 |

The fine-grained evaluation of the best-performing model in Approach 4 through per-label scores can be found in Table 4.20 below. The presented results indicate the weak sensitivity of the XGBoost model trained on the GoEmotion transferred feature representations in recognizing the nine labels composing the CBET dataset by recording low recall values ranging from 0.1142 for "Guilt" label to 0.4222 for "Thankfulness" label, while maintaining high precision values. This heavy imbalance between recall and precision scores generated low f1-scores, demonstrating the brittle effectiveness of transfer learning via feature extraction between GoEmotion and CBET datasets.

Table 4.20: Performance of XGBoost using BERT (GoEmotions) CLS embeddings

| Metrics | Anger | Fear | Joy | Love | Sadness | Surprise | Thankfulness | Disgust | Guilt | Macro Avg |
|---|---|---|---|---|---|---|---|---|---|---|
| Precision | 0.7213 | 0.8038 | 0.6943 | 0.7410 | 0.6162 | 0.7409 | 0.8214 | 0.6574 | 0.6577 | 0.7329 |
| Recall | 0.1704 | 0.3760 | 0.2708 | 0.3082 | 0.1230 | 0.2247 | 0.4222 | 0.2643 | 0.1142 | 0.2548 |
| F1-score | 0.2756 | 0.5124 | 0.3896 | 0.4353 | 0.2050 | 0.3448 | 0.5577 | 0.3771 | 0.1945 | 0.3658 |
| AUC | 0.7884 | 0.8808 | 0.8554 | 0.8425 | 0.7968 | 0.8348 | 0.9011 | 0.8703 | 0.8170 | 0.8430 |

### 4.3.4 Global Discussion of Multi-Label Emotion Recognition Approaches on the CBET Dataset

The evaluation of the CBET dataset, which is multi-label in nature with 9 emotion categories and 81,164 samples, reveals that the fine-tuning of RoBERTa (Approach 1) achieves the best overall performance (F1 = 0.6482; AUC = 0.9203), leveraging the power of transformers to capture nuanced contextual relationships.(see figure4.9).

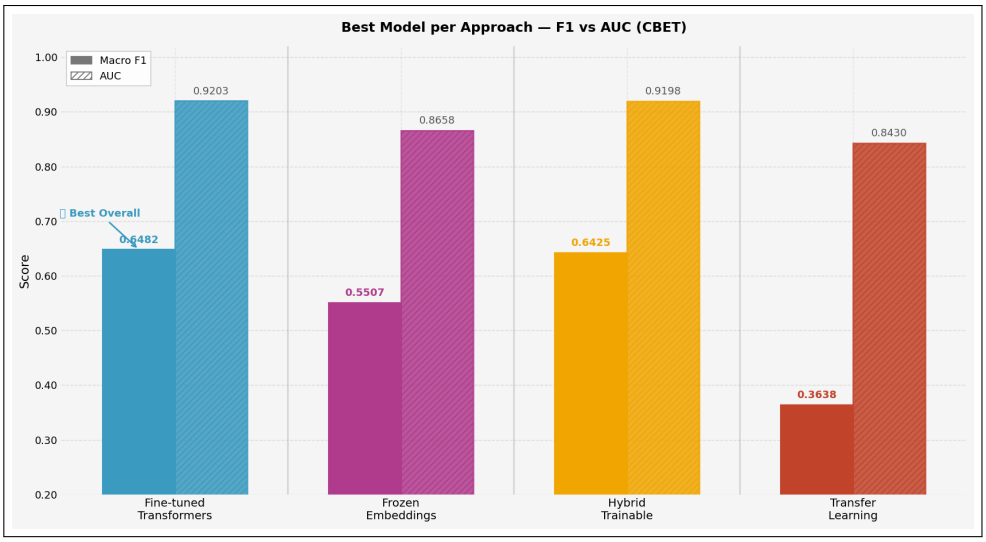



Figure 4.9: Comparative performance of all four approaches: Macro F1 and AUC.

Similar to the results obtained on the ISEAR dataset, models using FastText embeddings in Approach 2 consistently outperformed models based on BERT/RoBERTa embeddings. This superior performance can be attributed to the lower model complexity of FastText, which represents input tokens using lower-dimensional vectors compared to the high-dimensional contextualized representations generated by BERT/RoBERTa. Although the CBET dataset contains a larger number of samples than ISEAR, the similarity between certain emotions (e.g., "Anger" and "Sadness") and its multi-label nature still make the classification task challenging for models with a large number of parameters. In addition, the small learning rate (2e-5) used for training these large models based on BERT/RoBERTa embeddings may have slowed convergence and limited their ability to reach optimal solutions.

Comparison of Approach 3 results, mainly between hybrid BERT/RoBERTa models with and without self-attention highlights the superior performance of models which do not involve self-attention (best results obtained by RoBERTa + BiLSTM + MLP with F1 = 0.6392 and AUC = 0.9198). This degradation of performance after integrating self-attention indicates that this mechanism already inherent in RoBERTa does not significantly boost performance when included on top of the BiLSTM layer in Approach 3.

