People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Setif 1 University - Ferhat Abbas
Faculty of Sciences
Department of Computer Science



Doctoral Thesis

Presented by:

Hanane BOUTOUTA

As a requirement for obtaining the **Doctoral degree** in Computer Science

Option: Software Engineering & Human-Computer Interface

Implicit Emotion Recognition in Arabic Text

Committee of examiners

Nadjet Kamel	Prof.	Setif 1 University - Ferhat Abbas	President
Abdelaziz Lakhfif	Prof.	Setif 1 University - Ferhat Abbas	Supervisor
Chahrazed Mediani	MCA.	Setif 1 University - Ferhat Abbas	Co-Supervisor
Sadik Bessou	Prof.	Setif 1 University - Ferhat Abbas	Examiner
Farid Nouioua	Prof.	Bordj Bou Arreridj University	Examiner
Hadda Cherroun	Prof.	Ammar Telidji University - Laghouat	Examiner
Allaoua Refoufi	Prof.	Retired Professor	Invited

Defended on: july 2026

Implicit Emotion Recognition in Arabic Text, © july 2026

Author:
Hanane BOUTOUTA

Supervisors:
Prof. Abdelaziz Lakhfif
Dr. Chahrazed Mediani

Institute:
Setif 1 University - Ferhat Abbas, Setif, Algeria

ABSTRACT

Emotion recognition has become a critical task in Natural Language Processing (NLP) and affective computing, driven by the rapid increase in emotionally rich digital content and the growing demand for emotionally intelligent systems. While explicit emotion detection has advanced considerably, recognizing implicit emotions —those inferred from context— remains largely underexplored, especially for low-resource languages such as Arabic. This thesis investigates Implicit Emotion Recognition (IER) in Arabic text from three perspectives: resource development, modeling strategy, and linguistic diversity. First, we introduce Arabic Implicit Emotion Recognition (AIER), a publicly available dataset composed of real-world Arabic reviews written in Modern Standard Arabic (MSA) and annotated with emotion roles. We evaluate its quality using classical machine learning, deep learning, Pretrained Language Models (PLMs), and Large Language Models (LLMs), showing that contextualized transformer-based models consistently outperform traditional approaches, while nuanced implicit cues remain difficult to capture. Second, we propose a hybrid framework combining the transformer's bidirectional encoding with Recurrent Neural Network (RNN)'s sequential modeling to better capture implicit emotion in Arabic. Evaluated on a newly compiled Arabic Twitter corpus under both implicit and explicit recognition settings, the model consistently outperforms established deep learning baselines. Compared to State-of-the-Art (SOTA) models for explicit emotion recognition, our model achieves superior performance even when applied to IER. Third, we assess the emotional capabilities of modern LLMs by evaluating their performance on multi-label emotion recognition in the low-resource Algerian Arabic dialect and by benchmarking their implicit reasoning abilities in English under settings where lexical cues are absent. Together, these studies provide a cross-lingual assessment of LLMs' affective understanding. Overall, this thesis advances Arabic NLP and affective computing by introducing new public resources and comparative evaluations across modeling paradigms, providing an empirical foundation for future research on implicit emotion recognition.

Keywords: Implicit Emotion Recognition, Affective Computing, NLP, Arabic, Machine Learning, Deep learning, Transformer, Large Language Models

RÉSUMÉ

La reconnaissance des émotions est devenue une tâche essentielle du traitement automatique du langage naturel (TALN) et de l'informatique affective, portée par l'augmentation rapide des contenus numériques riches en émotions et par la demande croissante de systèmes capables de comprendre les émotions humaines. Bien que la détection des émotions explicites ait connu des avancées considérables, la reconnaissance des émotions implicites — celles qui sont déduites du contexte — demeure largement inexplorée, en particulier pour les langues à faibles ressources telles que l'arabe. Cette thèse étudie la reconnaissance des émotions implicites dans le texte arabe selon trois perspectives: le développement de ressources, les stratégies de modélisation et la diversité linguistique. Tout d'abord, nous présentons le corpus de reconnaissance implicite des émotions arabes (AIER), un jeu de données public composé d'avis réels rédigés en arabe standard moderne et annotés selon les rôles émotionnels. Nous évaluons sa qualité à l'aide des méthodes classiques d'apprentissage automatique, de l'apprentissage profond, des modèles de langue pré-entraînés et des grands modèles de langue génératifs (LLMs), en montrant que les modèles basés sur les transformeurs contextualisés surpassent systématiquement les approches traditionnelles, tandis que les indices émotionnels implicites les plus subtils restent difficiles à capturer. Ensuite, nous proposons un cadre hybride combinant l'encodage bidirectionnel des transformeurs avec la modélisation séquentielle des réseaux de neurones récurrents afin de mieux capturer les émotions implicites en arabe. Évalué sur un nouveau corpus de tweets arabes compilé pour cette étude, dans des scénarios de reconnaissance des émotions implicites et explicites, le modèle surpasse systématiquement les principales approches de référence en apprentissage profond. Comparé aux modèles de l'état de l'art pour la reconnaissance explicite des émotions, notre modèle atteint des performances supérieures, y compris lorsqu'il est appliqué à la reconnaissance des émotions implicites. Enfin, nous évaluons les capacités émotionnelles des LLMs en mesurant leurs performances en reconnaissance multi étiquette des émotions dans le dialecte algérien, une langue à faibles ressources, ainsi qu'en évaluant leurs capacités de raisonnement implicite en anglais dans des contextes où les indices lexicaux sont absents. Ensemble, ces études offrent une évaluation translinguistique de la compréhension affective des LLMs. Dans l'ensemble, cette thèse contribue à faire progresser le TALN arabe et l'informatique affective en introduisant de nouvelles ressources publiques et des évaluations comparatives de différents paradigmes de modélisation, fournissant ainsi une base empirique solide pour les futures recherches sur la reconnaissance des émotions implicites.

Mots-clés: Reconnaissance Des Émotions Implicites, Informatique Affective, Traitement Automatique du Langage Naturel, Language Arabe, Apprentissage Automatique, Apprentissage Profond, Transformer, Grands Modèles de Langue

ملخص

أصبح التعرف على المشاعر مهمة بالغة الأهمية في معالجة اللغات الطبيعية والحوسبة العاطفية، مدفوعةً بالزيادة السريعة في المحتوى الرقمي الغني بالمشاعر والطلب المتزايد على أنظمة قادرة على فهم المشاعر الإنسانية. وبينما شهد اكتشاف المشاعر الصريحة تقدماً ملحوظاً، لا يزال التعرف على المشاعر الضمنية ـ تلك المستنتجة من السياق ـ مجالاً غير مستكشف إلى حد كبير، لا سيما في اللغات ذات الموارد المحدودة كاللغة العربية. تبحث هذه الأطروحة في التعرف على المشاعر الضمنية في النصوص العربية من ثلاثة جوانب: تطوير الموارد، واستراتيجيات النمذجة، والتنوع اللغوي. أولاً، نقدم مجموعة بيانات التعرف على المشاعر الضمنية في اللغة العربية (AIER) ، وهي مجموعة بيانات عامة تتألف من آراء واقعية مكتوبة باللغة العربية الفصحى الحديثة ومصنفة وفقاً للأدوار العاطفية. نقوم بتقييم جودتها باستخدام أساليب التعلم الآلي التقليدية، والتعلم العميق، ونماذج اللغة المدربة مسبقاً، ونماذج اللغة التوليدية الكبيرة (LLMs) ، مما يُظهر أن النماذج القائمة على المحولات السياقية تتفوق باستمرار على الأساليب التقليدية، في حين أن أدق الإشارات العاطفية الضمنية لا تزال صعبة الرصد. بعد ذلك، نقترح إطار عمل هجينًا يجمع بين ترميز المحول ثنائي الاتجاه ونمذجة الشبكة العصبية المتكررة المتسلسلة (RNN) لتحسين فهم المشاعر الضمنية في اللغة العربية. وقد تم تقييم النموذج على مجموعة جديدة من التغريدات العربية جُمعت خصيصًا لهذه الدراسة، وفي سيناريوهات تتضمن التعرف على المشاعر الضمنية والصريحة، تفوق النموذج باستمرار على أساليب التعلم العميق الرائدة. وبالمقارنة مع أحدث النماذج في مجال التعرف على المشاعر الصريحة، يحقق نموذجنا أداءً فائقًا، حتى عند تطبيقه على التعرف على المشاعر الضمنية. وأخيرًا، نقوم بتقييم القدرات العاطفية للنماذج اللغوية الكبيرة الحديثة من خلال قياس أدائها في مهمة التعرّف متعدد التسميات على المشاعر في اللهجة العربية الجزائرية، وهي لهجة منخفضة الموارد، وكذلك من خلال تقييم قدراتها على الاستدلال الضمني باللغة الإنجليزية في سياقات تغيب فيها المؤشرات المعجمية الصريحة. وتوفر هذه الدراسات مجتمعةً تقييمًا عابرًا للغات لقدرات في الفهم العاطفي. بشكل عام، تساهم هذه الأطروحة في تطوير معالجة اللغة الطبيعية العربية والحوسبة العاطفية من خلال تقديم موارد عامة جديدة وتقييمات مقارنة لنماذج النمذجة المختلفة، مما يوفر أساسًا تجريبيًا متينًا للبحوث المستقبلية حول التعرف الضمني على المشاعر.

الكلمات المفتاحية: التعرف على المشاعر الضمنية، الحوسبة العاطفية، معالجة اللغة الطبيعية، اللغة العربية، التعلم الآلي، التعلم العميق، نموذج المحول، النماذج اللغوية الكبيرة

PUBLICATIONS

This thesis is based on five peer-reviewed articles. The articles are identical to their original published form, with the exception of minor changes in style, format, and the correction of typographical errors. The following articles are included in the thesis:

[1] H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, "Enhancement of Implicit Emotion Recognition in Arabic Text: Annotated Dataset and Baseline Models", IEEE Access, vol. 13, pp. 165096–165116, 2025, DOI: 10.1109/ACCESS.2025.3611337.

[2] H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, "A Transformer-based Hybrid Model for Implicit Emotion Recognition in Arabic Text", Engineering, Technology & Applied Science Research, vol. 15, no. 3, pp. 23834–23839, 2025. DOI: 10.48084/etasr.10261. [Online]. Available: https://etasr.com/index.php/ETASR/article/view/10261.

[3] H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, "From Context to Emotion: Leveraging LLMs for Recognizing Implicit Emotions", in Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025), Southern Denmark University, Odense, Denmark: Association for Computational Linguistics, Aug. 2025, pp. 399–409. [Online]. Available: https://aclanthology.org/2025.icnlsp-1.39/.

[4] H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, Do LLMs Grasp Emotion or Just Keywords? A Benchmark for Implicit Emotion Recognition." In: 2025 1st International Conference on Artificial Intelligence and Cyber Physical Systems (AICPS), 2025, pp. 104–109. DOI: https://doi.org/10.1109/AICPS66617.2025.11513483.

[5] H. Boutouta, F. Senator, Z. Lakab, C. Mediani, and A. Lakhfif, "Multi-Label Emotion Recognition in Low-Resource Dialects: A Case Study on Algerian Arabic with Large Language Models", Procedia Computer Science, vol. 275, pp. 817–824, 2026, 7th International Conference on AI in Computational Linguistics, issn: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2026.01.094. 2026.01.094. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050926000943.

Note on Publications Not Included in This Thesis: In addition to the above papers, the following works have been published during the research period for this thesis; they have contributed to our deeper understanding of the work towards this thesis; however, these publications do not fit into the narrative of this thesis and have not been included in the text.

[1] F. Senator, A. Lakhfif, I. Zenbout, H. Boutouta, and C. Mediani, "Leveraging chatgpt for enhancing arabic nlp: Application for semantic role labeling and cross-lingual annotation projection", IEEE Access, vol. 13, pp. 3707–3725, 2025. doi: 10.1109/ACCESS.2025.3525493 [1].

[2] F. Senator, H. Boutouta, A. Lakhfif, and C. Mediani, "Semantic role labeling of arabic emotional text in tweets", in 2023 International Conference on Advances in Electronics, Control and Communication Systems (ICAECCS), 2023, pp. 1–6. doi: 10.1109/ICAECCS56710.2023.10104732 [2].

[3] Senator, F., Boutouta, H., Lakhfif, A., Mediani, C. (2025). The impact of semantic roles in emotion detection using the AraBERT model. In Proceedings of the Mediterranean Conference on Computer Science and Artificial Intelligence (MCCSAI 2025) (accepted, to appear). Springer.

ACKNOWLEDGMENTS

First and foremost, all praise and gratitude are due to ***ALLAH***, the Almighty, for granting me strength, patience, guidance, and perseverance throughout this long and challenging academic journey. Without His blessings and mercy, the completion of this work would not have been possible.

الْحَمْدُ لِلَّهِ الَّذِي بِنِعْمَتِهِ تَتِمُّ الصَّالِحَاتُ

I would like to express my deepest gratitude to my supervisor, **Prof. Abdelaziz Lakhfif**, for his invaluable supervision, support, and tutelage during my PhD study over the past few years. His continuous support and encouragement help me through the difficulties. I also extend my sincere appreciation to my co-supervisor **Dr. Chahrazed Mediani** for her support, valuable advice, and insightful contributions.