Approach 4, which transferred emotion-related knowledge learned by BERT and RoBERTa on GoEmotion dataset to CBET via feature extraction, showcased the lowest performance across F1 and AUC scores (see figure4.10). This transfer learning technique based on extracting emotion-aware feature representations using GoEmotion's BERT/RoBERTa models is demonstrated to be less effective than fine-tuning of these models on the target dataset.

A per-class analysis highlights that ”Thankfulness” and ”Fear” are the best-performing labels, with the former benefiting from strong lexical markers like ”grateful” or ”thanks,” and the latter from intense, unambiguous vocabulary that is less prone to confusion, whereas ”Sadness” and ”Anger” emerge as the most challenging classes due to their frequent implicit expression, semantic overlap with other negative emotions, and the prevalence of irony or sarcasm that complicates detection.

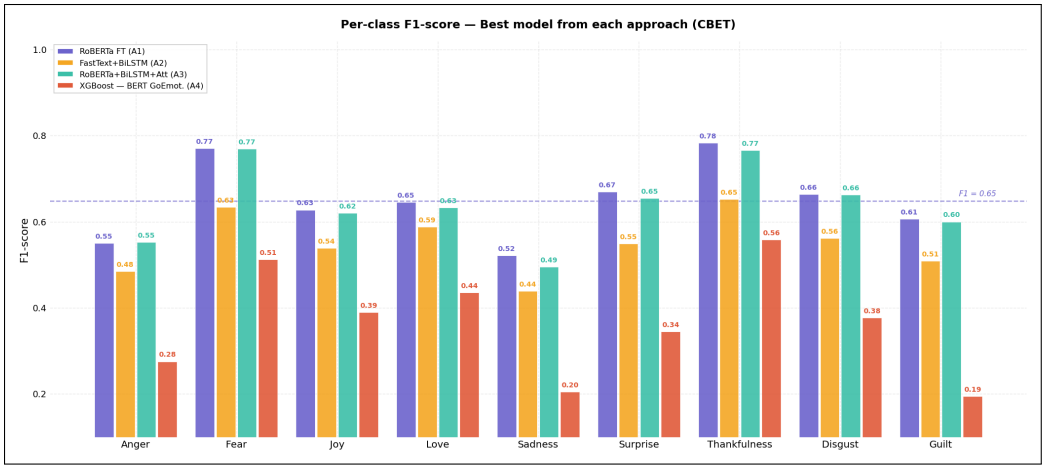



Figure 4.10: Per-class F1 scores across of approaches 1,2,3 and 4 on CBET dataset

Experimental results on the ISEAR and CBET datasets indicate that the multi-label classification task of CBET (9 emotion labels and 81,164 samples) is more challenging than the multi-class classification task of ISEAR (7 emotion classes and 7,516 samples) for all proposed models. This increased difficulty stems from the multi-label nature of CBET, which requires the simultaneous prediction of multiple co-occurring emotions within a single text, whereas ISEAR involves assigning only one emotion category to each text instance.

### 4.3.5 Multi-Label Emotion Classification of Arabic Texts : SemEval-2018 Dataset

#### *Approach 1: Fine-tuning Transformer-based Models*

The following table presents the results obtained with Approach 1 on the SemEval-2018 dataset evaluated across the macro-averaged performance metrics. These results demonstrated the superior performance of MarBERT over AraBERT CAMeLBERT and Xlm-RoBERTa across all the performance metrics when fine-tuned on SemEval-2018 dataset for fine-grained multi-label emotion classification. MarBERT model ,it is the best model outperformed in terms of f1-score and AUC by obtaining 0.6251 and 0.9159, respectively.

Per-label scores for the best-performing model MARBERT in measure of precision, recall, f1-score and AUC are presented in Table 4.22 below.

Table 4.21: Macro-average scores obtained by approach 1 on SemEval-2018 dataset.

| Approach | Model | Accuracy | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|---|
| Fine-Tuning Transformer-Based Models | AraBERT | 0.1658 | 0.5759 | 0.7728 | 0.5924 | 0.9017 |
| | XLM-RoBERTa | 0.0000 | 0.2094 | 0.6815 | 0.2770 | 0.4722 |
| | CAMeLBERT | 0.2007 | 0.6126 | 0.7864 | 0.5945 | 0.9138 |
| | **MARBERT** | **0.2234** | **0.6285** | **0.8030** | **0.6251** | **0.9159** |

Table 4.22: Per-label performance of MarBERT on SemEal-2018 dataset.

| Metrics | Anger | Anticipation | Disgust | Fear | Joy | Love | Optimism | Pessimism | Sadness | Surprise | Trust | Macro Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Precision | 0.8060 | 0.2752 | 0.5172 | 0.7339 | 0.7500 | 0.7407 | 0.7024 | 0.4415 | 0.7711 | 0.1471 | 0.2706 | 0.6285 |
| Recall | 0.8779 | 0.5357 | 0.8182 | 0.8163 | 0.9329 | 0.8805 | 0.8252 | 0.6535 | 0.7828 | 0.3333 | 0.5476 | 0.8030 |
| F1-score | 0.8404 | 0.3636 | 0.6338 | 0.7729 | 0.8315 | 0.8046 | 0.7588 | 0.5270 | 0.7769 | 0.2041 | 0.3622 | 0.6251 |
| AUC | 0.9420 | 0.7713 | 0.8964 | 0.9371 | 0.9570 | 0.9524 | 0.9366 | 0.7888 | 0.9077 | 0.7914 | 0.8558 | 0.9159 |

The per-label scores of Table 4.22 above obtained by fine-tuning MARBERT on SemEal-2018 dataset demonstrate the high performance of this model in recognizing ."Anger" and "Joy" labels. Whereby, the "Anger" label recorded an F1-score of (0.8404)and a precision of (0.8060), while "Joy" reached a high recall of (0.9329) and a high F1-score of (0.8315). In contrast, the "Surprise" and "Trust" showcased to be the weakest labels by obtaining the lowest recall (0.3333) and F1-score (0.2041) for "Surprise" and the lowest precision (0.2706) with a low recall (0.5476) for "Trust". Guilt also received moderate scores across all metrics.

### *Approach 2: Static and Contextual Word Embeddings with BiGRU/BiLSTM and Self-attention*

The following table presents the results obtained by Static Vs. Contextual Word Embedding with BiGRU/BiLSTM and Self-attention for Multi-label Emotion Classification on the SemEval-2018 dataset (Approach 2) evaluated across multiple macro-averaged performance metrics. These results show the best performing models to be AraBERT (not trainable) plus (BiLSTM) plus attention, yielding an F1 score of 0.5317 and area under the curve (AUC) of 0.8600. MarBERT models show lower performance metrics than FastText, AraBERT, and CAMeLBERT based models.

Table 4.23: Macro-average scores obtained with Approach 2 on SemEval-2018 dataset.

| Approach | Model | Accuracy | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|---|
| FastText Embedding | FastText+BiGRU | 0.0070 | 0.3047 | 0.7215 | 0.3929 | 0.6946 |
| | FastText+BiGRU+Att | 0.1675 | 0.2142 | 0.1653 | 0.0968 | 0.5565 |
| | FastText+BiLSTM | 0.0035 | 0.3017 | 0.7328 | 0.3818 | 0.6966 |
| | FastText+BiLSTM+Att | 0.0000 | 0.2248 | 0.6219 | 0.2480 | 0.5245 |
| AraBERT Contextual Embedding | AraBERT(Embedding)+ BiGRU | 0.0384 | 0.4199 | 0.7977 | 0.5037 | 0.8320 |
| | AraBERT(Embedding)+ BiGRU+ Att | 0.0716 | 0.4593 | 0.7955 | 0.5238 | 0.8594 |
| | AraBERT(Embedding)+ BiLSTM | 0.0558 | 0.4179 | 0.8121 | 0.5063 | 0.8459 |
| | **AraBERT(Embedding + BiLSTM + Att** | **0.0506** | **0.4486** | **0.8098** | **0.5317** | **0.8600** |
| CAMeLBERT Contextual Embedding | CAMeLBERT (Embedding)+ BiGRU | 0.0192 | 0.4073 | 0.8053 | 0.4967 | 0.8362 |
| | CAMeLBERT(Embeddin + BiGRU + Att | 0.0593 | 0.4313 | 0.7796 | 0.5060 | 0.8417 |
| | CAMeLBERT (Embedding) + BiLSTM | 0.0262 | 0.4086 | 0.7917 | 0.4920 | 0.8371 |
| | CAMeLBERT(Embeddin + BiLSTM + Att | 0.0366 | 0.4124 | 0.8015 | 0.5047 | 0.8422 |
| MarBERT Contextual Embedding | MarBERT(Embedding) + BiGRU | 0.0140 | 0.3228 | 0.7328 | 0.4240 | 0.7406 |
| | MarBERT(Embedding) + BiGRU + Att | 0.0558 | 0.3801 | 0.7109 | 0.4412 | 0.7727 |
| | MarBERT(Embedding) + BiLSTM | 0.0541 | 0.3432 | 0.8136 | 0.4567 | 0.7968 |
| | MarBERT(Embedding) + BiLSTM + Att | 0.0175 | 0.3612 | 0.7162 | 0.4422 | 0.7627 |

Table 4.24 below presents the per-label scores of the best performing model FastText+BiLSTM on the SemEval-2018 dataset in measure of recall, precision, f1-score, and AUC.

Table 4.24: Per-label scores of AraBERT (Not Trainable) + BiLSTM + Self-Attention on SemEval-2018 dataset.

| Metrics | Anger | Anticipation | Disgust | Fear | Joy | Love | Optimism | Pessimism | Sadness | Surprise | Trust | Macro Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Precision | 0.6803 | 0.1469 | 0.4271 | 0.3401 | 0.6837 | 0.6716 | 0.6421 | 0.3441 | 0.6468 | 0.0802 | 0.1600 | 0.4486 |
| Recall | 0.8592 | 0.5536 | 0.7455 | 0.8571 | 0.8171 | 0.8616 | 0.8531 | 0.8425 | 0.7677 | 0.8667 | 0.6667 | 0.8098 |
| F1-score | 0.7593 | 0.2322 | 0.5430 | 0.4870 | 0.7444 | 0.7548 | 0.7327 | 0.4886 | 0.7021 | 0.1469 | 0.2581 | 0.5317 |
| AUC | 0.8914 | 0.7121 | 0.8370 | 0.8356 | 0.9275 | 0.9159 | 0.9073 | 0.7591 | 0.8453 | 0.8706 | 0.7538 | 0.8600 |

The table above demonstrates that the label with the best performance in terms of recall (0.8592) is the "Anger" class, which has the highest recall rate. This indicates that the model's ability to correctly identify actual instances of anger is very high. In contrast," Surprise" had the lowest recall value, with a score of 0.1469 . Therefore, Surprise is considered the most challenging emotion for the model to detect.