I would like to thank all members of the Department of Computer Science and all **my teachers** for their direct and indirect contributions to the success of my studies.

My heartfelt thanks also go to the members of the **dissertation committee** for accepting to evaluate this work and for their time, valuable comments, and constructive suggestions.

On a personal note, I would like to express my heartfelt appreciation to my source of inspiration, **my father**, and **the pure soul of my beloved mother** for their unconditional love, sacrifices, and unwavering belief in me. Without your encouragement and sacrifices, this achievement would not have been possible.

My sincere thanks go to my husband, **Abdelhak**, for his patience, understanding, and continuous support throughout this journey. I am especially grateful to my daughters, the light of my life, **Farah** and **Rital**, whose love, smiles, and presence have been my greatest motivation and inspiration. Your trust in me has been the foundation for this thesis.

I extend special thanks to my dear sister, **Fatima**, whose emotional support, encouragement, and constant presence have meant more to me than words can express. I also extend my heartfelt thanks to my brothers, **Ramy** and **Abderazek**, for their support and encouragement throughout my academic path.

Finally, I would like to thank my dear friends, **Ferial Senator**, **Zakia Lakab**, **Ghozlene Chenni**, and **Imene Zenbout**, who have been constant companions in my daily life and unwavering supporters throughout.

Hanane BOUTOUTA, *july 2026*

DEDICATION

إلى أمي الغالية رحمها الله وطيّب ثراها.

إلى أبي الغالي أطال الله في عمره.

أسأل الله تعالى أن يتقبل هذا العمل صدقة جارية عنهما

وأن يجعلها فى ميزان حسناتهما

STATEMENT OF AUTHORSHIP

I, Hanane BOUTOUTA, hereby declare that this thesis, titled *Implicit Emotion Recognition in Arabic Text*, and the work presented herein are the result of my own original research. I confirm that this work was done mainly while in candidature for a research degree at the Setif 1 University - Ferhat Abbas.

Except where reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or in part from a thesis accepted for the award of any other degree or diploma. No other person's work has been used without due acknowledgment in the main text of the thesis. This thesis has not been submitted for the award of any academic degree or qualification at this or any other tertiary institution.



Hanane BOUTOUTA july, 2026

CONTENTS

Abstract . . . iii

Publications . . . ix

Acknowledgments . . . xi

Declaration of Authorship . . . xv

Contents . . . xv

List of Figures . . . xx

List of Tables . . . xxi

Acronyms . . . xxiii

1 GENERAL INTRODUCTION . . . 3
- 1.1 Background and Motivation . . . 3
- 1.2 Research Questions . . . 5
- 1.3 Research Objectives . . . 6
- 1.4 Main Contributions . . . 7
- 1.5 Thesis Structure . . . 7

2 THEORETICAL BACKGROUND . . . 9
- 2.1 Emotions in Computational Linguistics . . . 10
  - 2.1.1 Theories of Emotion . . . 10
  - 2.1.2 Emotion modeling . . . 11
  - 2.1.3 Emotion Expression in Text . . . 12
- 2.2 Arabic Language Characteristics . . . 12
  - 2.2.1 Arabic varieties . . . 13
  - 2.2.2 Key Challenges in Arabic Emotion Recognition . . . 13
- 2.3 Explicit vs. Implicit Emotion Expression . . . 16
  - 2.3.1 Emotion Words . . . 16
  - 2.3.2 Explicit Emotion Expression . . . 17
  - 2.3.3 Implicit Emotion Expression . . . 18
  - 2.3.4 Linguistic Cues of Implicit Emotion . . . 20
- 2.4 Computational Approaches in NLP . . . 21
  - 2.4.1 Text Representations . . . 21
  - 2.4.2 Classical Machine Learning Methods . . . 28

2.4.3 Traditional Deep Learning Approaches . . . 30
2.4.4 Transformers-Based Models . . . 33
2.4.5 Evaluation Metrics for Emotion Classification . . . 41
2.5 Gaps and Challenges in Implicit Emotion Recognition . . . 43
2.6 Conclusion . . . 44

3 LITERATURE REVIEW . . . 45
3.1 Resources and Datasets . . . 45
3.1.1 Arabic Emotion Datasets . . . 46
3.1.2 Implicit Emotion Datasets . . . 47
3.2 Implicit Emotion Recognition: State of the Art . . . 49
3.3 General Considerations in the Thesis . . . 52
3.4 Conclusion . . . 55

4 BUILDING THE FOUNDATION: ANNOTATED DATASET AND BASELINE MODELS FOR ARABIC IMPLICIT EMOTION RECOGNITION . . . 57
4.1 Introduction . . . 57
4.2 Datasets in Emotion recognition . . . 59
4.3 The AIER dataset . . . 60
4.3.1 Data Collection . . . 61
4.3.2 Annotation Process . . . 63
4.3.3 Validation . . . 63
4.3.4 Annotator Agreement . . . 64
4.3.5 Implicit Scenario Simulation . . . 66
4.3.6 Dataset Description . . . 66
4.4 Implicit Emotion Modeling . . . 68
4.4.1 BERT-based models . . . 68
4.4.2 Generative LLMs . . . 70
4.4.3 Classical ML models . . . 72
4.4.4 DL-based models . . . 73
4.4.5 Evaluation . . . 74
4.5 Results and Discussion . . . 74
4.5.1 Performance of ML classifiers . . . 74
4.5.2 Performance of DL classifiers . . . 76
4.5.3 Performance of BERT-based Models . . . 78
4.5.4 Performance of Generative LLMs . . . 80
4.5.5 Cost-effectiveness Analysis . . . 82
4.6 Conclusions . . . 84

5 A TRANSFORMER-BASED HYBRID MODEL FOR IMPLICIT EMOTION RECOGNITION IN ARABIC TWEETS . . . 87
5.1 Introduction . . . 87
5.2 Methodology . . . 89

5.2.1 Baselines . . . 89
5.2.2 Proposed Model . . . 89
5.2.3 Hyperparameter Settings . . . 92
5.3 Experiments . . . 92
5.3.1 Dataset . . . 92
5.3.2 Experimental Setup and Evaluation . . . 94
5.4 Results and Discussion . . . 94
5.4.1 Baseline Results . . . 95
5.4.2 Proposed Model vs. Baseline Results . . . 95
5.4.3 Implicit vs. Explicit Affective Cues . . . 96
5.4.4 Per-emotion Performance . . . 96
5.4.5 Error Analysis . . . 97
5.4.6 Comparison with state-of-the-art Methods . . . 98
5.5 Conclusion . . . 99