### *Approach 3: Hybrid Transformer-based and BiLSTM/BiGRU and Self-attention*

The experimental results of Approach 3 when applied on the SemEval-2018 dataset are presented in Table 4.25 below. Competitive scores are depicted between hybrid models using AraBERT/MarBERT and CAMeLBERTmodels as backbone.MarBERT (trainable) + BiGRU + Attention is the best-performing model. It had the highest value for the F1 metric at 0.6210 and AUC at 0.9178. In contrast to AraBERT- and CAMeLBERT-based models, MarBERT models consis- tently showed better performance than all other models.

Table 4.25: Macro-average scores obtained by Approach 3 on SemEval-2018 dataset.

| Approach | Model | Accuracy | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|---|
| Hybrid Transformer-based [AraBERT] and BiLSTM/BiGRU models | AraBERT + BiGRU | 0.1134 | 0.5142 | 0.7374 | 0.5408 | 0.8652 |
| | AraBERT + BiGRU + Att | 0.1414 | 0.5059 | 0.7442 | 0.5338 | 0.8670 |
| | AraBERT + BiLSTM | 0.1152 | 0.5095 | 0.7909 | 0.5494 | 0.8745 |
| | AraBERT + BiLSTM + Att | 0.0785 | 0.4977 | 0.7479 | 0.5461 | 0.8627 |
| Hybrid Transformer-based [CAMeLBERT] and BiLSTM/BiGRU models | CAMeLBERT+BiGRU | 0.1798 | 0.5761 | 0.7653 | 0.5713 | 0.8974 |
| | CAMeLBERT + BiGRU + Att | 0.1728 | 0.5836 | 0.7721 | 0.5866 | 0.8999 |
| | CAMeLBERT+BiLSTM | 0.1483 | 0.5543 | 0.7585 | 0.5644 | 0.8855 |
| | CAMeLBERT+BiLSTM + Att | 0.1571 | 0.5461 | 0.7683 | 0.5805 | 0.8868 |
| Hybrid Transformer-based [MarBERT] and BiLSTM/BiGRU models | MarBERT+BiGRU | 0.2094 | 0.6103 | 0.7872 | 0.6028 | 0.9115 |
| | **MarBERT + BiGRU + Att** | **0.2077** | **0.5987** | **0.8242** | **0.6210** | **0.9178** |
| | MarBERT + BiLSTM | 0.2077 | 0.6167 | 0.7977 | 0.6150 | 0.9141 |
| | MarBERT + BiLSTM + Att | 0.1850 | 0.6098 | 0.7525 | 0.5789 | 0.9037 |

Table 4.26 presents per-class scores of precision, recall, f1-score, and AUC metrics achieved by MarBERT (Trainable) + BiGRU + Self-Attention on SemEval-2018 dataset.Based on the results above,MarBERT (Trainable) + BiGRU + Self-Attention model achieved the highest values for prescion and F1- score for the "Anger " class by obtaining 0.7966 and 0.8374, respectively. Whereas, this model recorded the lowest scores for "Surprise " class (0.3333 for recall and 0.1961 for f1-score).

Table 4.26: Per-label scores MarBERT (Trainable) + BiGRU + Self-Attention on SemEval-2018 dataset.

| Metrics | Anger | Anticipation | Disgust | Fear | Joy | Love | Optimism | Pessimism | Sadness | Surprise | Trust | Macro Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Precision | 0.7966 | 0.2081 | 0.5172 | 0.8152 | 0.7598 | 0.7500 | 0.6625 | 0.3864 | 0.6856 | 0.1399 | 0.3519 | 0.5987 |
| Recall | 0.8826 | 0.5536 | 0.8182 | 0.7653 | 0.9217 | 0.8689 | 0.8881 | 0.8031 | 0.8333 | 0.3333 | 0.4762 | 0.8242 |
| F1-score | 0.8374 | 0.3024 | 0.6348 | 0.7905 | 0.8320 | 0.8057 | 0.7582 | 0.5227 | 0.7527 | 0.1961 | 0.4040 | 0.6210 |
| AUC | 0.9479 | 0.7326 | 0.9081 | 0.9406 | 0.9645 | 0.9538 | 0.9446 | 0.8001 | 0.9076 | 0.7981 | 0.8346 | 0.9178 |

### 4.3.6 Global Discussion of Multi-Label Emotion Recognition Approaches on the SemEval-2018 Dataset

The evaluation of the SemEval-2018 Arabic dataset, which presents a challenging multi-label classification task (11 labels, 2,866 samples), reveals competitive results between fine-tuning MarBERT (Approach1), achieving the best F1-score 0.6251, and fine-tuning Hybrid MarBERT + BiGRU + Attention (Approach 3), obtaining the best AUC score 0.9159(see figure **??**).