6 DECODING HIDDEN EMOTIONS ACROSS LANGUAGES: LLMS-BASED INSIGHTS FROM ARABIC DIALECTS TO ENGLISH TEXTS . . . 101
6.1 Introduction . . . 101
6.2 Multi-label Emotion Recognition in Algerian Dialect . . . 103
6.2.1 Introduction . . . 103
6.2.2 Experiments . . . 104
6.2.3 Results and Discussion . . . 107
6.2.4 Conclusion . . . 111
6.3 Benchmarking Implicit Reasoning in English . . . 111
6.3.1 Introduction . . . 111
6.3.2 Experiments . . . 113
6.3.3 Results and Discussion . . . 120
6.3.4 Conclusion . . . 124
6.4 Conclusion and Cross-linguistic Insights . . . 125

7 GENERAL CONCLUSION . . . 127

BIBLIOGRAPHY . . . 131

LIST OF FIGURES

Figure 2.1 The Transformer architecture, consisting of an encoder and a decoder (Vaswani et al. [68]. 27
Figure 4.1 An overview of the AIER dataset building process 61
Figure 4.2 Observed agreement result for each emotion. 65
Figure 4.3 Text length distribution. 68
Figure 4.4 Emotion distribution within the AIER dataset. 69
Figure 4.5 Classification results of classical ML-based models on AIER dataset, for BoW and TF-IDF feature extraction. 75
Figure 4.6 Classification results of DL-based models on AIER dataset. 77
Figure 4.7 Classification results of implicit emotion using Fine-tuned pre-trained models. 79
Figure 4.8 Classification results of generative LLMs on AIER dataset. 81
Figure 4.9 Performance comparison of different classes of classifiers. 83
Figure 5.1 Architecture of the proposed Hybrid Model 90
Figure 5.2 Emotions distribution within the AIEmoD dataset. 94
Figure 6.1 Distribution of emotions in the IEST dataset. 116
Figure 6.2 Distribution of emotion in the ISEAR dataset. 116
Figure 6.3 Prompt example for IER. The examples inside the dashed box are demonstrations of the few-shot setting and would be removed under the zero-shot setting. 119
Figure 6.4 Comparative analysis of Zero-Shot and Few-Shot Results across ISEAR and IEST datasets. 121
Figure 6.5 Performance Variance across the ISEAR and IEST Datasets (LLMs results are reported using the few-shot setting). 122

LIST OF TABLES

Table 3.1 Related resources for Implicit Emotion Recognition. 49
Table 3.2 A summary of previous works on IER in textual data. 53
Table 4.1 Overview of publicly available Arabic and non-Arabic datasets used in emotion recognition, indicating the corresponding task (EER: Explicit Emotion Recognition; IER: Implicit Emotion Recognition) and language. 60
Table 4.2 Query words used for data collection. 62
Table 4.3 Agreement statistics for emotion classification. 64
Table 4.4 Pairwise observed agreement for each emotion class. 65
Table 4.5 Selective examples from the AIER dataset. 67
Table 4.6 General statistics of the AIER dataset. 67
Table 4.7 Dataset statistics across different emotion classes. 68
Table 4.8 Dataset Train/Validation/Test Split. 69
Table 4.9 Tested and optimal Hyperparameter values. 70
Table 4.10 Hyperparameter used for Classical ML models. 72
Table 4.11 Performance comparison of classical ML-based classifiers on the AIER dataset. 75
Table 4.12 Performance comparison of ML-based Approaches with BoW per emotion class. 76
Table 4.13 Performance of DL-based classifiers on the AIER dataset. 77
Table 4.14 Performance comparison of DL-based approaches per emotion class. 77
Table 4.15 performance comparison of different transformer-based models, on the AIER dataset. 78
Table 4.16 Performance comparison of BERT-based models per emotion class. . 79
Table 4.17 Performance comparison of different generative LLMs, on the AIER dataset. 80
Table 4.18 Performance comparison of generative LLMs per emotion class. . . 81
Table 4.19 Cost-effectiveness comparison 85
Table 5.1 The hyperparameters of our model 92
Table 5.2 Comparative results of the proposed model against baseline models on the implicit SemEval-2018, AETD, and AIEmoD datasets. . . . 95
Table 5.3 Explicit vs. Implicit performance of the proposed model on the SemEval-2018, AETD, and AIEmoD datasets 96
Table 5.4 Per-emotion evaluation results of the proposed model on the SemEval-2018, AETD, and AIEmoD datasets 97

Table 5.5 Examples of misclassified tweets and their corresponding error causes 98
Table 5.6 Comparison with the state-of-the-art methods on the AETD dataset. 99
Table 6.1 Hyperparameters used for fine-tuning pretrained models. 106
Table 6.2 Performance results of fine-tuned PLMs and zero-shot LLMs on multi-label emotion classification. 108
Table 6.3 Multi-label classification performance: Comparing Fully Correct Samples, Exact-match Accuracy, Jaccard Index, and Hamming Loss across models. 109
Table 6.4 Per-emotion F1-scores (%) across models. 109
Table 6.5 Error analysis of texts misclassified by the Gemini model. 110
Table 6.6 Performance comparison of baselines and top-ranked systems on SemEval-2025 Task 11 in the Algerian Arabic dataset (Arq). 111
Table 6.7 Example tweets from the IEST dataset, with the emotion word masked as [#TARGETWORD#]. 114
Table 6.8 Example from the ISEAR dataset for each emotion label. 115
Table 6.9 Key differences between the IEST and ISEAR datasets. 115
Table 6.10 Hyperparameters tuning 117
Table 6.11 Comparative results of different models on both ISEAR and IEST datasets. Acc, F1, Prec, and Rec denote accuracy, macro F1 score, precision, and recall, respectively. 121
Table 6.12 Per-emotion Result on the ISEAR dataset. 123
Table 6.13 Per-emotion Result on the IEST dataset. 124

ACRONYMS

NLP	Natural Language Processing
IER	Implicit Emotion Recognition
NLU	Natural Language Understanding
AIER	Arabic Implicit Emotion Recognition
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly optimized BERT approach
MLM	Masked Language Modeling
NSP	Next Sentence Prediction
ML	Machine Learning
DL	Deep Learning
MSA	Modern Standard Arabic
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
XGBoost	eXtreme Gradient Boosting
RF	Random Forest
DT	Decision Tree
LR	Logistic Regression
MNB	Multinomial Naïve Bayes
NB	Naïve Bayes
BoW	Bag-of-Words
TF-IDF	Term Frequency-Inverse Document Frequency
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Units
BiLSTM	Bidirectional Long Short-Term Memory
BiGRU	Bidirectional Gated Recurrent Units
RNN	Recurrent Neural Network
ISEAR	International Survey on Emotion Antecedents and Reactions
VSM	Vector Space Model

IEST	Implicit Emotions Shared Task
CIEA	Chinese Implicit Emotion Analysis
TL	Transfer Learning
SOTA	State-of-the-Art
LLMs	Large Language Models
PLMs	Pretrained Language Models
AETD	Arabic Emotion Tweets Dataset
AIEmoD	Arabic Implicit Emotion Dataset
ANN	Artificial Neural Networks
GPT	Generative Pre-trained Transformer

1 GENERAL INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

The automatic understanding of human emotions has long been considered a fundamental goal in the development of intelligent systems capable of interacting naturally with humans. With the exponential growth of digital communication over the past two decades, vast amounts of emotionally rich textual data are generated daily across social media platforms, online reviews, and conversational systems. This growth has created both the necessity and the opportunity for computational systems capable of automatically detecting and interpreting the emotional content in textual data. Consequently, the field of affective computing, introduced by Picard [3], has gained increasing attention, alongside key areas in NLP such as sentiment analysis and emotion recognition [4, 5], which aim to model and analyze affective information at scale.

Emotion recognition refers to the task of automatically identifying and classifying the emotional content conveyed in a given text [3]. It extends beyond binary sentiment classification by assigning fine-grained emotion categories, such as Ekman's six basic emotions of joy, sadness, anger, fear, disgust, and surprise [6], to written expressions. As a research area, it has undergone substantial development since its early formulations, evolving from rule-based approaches [7] and classical machine learning methods [8] to sophisticated deep learning architectures [9] and, most recently, large-scale PLMs [10] and LLMs [11]. These advances have yielded impressive results on standard benchmarks, with emotion-aware systems now widely applied in diverse real-world applications, including mental health monitoring [12], customer feedback analysis [13], human-computer interaction [14], and the automated moderation of harmful online content [15].

However, the majority of existing emotion recognition systems focus primarily on explicit emotions, which are conveyed directly through emotional words (e.g., happy, angry) [8, 16–18]. In contrast, a significant portion of human emotional expression is implicit, where emotions are not directly stated but must be inferred from linguistic cues such as contextual descriptions, metaphorical expressions, situational events, figurative language, or culturally specific expressions [16, 18–23]. Consider the statement "She stared at the empty mailbox for hours"; no emotion word is present, yet the frustration and disappointment it implies are immediately legible to any competent reader. This capacity for recognizing such implicit emotional cues, effortless in human comprehension, remains a significant challenge for current systems, as it requires contextual reasoning and world knowledge that go far beyond simple lexical pattern matching [24, 25].

This challenge becomes considerably more difficult when the target language is Arabic. As a Semitic language and one of the most widely spoken in the world, Arabic has over 420 million speakers across more than 22 countries and serves as the liturgical language of approximately 1.8 billion Muslims worldwide [26]. However, Arabic presents a unique set of linguistic characteristics and complexities that make it particularly challenging for NLP research fields[27, 28].

One of the most distinctive features of Arabic is its diglossia [29], which refers to the coexistence of two broad language varieties used in different social contexts. MSA is the formal variety used in writing, journalism, education, and official discourse, while a range of regional colloquial dialects, such as Egyptian, Levantine, Gulf, and Algerian Arabic, are used in everyday speech and informal digital communication. These dialects differ from MSA and from each other in pronunciation, morphology, syntax, and vocabulary, sometimes to the point where speakers of different dialect groups struggle to understand one another. Regarding digital content, particularly on social media, it often mixes MSA and dialectal forms within a single text and may further incorporate code-switching with different languages such as English and French, the use of Romanized Arabic script (Arabizi), and a heavy reliance on culturally specific rhetorical conventions, including proverbs, religious formulae, and inter-textual allusions [30].

Beyond diglossia, Arabic is characterized by rich morphology, complex syntax, and a high degree of orthographic ambiguity, where words with identical spellings may carry entirely different meanings depending on context, diacritical marks, and punctuation [31].

This ambiguity is further compounded by the language's agglutinative nature, which allows complex words to be segmented and interpreted in multiple ways, making accurate lexical analysis particularly demanding [32]. These structural and morphological characteristics make it considerably harder to build comprehensive linguistic resources and labeled datasets, amplifying the problem of data scarcity, a challenge that is already acute and hinders the development of reliable, accurate, and powerful tools [30].

These challenges extend to the broader landscape of Arabic NLP research. Despite the demographic and cultural significance of Arabic, it remains substantially underrepresented compared to highly resourced languages such as English, Chinese, and other European languages [28, 30, 32]. Nevertheless, the Arabic NLP community has made remarkable progress in recent years in several core tasks, including named entity recognition, machine translation, and summarization [31, 32]. In the affective domain, a growing body of work has addressed Arabic sentiment analysis and emotion recognition, and several annotated corpora have been constructed for this purpose [9, 33]. The advent of PLMs such as AraBERT [34] and CAMeLBERT [35] has further accelerated progress by enabling transfer learning across a wide range of Arabic NLP tasks.

However, the specific challenge of IER in Arabic text has received remarkably little attention. Most existing Arabic emotion datasets have been constructed with explicit emotional expressions as their primary focus, and the annotation schemes employed have not been designed to capture the underlying dimensions of emotional communication. This issue represents a critical gap, as a growing proportion of emotionally charged Arabic digital content, such as political commentary, personal narratives, and social media interaction, relies precisely on the kind of indirect, culturally mediated expression that current systems are ill-equipped to handle. Beyond the absence of dedicated annotated resources, the modeling strategies required to address this problem remain underdeveloped, and the generalization capacity of current models across Arabic varieties is poorly understood. It is precisely this confluence of gaps – in resources, modeling approaches, and evaluation – that motivates and defines the research undertaken in this thesis.

1.2 RESEARCH QUESTIONS

The main research questions of this thesis can be summarized as follows:

RQ1 How can implicit emotional meaning be identified and annotated in Arabic text, given the language's morphological complexity?

RQ2 What computational approaches are most effective for capturing the latent emotional cues that characterize implicit expression?

RQ3 To what extent can state-of-the-art models, including Arabic PLMs and generative LLMs, generalize across Arabic varieties and to other languages?

RQ4 what are the fundamental limitations of current systems when confronted with the linguistic and contextual depth of implicit emotional communication?