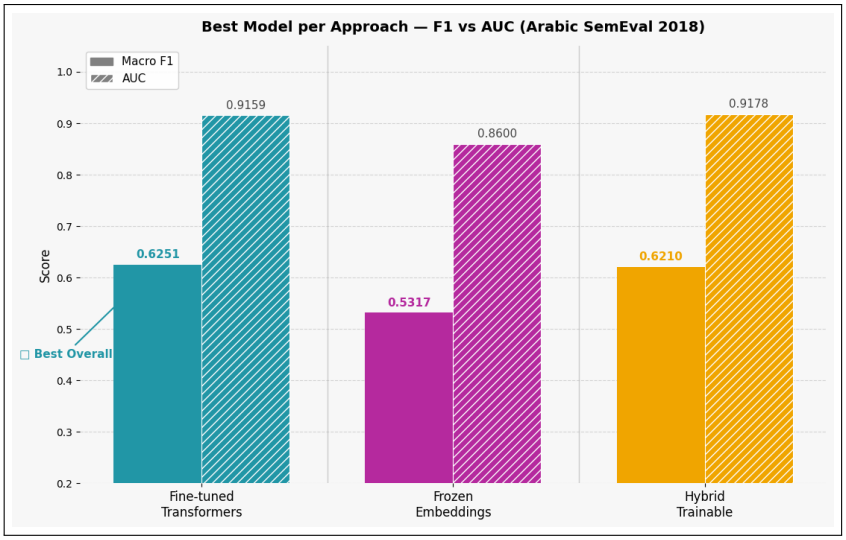



Figure 4.11: Macro F1 and AUC scores for Fine-tuned Transformers, Frozen Embeddings, and Hybrid Trainable approaches on the SemEval-2018 dataset.

In contrast to English datasets, experimental results revealed the outperformance of contextual AraBERT embeddings combined with BiLSTM and self-attention over the rest of models using MarBERT, CAMelBERT, and FastText embeddings.This superior performance can be attributed to ability of AraBERT to provide richer contextual representations as it is pretrained on large Arabic textual data, compared to Arabic FastText relying on static word vectors that cannot handle context-dependent meanings in Arabic.

The best performing model in Approach 3, which integrates MarBERT with BiGRU and self-attention, confirms that self-attention has a distinctly positive effect, yielding competitive results (F1 = 0.6210; AUC = 0.9178), as the attention layer allows the model to weigh relevant sequential features across the bidirectional GRU outputs. Compared to the hybrid transformer-based models which do not employ self-attention, the MarBERT + BiGRU + attention model recorded an improvement of 3.32

A per class analysis highlights that ”Anger” and ”Joy” achieve the highest scores, largely because these classes contain substantially larger numbers of training examples, enabling the model to learn robust and discriminative patterns (see figure**??**). On the other hand, the ”Surprise” and ”Trust” perform poorly due to severe underrepresentation (”Surprise” = 60 samples and ”Trust” = 156 samples).

Finally, the Arabic multi-label setup (11 labels, 2,866 samples) underperforms compared to the English multi-label configuration (9 labels, 81,164 samples). This performance gap is due to three main factors: first, the English dataset is significantly larger, providing richer training signals and more diverse linguistic patterns; second, increasing the number of labels from 9 to 11 inherently complicates the classification task by expanding the possible label combinations and increasing inter-label ambiguity; and third, Arabic remains a relatively low-resource language compared to English, with fewer pretrained models, limited annotated datasets, and greater morphological complexity, all of which make it inherently more challenging for DL models to achieve comparable performance.




Figure 4.12: Class-wise F1-scores for MARBERT FT (A1), AraBERT+BiLSTM+Att (A2), and MarBERT+BiGRU+Att (A3) on the SemEval dataset.

### 4.3.7 Comparison with Literature

The following table presents a comparison between the results of three previous studies and the proposed models, across three different datasets: ISEAR, CBET, and SemEval-2018.

Table 4.27: Comparison with existing literature

| **Study** | **Dataset** | **Size** | **Model** | **Acc** | **Prec** | **Rec** | **F1** |
|---|---|---|---|---|---|---|---|
| [36] | ISEAR | 5,431 | Lexicon-Enhanced Multi-Task CNN (LMT-CNN) | — | 0.7300 | 0.7257 | 0.7350 |
| [37] | ISEAR | — | BERT | 0.63 | — | — | — |
| [35] | CBET | 76,860 | BERT Base model | 0.609 | 0.608 | 0.611 | 0.609 |
| [36] | CBET | 51,240 | Lexicon-Enhanced Multi-Task CNN (LMT-CNN) | — | 0.6385 | 0.6331 | 0.6284 |
| [37] | CBET | >30k | BERT | 0.56 | — | — | — |
| [43] | SemEval-2018 | 4,381 | MARBERT, Bi-LSTM, and Bi-GRU | 0.485 jaccard | 0.522 | 0.559 | 0.509 |
| Proposed (English) | ISEAR | 7,516 | Hybrid RoBERTa (Trainable) + BiGRU | 0.7307 | 0.7347 | 0.7301 | 0.7308 |
| Proposed (English) | CBET | 81,164 | Fine-tuning RoBERTa | 0.5749 | 0.6994 | 0.6059 | 0.6482 |
| Proposed (Arabic) | SemEval-2018 | 2,866 | Fine-tuning MarBERT | 0.2234 | 0.6285 | 0.8030 | 0.6251 |