1.3 RESEARCH OBJECTIVES

The overarching objective of this thesis is to advance the SOTA in IER with a particular focus on Arabic text and its varieties. To this end, the research pursues three specific objectives:

1. The first objective is to address the critical absence of dedicated resources for IER in Arabic by constructing a publicly available annotated dataset specifically designed to capture emotional meaning inferred from contextual cues rather than explicit emotional vocabulary.

2. The second objective is to evaluate the noval dataset using different computational approaches, thereby establishing reproducible baselines for the research community.

3. The third objective is to develop an effective modeling strategy for IER in Arabic that goes beyond the limitations of standard approaches, by designing a hybrid architecture capable of capturing both the contextual and sequential dimensions of latent emotional meaning in Arabic text.

4. The fourth objective is to assess the generalization capacity of latest LLMs for emotion recognition across languages and dialects — specifically Algerian Arabic and English — probing their ability to perform genuine semantic and pragmatic emotional reasoning rather than exploiting superficial lexical patterns.

1.4 MAIN CONTRIBUTIONS

In pursuit of these objectives, the main contributions of this thesis can be summarized as follows:

Contribution 1 We introduce the AIER dataset, a publicly available benchmark composed of real-world Arabic reviews in MSA, annotated with emotion roles including emotions, cues, and causes. The dataset is evaluated using four computational approaches — classical machine learning, deep learning, Arabic PLMs, and generative LLMs — establishing reproducible baselines for the research community.

Contribution 2 We propose a hybrid contextual–sequential architecture that integrates the deep bidirectional encoding capacity of AraBERT with the temporal modeling capabilities of a Bidirectional Gated Recurrent Units (BiGRU) network. Evaluated on a newly compiled Arabic Twitter corpus under both implicit and explicit recognition settings, the model consistently outperforms established deep learning baselines, demonstrating the value of combining transformer-based contextual representations with sequential modeling.

Contribution 3 We investigate the emotional reasoning capabilities of SOTA generative LLMs across two complementary settings: emotion recognition in the Algerian Arabic dialect, including a multi-label classification framework, and context-based IER in English. Together, these studies provide a rigorous cross-lingual assessment of LLM capabilities and limitations in affective understanding across linguistically diverse and emotionally complex textual contexts.

1.5 THESIS STRUCTURE

This thesis contains in total of 7 Chapters and is organized as follows:

Chapter 1 : Presents the general background and motivation behind this research, research questions and their corresponding research objectives, main contributions, as well as the structure of this thesis.

Chapter 2 : Presents the theoretical background of this work, covering emotion modeling and conceptualization, challenges of Arabic for emotion recognition, the distinction between explicit and implicit emotion expression, and the main modeling approaches for emotion recognition, concluding with key research gaps in IER for Arabic.

Chapter 3 : Provides the literature review, covering existing approaches to IER , the landscape of Arabic NLP, highlighting the lack of existing annotated datasets and tailored models for IER, and establishes the conceptual and empirical gaps in the literature that the thesis aims to address.

Chapter 4 : Details the first core contribution: the construction of the AIER dataset and the establishment of baseline results across multiple computational approaches.

Chapter 5 : Presents the second contribution: the proposed hybrid AraBERT–BiGRU architecture and its empirical evaluation on Arabic implicit and explicit emotion recognition settings.

Chapter 6 : Presents the third contribution, structured around two complementary studies examining LLMs performance on multi label emotion recognition in Algerian Arabic dialect and in English.

Chapter 7 : Provides a summary of the thesis, highlighting key findings and offering valuable perspectives on potential future work in IER , Arabic NLP, and multilingual affective computing.

2

THEORETICAL BACKGROUND

This chapter aims to supplement the forthcoming chapters with background on key research milestones and challenges in the field of emotion recognition in text and Arabic Natural Language Processing (NLP). The supplementary background provides context behind the methods, representations, and modeling choices that have motivated the work presented in this thesis.

This introductory background comprises four parts. The first part provides an overview of emotion and affect as linguistic and computational phenomena and reviews core concepts related to emotion representation and emotion recognition in text. The second part introduces the linguistic characteristics of Arabic and discusses the challenges relevant to emotion recognition in Arabic text. The third part discusses explicit and implicit emotion expression, outlining their properties, followed by the main linguistic cues used to convey implicit emotions. The fourth part presents modeling approaches and word representation techniques used for emotion recognition in NLP, ranging from classical representations and traditional Machine Learning (ML) methods to Deep Learning (DL), transformer-based models, and Large Language Models (LLMs). Finally, we highlight the major gaps and challenges in implicit emotion recognition, including contextual ambiguity, subjectivity, data scarcity, and the limitations of existing approaches, particularly for Arabic.

2.1 EMOTIONS IN COMPUTATIONAL LINGUISTICS

2.1.1 *Theories of Emotion*

Emotions are a complex phenomenon that plays a central role in our experiences and daily communications. In recent decades, the study of emotions has attracted increasing attention across multiple disciplines, including philosophy, psychology, sociology, computer science, and linguistics [18].

Despite the diversity of theoretical perspectives, there is broad consensus that emotions constitute organized, multi-component processes that arise in response to the evaluation of an internal or external stimulus event relevant to an individual's goals, needs, and values [36, 37]. According to Plutchik [38], emotions evolved as survival mechanisms that help organisms respond to environmental challenges. Complementing this view, Scherer [36] describes emotion as a synchronized process involving multiple organismic subsystems, including information processing, support, executive, action, and monitoring systems. In contrast, appraisal-based theories, particularly those proposed by Lazarus [39], emphasize the role of cognitive interpretation in shaping emotional responses, suggesting that emotions arise from an individual's evaluation of events in relation to personal goals and expectations. Ekman [6] argues that emotions are not continuous or fluid phenomena, but rather a structured set of distinct affective categories that are biologically grounded, universal, and shared across cultures. Together, these perspectives frame emotion as a dynamic, multi-component process integrating cognitive, physiological, and behavioral dimensions.

In NLP, emotion is typically understood as a subjective affective state that can be inferred from various linguistic, semantic, and contextual cues present within textual data [40]. Emotion recognition, also known as 'emotion detection' or 'emotion classification', lies at the intersection of many fields, including NLP, computational linguistics, and data mining. It refers to a specific type of text categorization that focuses on classifying words, phrases, or documents into predefined emotion categories [41]. In contrast to sentiment analysis, which typically focuses on determining the polarity of a text (positive, negative, neutral), emotion recognition seeks to identify more fine-grained affective categories, such as joy, anger, sadness, fear, surprise, and disgust [6].

2.1.2 *Emotion modeling*

Several conceptual models proposed by psychologists and neuro-scientists have emerged to categorize and interpret the complex nature of human emotion. Among these, three prominent frameworks have gained significant attention in both psychological research and Natural Language Processing: the discrete or categorical model, the dimensional model, and the componential model. Each is described in the following.

2.1.2.1 *Discrete or categorical model*

The discrete model conceptualizes emotions as a set of distinct, independent, and universally recognizable affective categories, wherein each emotion represents a qualitatively separate psychological state with its own characteristic features and expressive patterns. Two of the most prominent frameworks of this approach include Ekman's model, which identifies six basic emotions (anger, fear, sadness, joy, disgust, and surprise) [42], and Plutchik's psychoevolutionary model, which extends the categorical framework to encompass eight primary emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust) arranged in opposing pairs of varying intensity that can combine to form complex secondary emotional states [43].

2.1.2.2 *Dimensional model*

In contrast to the discrete model, dimensional models represent emotions not as independent categories but as interconnected affective states that can be represented as points within a continuous multidimensional space. This perspective acknowledges that emotions vary in intensity and share underlying psychological properties, and that the boundaries between them are fluid rather than fixed. The most widely adopted dimensional framework is the Valence-Arousal-Dominance (VAD) model proposed by Russell and Mehrabian [44], in which emotion names are located within a vector space defined by three continuous dimensions: valence (the degree of pleasantness or unpleasantness), arousal (the level of physiological and psychological activation), and dominance (the degree of perceived control over the emotional situation).

2.1.2.3 *Componential model*

The componential model comprises appraisal-based theories [45, 46], which emphasize that emotions are influenced by an individual's subjective evaluation or appraisal of a situation or experience. According to this perspective, emotional responses are shaped by several appraisal factors, including pleasantness, certainty, attention, control, and responsibility, all of which influence how individuals perceive and react emotionally to specific situations [47].

2.1.3 *Emotion Expression in Text*

The expression of emotion in written text is through the use of words, and most often emotion-bearing words like 'happy' or 'sad'. However, emotions can be conveyed indirectly through narrative, metaphor, irony, or situational description, without any overt emotional markers [18, 19, 22, 48]. For example, while the sentence "I'm so happy to meet you again" explicitly states the emotion, the sentence "It's like the sun is shining just for me" implies happiness indirectly through contextual and figurative cues. This distinction is particularly relevant to the present work, as it motivates the focus on IER in Arabic text, where emotional meaning is frequently encoded through culturally specific and linguistically indirect expressions. A more detailed discussion of explicit and implicit emotion expression is presented in Section 2.3.

2.2 ARABIC LANGUAGE CHARACTERISTICS

As a Semitic language and one of the most widely spoken in the world, Arabic has over 420 million speakers across more than 22 countries and serves as the liturgical language of approximately 1.8 billion Muslims worldwide [26]. However, Arabic presents a unique set of linguistic characteristics and complexities that make it particularly challenging for NLP research fields [27, 28].

2.2.1 *Arabic varieties*

Arabic is characterized by a profound and pervasive linguistic diversity, encompassing multiple co-existing varieties that differ from one another at every level of linguistic structure: phonology, morphology, syntax, and lexicon, and that are associated with distinct communicative domains, social functions, and computational challenges. Arabic is represented in three distinct forms: Classical Arabic, Modern Standard Arabic (MSA), and Dialectal Arabic (DA).

Classical Arabic also known as Quranic Arabic or Fus'ha, is the variety of Arabic in which the Quran was revealed and in which the classical literary, religious, and scholarly tradition of the Arab world was produced over more than fourteen centuries. Classical Arabic is distinguished by its complex grammatical and syntactical rules and serves as the foundation for MSA.

Modern Standard Arabic (MSA) It is the standardized, pan-regional variety of Arabic used in formal written communication, education, journalism, literature, official discourse, and broadcast media across the Arab world. MSA is the variety for which the overwhelming majority of Arabic NLP resources have been developed, including sentiment lexicons and pre-trained language models, making it the best-supported Arabic variety for computational processing. However, MSA is primarily a written language and is not commonly spoken in everyday conversations.

Dialectal Arabic It refers to the diverse set of regional colloquial varieties spoken as native languages across the Arab world, used in everyday conversation, informal written communication, and increasingly in digital media and social platforms. Dialectal varieties differ from MSA and from each other at every level of linguistic structure (vocabulary, pronunciation, and grammar) and can be broadly grouped into five major regional families: Egyptian, Levantine, Gulf, Maghrebi, and Iraqi Arabic.

2.2.2 *Key Challenges in Arabic Emotion Recognition*

Arabic presents unique and multifaceted challenges in the tasks of emotion recognition that make it different from high-resource languages such as English. These challenges

arise from the interplay of the language's exceptional morphological complexity, its strong dialectal variation, its script-specific properties, and the culturally embedded nature of its communicative conventions. Together, these factors make all stages of the NLP pipeline more difficult. The main challenges are described below:

Script and Alphabet unlike English and many other languages that are written from left to right, Arabic is written and read from right to left. In terms of Alphabet, Arabic uses a special script of 28 letters. The script is cursive; that is most letters are connected within words. Also, Arabic does not have capital and lowercase letters; all letters are in a similar form regardless of their position in the sentence. The right-to-left directionality and position-dependent letter forms complicate tokenization and the alignment of Arabic text with NLP tools designed primarily for left-to-right Latin-script languages [28].

Morphological richness Arabic is a highly inflectional and derivational language, where a single root can generate numerous word forms through the application of inflectional patterns, inflectional morphology, and attached pronouns. This root-and-pattern system allows for a wide range of lexical and grammatical variations, contributing to the language's morphological richness and complexity. This morphological richness produces severe data sparsity, as the same root surfaces in dozens of distinct morphological forms that systems may fail to recognize without prior morphological normalization and lemmatization.

Orthographic Ambiguity The Arabic script employs optional diacritical marks, known as *harakat*, to indicate short vowels and other phonological features. These diacritics are essential for disambiguating words that share the same consonantal root but differ in pronunciation and meaning. For emotion recognition, orthographic ambiguity is particularly problematic because the same unvocalized sequence may correspond to words with entirely different affective meaning, and incorrect disambiguation can produce systematically wrong emotion predictions.

Orthographic inconsistency Both Standard and dialectal Arabic exhibit significant spelling variability, particularly in informal contexts such as social media. This inconsistency is often due to the lack of standardized orthographic conventions for

dialects, as well as the frequent omission of diacritics and non-standardized transliterations. This orthographic inconsistency significantly reduces the coverage of emotion lexicons and the generalization ability of models trained on standardized text, as the same emotion word may appear in dozens of orthographic variants that are treated as distinct and unrelated tokens.