The proposed model (Hybrid Transformer RoBERTa + BiGRU) outperforms [36] on the ISEAR dataset not only because it relies on a hybrid architecture that combines a trainable RoBERTa transformer with a BiGRU layer, allowing it to extract both contextual representations and sequential relationships more effectively than the lexicon-enhanced CNN which uses only fixed lexical representations, but also because the two models are evaluated on different numbers of classes. Specifically, the [36] used a dataset with only 5 emotions, whereas our model is evaluated on 7 emotions, making the classification task more challenging. Despite this increased difficulty, our model achieves F1 0.7308 vs. 0.7350
Additionally, our model also outperforms another recent study [37] that employed a standard BERT model, which reported an accuracy of only 0.63 on a similar task. In contrast, our hybrid approach achieves a significantly higher accuracy of0.7307 demonstrating the clear advantage of combining RoBERTa with BiGRU to capture both deep contextual and sequential dependencies more effectively than using BERT alone.
The proposed model (Fine-tuning transformer RoBERTa) outperforms [36] on the CBET dataset because it leverages a larger dataset (81,164 samples compared to 51,240) and uses a fine-tuned RoBERTa transformer, which captures deeper contextual representations than the lexicon-enhanced CNN. As a result, the proposed model achieves higher precision (0.6994 vs. 0.6385) and a better F1-score (0.6482 vs. 0.6284), demonstrating the effectiveness of fine-tuning large pre-trained transformers on emotion recognition tasks.

In addition to this, our model also surpasses two other recent studies that both employed standard BERT models. Specifically, compared to [35], which reported an F1 of only 0.609 , our model achieves a significantly higher F1 of 0.6482. our approach outperforms[37], where the authors obtained an Accuracy of 0.56 using BERT, whereas our model achieves0.5749 accuracy.
These results confirm that fine-tuning RoBERTa, which benefits from improved pre-training strategies and larger-scale data, provides substantial advantages over standard BERT-based models for emotion recognition.

The proposed model (Fine-tuning MarBERT) on the SemEval dataset achieves a competitive performance compared to MARBERT, Bi-LSTM, and Bi-GRU not only because it leverages a transformer specifically pre-trained on Arabic language data, enabling better capture of linguistic nuances and emotional expressions in Arabic text, but also because the two models are evaluated on different dataset sizes and architectures. Specifically, [43] employed a combination of MARBERT, Bi-LSTM, and Bi-GRU on a larger dataset of 4,381 samples, whereas our model is fine-tuned on a smaller dataset of only 2,866 samples, making the training process more challenging due to limited data availability. Despite this increased difficulty, our model achieves a macro F1-score of 0.6251 vs. 0.701.

# General Conclusion

This thesis investigated the application of deep learning techniques for fine-grained emotion recognition from textual data, addressing the limitations of conventional machine learning methods in handling noisy social media text, emotion variability, and limited annotated datasets, particularly for morphologically rich languages like Arabic. Through a comprehensive literature review and a methodologically grounded experimental study, the work evaluated and proposed four distinct approaches across three datasets: ISEAR (English, multi-class), CBET (English, multi-label), and SemEval-2018 (Arabic, multi-label). These approaches included fine-tuning transformer-based models, static and contextual word embeddings with BiLSTM/BiGRU architectures, hybrid transformer-recurrent frameworks, and transfer learning from GoEmotions.

After systematically evaluating all approaches on each dataset, the best performing model was identified for each one. On the English multi-class ISEAR dataset, the hybrid architecture combining trainable RoBERTa with BiGRU achieved the highest macro F1-score (0.7308). On the English multi-label CBET dataset, fine-tuning the RoBERTa transformer alone proved to be the most effective, obtaining the best macro F1-score (0.6482). On the Arabic SemEval-2018 dataset, fine-tuning the MARBERT transformer alone achieved the best macro F1-score (0.6251). These results demonstrate that the optimal architecture depends on the specific characteristics of each dataset, including language, task type, and annotation schema.

The proposed models demonstrated competitive performance in terms of F1-score, precision, recall, and AUC, offering improved robustness in recognizing complex and subtle emotional expressions. The integration of preprocessing strategies, early stopping, and appropriate loss functions proved essential in optimizing performance, particularly in the presence of class imbalance and limited data. The two most effective configurations were fine-tuning Transformer models alone, which proved sufficient for the CBET and SemEval datasets, and the hybrid architecture combining trainable RoBERTa with BiGRU, which excelled on the multi-class ISEAR dataset.

Despite promising results, one of the persistent challenges highlighted in this study is the high computational cost required to implement advanced transformer architectures. While many recent state-of-the-art models outperform simpler networks in emotion recognition accuracy, their real-world deployment is constrained by the need for powerful hardware and long training timesresources that are not always available in academic or research environments. This gap underscores the importance of developing efficient yet effective models that can balance performance and practicality.

Overall, this research contributes to the growing field of AI-driven natural language processing by offering insights, benchmarks, and practical considerations for deploying

deep learning solutions in text-based emotion recognition across different languages and annotation schemas. The findings confirm that hybrid architectures are particularly effective for multi-class classification, while fine-tuned transformers alone suffice for multi-label tasks and for morphologically rich languages like Arabic when specialized pre-trained models are available. Future work may focus on lightweight models, domain adaptation techniques, cross-lingual transfer learning, expanding Arabic datasets, and clinical validation to enhance the usability and reliability of such systems in real-world applications such as mental health monitoring and human-computer interaction. Future work may also focus on data expansion through collecting larger and more diverse real-world datasets while addressing class imbalance, exploring large language models such as GPT-4 and LLaMA under few-shot and zero-shot settings for emotion classification, and multimodal analysis that combines text, audio, and facial expressions for richer emotion understanding.

# Bibliography


[1] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). *A training algorithm for optimal margin classifiers*. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT)*, pp. 144–152. https://doi.org/10.1145/130385.130401

[2] Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. https://doi.org/10.1145/2939672.2939785

[3] Miao, M., Yan, W., Gao, X., Yang, L., Zhou, J., & Zhang, W. (2025). *Attention-Based CNN-BiGRU-Transformer Model for Human Activity Recognition*. *Applied Sciences*, 15(23), 12592.

[4] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). *LSTM: A search space odyssey*. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

[5] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186. https://arxiv.org/abs/1810.04805

[6] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv preprint arXiv:1907.11692. https://arxiv.org/abs/1907.11692

[7] Antoun, W., Baly, F., & Hajj, H. (2020). *AraBERT: Transformer-based Model for Arabic Language Understanding*. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*, pp. 9–15. https://aclanthology.org/2020.osact-1.2/

[8] Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2021). *The interplay of variant, size, and task type in Arabic pre-trained language models*. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP)*, pp. 92–104. https://aclanthology.org/2021.wanlp-1.11

[9] Saxena, A., Khanna, A., & Gupta, D. (2020). *Emotion recognition and detection methods: A comprehensive survey*. Journal of Artificial Intelligence and Systems, 2(1), 53–79. https://iecscience.org/jpapers/46

[10] Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). *Current state of text sentiment analysis from opinion to emotion mining. ACM Computing Surveys (CSUR)*, 50(2), 1–33. https://dl.acm.org/doi/10.1145/3057270

[11] Goyal, S., & Tiwari, N. (2017). *Emotion recognition: a literature survey.* Int. J. Technol. Res. Eng, 4(9), 1502–1524. https://www.academia.edu/33169558/EMOTION$_R$ECOGNITION$_{AL}$ITERATURE$_S$URVE

[12] Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., Scherer, K., Tomita, M., & Tzavaras, A. (1987). *Universals and cultural differences in the judgments of facial expressions of emotion.* Journal of Personality and Social Psychology, 53(4), 712–717. https://doi.org/10.1037/0022-3514.53.4.712

[13] Plutchik, R. (2001). *The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice.* American Scientist, 89(4), 344–350. http://www.jstor.org/stable/27857503

[14] Can, Y. S., Mahesh, B., & André, E. (2023). *Approaches, applications, and challenges in physiological emotion recognition—a tutorial overview.* Proceedings of the IEEE, 111(10), 1287–1313. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10171229

[15] Bachate, M., & Suchitra, S. (2025). *Sentiment analysis and emotion recognition in social media: A comprehensive survey.* Applied Soft Computing, 174, 112958. https://www.sciencedirect.com/science/article/abs/pii/S1568494625002698

[16] Domingos, P. (2012). *A few useful things to know about machine learning.* Communications of the ACM, 55(10), 78–87. https://doi.org/10.1145/2347736.2347755

[17] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387

[18] Breiman, L. (2001). *Random forests. Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[19] LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning. Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

[20] Du, K. L., Leung, C. S., Mow, W. H., & Swamy, M. N. S. (2022). *Perceptron: Learning, generalization, model selection, fault tolerance, and role in the deep learning era. Mathematics, 10*(24), 4730. https://doi.org/10.3390/math10244730

[21] Nti, I. K., Samuel, A.-A., & Michael, A. (2019). *Predicting monthly electricity demand using soft-computing. Rep. Econ. Financ*, 6, 1967–1973.

[22] Du, P., Li, X., & Gao, Y. (2020). *Dynamic music emotion recognition based on CNN-BiLSTM.* In *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)* (pp. 1372–1376). IEEE. https://ieeexplore.ieee.org/abstract/document/9141729

[23] Eren, A. O., & Sert, M. (2020). Audio captioning using gated recurrent units. *arXiv preprint arXiv:2006.03391*. https://arxiv.org/abs/2006.03391

[24] Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). *Bag of tricks for efficient text classification. Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 2, short papers*, 427–431. https://aclanthology.org/E17-2068.pdf

[25] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need. Advances in Neural Information Processing Systems, 30*. https://arxiv.org/abs/1706.03762

[26] Huang, Z., Low, C., Teng, M., Zhang, H., Ho, D. E., Krass, M. S., & Grabmair, M. (2021). *Context-aware legal citation recommendation using deep learning*. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law* (pp. 79–88). Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3462757.3466066

[27] Habbat, N., Anoun, H., & Hassouni, L. (2021). *Arabertopic: A neural topic modeling approach for news extraction from arabic facebook pages using pre-trained bert transformer model. International Journal Of Computing and Digital System*, 14. https://www.uob.edu.bh/journals/ijcds.aspx

[28] Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2021). *"The interplay of variant, size, and task type in Arabic pre-trained language models."* In *Proceedings of the sixth Arabic natural language processing workshop*, pp. 92–104. https://aclanthology.org/2021.wanlp-1.10/

[29] Rathwa, H., Mishra, P., & Malviya, S. (2025). *"AGI" Team at SHROOM-CAP: Data-Centric Approach to Multilingual Hallucination Detection using XLM-RoBERTa*. In *CHOMPS 2025*, pp. 96–100. https://aclanthology.org

[30] Scherer, K. R., & Wallbott, H. G. (1994). *Evidence for universality and cultural variation of differential emotion response patterning*. Journal of Personality and Social Psychology, 66(2), 310–328. https://psycnet.apa.org/buy/1994-29654-001

[31] Shahraki, A. G., & Zaiane, O. R. (2017). *Lexical and learning-based emotion mining from text*. In *Proceedings of the international conference on computational linguistics and intelligent text processing (CICLing 2017)*, Vol. 9, pp. 24–55. https://webdocs.cs.ualberta.ca/ zaiane/postscript/CICLing17.pdf

[32] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). *GoEmotions: A dataset of fine-grained emotions*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pp. 4040–4054. https://aclanthology.org/2020.acl-main.372.pdf

[33] Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). *SemEval-2018 task 1: Affect in tweets*. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pp. 1–17. https://aclanthology.org/S18-1001.pdf

[34] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). *MELD: A multimodal multi-party dataset for emotion recognition in conversations*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pp. 527–536. https://aclanthology.org/P19-1050.pdf

[35] Kodiyala, V. S., & Mercer, R. E. (2021). *Emotion recognition and sentiment classification using BERT with data augmentation and emotion lexicon enrichment*. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 191–198. https://ieeexplore.ieee.org/abstract/document/9679993

[36] Dong, Y., & Zeng, X. (2022). *Lexicon-enhanced multi-task convolutional neural network for emotion distribution learning*. *Axioms*, 11(4), 181. https://www.mdpi.com/2075-1680/11/4/181

[37] Andrikakis, E., Perikos, I., Paraskevas, M., & Hatzilygeroudis, I. (2023). *Text analysis and recognition of emotional content using deep learning methods and BERT*. In *2023 IEEE/ACIS 23rd International Conference on Computer and Information Science (ICIS)*, pp. 37–42. https://ieeexplore.ieee.org/abstract/document/10210232/

[38] Ahanin, Z., Ismail, M. A., Singh, N. S. S., & AL-Ashmori, A. (2023). *Hybrid feature extraction for multi-label emotion classification in English text messages*. *Sustainability*, 15(16), 12539. https://www.mdpi.com/2071-1050/15/16/12539

[39] Wang, K., Jing, Z., Su, Y., & Han, Y. (2024). *Large language models on fine-grained emotion detection dataset with data augmentation and transfer learning*. *arXiv preprint arXiv:2403.06108*. https://arxiv.org/pdf/2403.06108

[40] Hinojosa Lee, M. C., Braet, J., & Springael, J. (2024). *Performance metrics for multilabel emotion classification: comparing micro, macro, and weighted f1-scores*. *Applied Sciences*, 14(21), 9863. https://www.mdpi.com/2076-3417/14/21/9863

[41] Maazallahi, A., Asadpour, M., & Bazmi, P. (2025). *Advancing emotion recognition in social media: A novel integration of heterogeneous neural networks with fine-tuned language models*. *Information Processing & Management*, 62(2), 103974. https://www.sciencedirect.com/science/article/abs/pii/S0306457324003339

[42] Alshdaifat, E., Altarawneh, E., Aloshaibat, M. A., Hussein, A., Dawood, M., & Tahaineh, T. (2026). *From machine learning to heterogeneous models: A comprehensive study on fine-grained emotion detection in Arabic text*. *Engineering, Technology & Applied Science Research*, 16(2), 34396–34402. https://www.etasr.com/index.php/ETASR/article/view/17705

[43] Mansy, A., Rady, S., & Gharib, T. (2022). *An ensemble deep learning approach for emotion detection in Arabic tweets*. *International Journal of Advanced Computer Science and Applications*, 13(4), 980–989. https://doi.org/10.14569/IJACSA.2022.01304112

[44] Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., & Chen, Y.-S. (2018). *CARER: Contextualized affect representations for emotion recognition*. In *Proceedings of the*

*2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pp. 3687–3697. https://aclanthology.org/D18-1404.pdf

[45] Al-Khatib, A., & El-Beltagy, S. R. (2017). *Emotional tone detection in Arabic tweets*. In *International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017)*, pp. 105–114. https://link.springer.com/chapter/10.1007/978-3-319-77116-8$_8$