Dialects and Variants Emotion recognition in Arabic dialects presents significant challenges due to both linguistic variability and resource scarcity. A major limitation lies in the scarcity of annotated emotional datasets, which restricts the effectiveness and generalization capabilities of machine learning approaches [49]. Furthermore, Arabic dialects exhibit considerable variation across regions at multiple linguistic levels, including lexicon, morphology, and syntax, often combined with informal writing and inconsistent spelling conventions, which make text processing and emotion detection particularly difficult [28, 50]. In addition to these challenges, cultural and social diversity across dialect-speaking communities play a crucial role in shaping emotional expression, which is often implicit and context-dependent, reflecting local norms and communication styles [51].

User-Generated Content Emotional expression in real-world communication rarely follows standardized linguistic norms, particularly on social media platforms. While the platform enables users from diverse backgrounds to share opinions and experiences, recognizing emotions in tweets remains highly challenging due to their characteristics and the complexity of emotional expression [52]. Tweets are short, informal, and often contain noisy linguistic elements such as dialectal variations, abbreviations, slang, misspellings, emojis, and hashtags, which complicate semantic understanding [53]. Moreover, emotions are not always expressed explicitly through emotion words; instead, they are often conveyed implicitly through figurative language such as metaphors, idiomatic expressions, similes, and contextual cues that require deeper interpretation [20]. Compared to MSA, tweets introduce additional layers of lexical, morphological, and structural complexity, highlighting the need for context-aware and semantically rich models capable of effectively capturing emotional cues.

Scarcity of Arabic Emotion Resources Arabic NLP remains significantly under-resourced relative to English, and this resource gap is particularly acute in the task of emotion recognition. Existing Arabic emotion corpora are limited in scale, predominantly focused on explicit emotional expression, and largely annotated using emotion schema borrowed from English benchmarks without adaptation to Arabic cultural and linguistic specificity.

2.3 EXPLICIT VS. IMPLICIT EMOTION EXPRESSION

A fundamental distinction in emotion recognition is that between explicit and implicit emotional expression. Understanding this distinction is crucial for framing the computational challenge addressed in this thesis, as it highlights the difference between conventional emotion recognition tasks and the more complex problem of inferring emotions from indirect linguistic, semantic, and contextual cues rather than from explicitly stated emotional terms.

2.3.1 *Emotion Words*

The most direct form of expressing emotions in language is through the use of emotion words, namely lexical items that convey or evoke emotional meaning. Emotion lexicons can generally be divided into three categories: emotion words, emotion-laden words, and emotion-related words [54, 55].

Emotion words directly express a particular emotional state, e.g., "سعيد" (*happy*), "غاضب" (*angry*), and "خائف" (*afraid*). These words explicitly communicate emotions and are, therefore, commonly used in traditional emotion recognition systems and affective

lexicons. Emotion-laden words, in contrast, do not explicitly name emotions but evoke emotional reactions through their semantic associations. Examples include "جنازة" (*funeral*), خيانة (*betrayal*), and "نجاح" (*success*). The broader category of emotion-related words includes lexical items associated with emotional situations, behaviors, or experiences without necessarily expressing or evoking a specific emotion directly. For example, "يبكي" (*crying*), "يبتسم" (*smiling*), and "يصرخ" (*screaming*). Unlike emotion words, emotion-related words are not uniquely associated with a single emotion. For example, the action "يبكي" (*crying*) is commonly linked to sadness; however, it may also occur in contexts of happiness or fear. A person may cry tears of joy after achieving success, or cry out of fear or anxiety. Therefore, most emotion-related words do not correspond exclusively to a particular emotion but rather exhibit contextual associations with multiple emotional states.

This distinction highlights that emotional meaning is distributed across language with varying degrees of explicitness and contextual dependency. While explicit emotion recognition often relies heavily on emotion-bearing words, IER depends more on emotion-laden and emotion-related expressions, contextual reasoning, and background knowledge to infer underlying emotional states. An emotion-bearing word is described as a word that can convey emotions on its own [22]. In this work, the term “emotion-bearing words” is interchangeable with the term “emotion word”.

2.3.2 *Explicit Emotion Expression*

Explicit emotions are often expressed through clear and direct emotional cues at the lexical, semantic, or symbolic levels. One of the most common forms of explicit emotion expression is the use of emotion-bearing words such as "حزين" (*sad*), غاضب" (*angry*), and "خائف" (*afraid*), which explicitly encode emotional meaning within the text.

In social media and informal online communication, users often reinforce explicit emotions using repeated letters and punctuation marks. For instance, elongated forms such as "سعيييييييد" and repeated punctuation marks like "!!!!!!" are commonly employed to emphasize emotional intensity. Explicit emotions may also be conveyed through interjections and exclamatory expressions such as "يا سلام!" or "الحمد لله!", which directly communicate affective reactions and emotional engagement.

Beyond textual expressions, emojis and emoticons constitute important indicators of explicit emotions in digital communication. Emojis directly represent emotional states and often complement or even replace textual emotional expressions. Similarly, emoticons such as :) and :(provide simplified visual representations of affective states. Previous studies have demonstrated that emojis carry valuable affective information and play a significant role in emotion recognition within social media texts [56, 57].

The presence of such explicit markers renders emotion recognition relatively tractable from a computational perspective, as models can leverage lexicon-based methods, keyword spotting, and shallow syntactic analysis to identify and classify the emotional content of a text with reasonable accuracy [16, 41].

2.3.3 *Implicit Emotion Expression*

Implicit emotion refers to the expression of an emotional state without the use of explicit emotional cues. In this form, emotions must be inferred from contextual descriptions, events, behaviors, figurative language, or culturally specific expressions rather than from direct emotion-bearing words. In Arabic, implicit emotion is particularly prevalent in literary prose, poetry, social media discourse, and everyday conversations, where emotional meaning is frequently embedded in culturally specific metaphors, rhetorical devices, and indirect speech acts [9]. For instance, the expression "الدنيا أغلقت أبوابها في وجهي" (*The world closed its doors in my face*) implicitly conveys despair or disappointment without directly

naming the emotion. Similarly, sarcasm and irony frequently encode implicit emotions that may differ from the literal meaning of the text [58–60].

At the semantic level, words that implicitly express emotions fall into two categories: emotion-related words and emotion-laden words. Unlike explicit emotion-bearing words, these lexical items do not directly name emotional states, making emotional interpretation more dependent on contextual and semantic understanding.

Emotion-related words, as defined by Pavlenko [55], describe behaviors, actions, or situations associated with particular emotions without explicitly naming the emotions themselves. For example, the verb "ابتسم" (*to smile*) may indicate joy, relief, or even sarcastic satisfaction, depending on the surrounding context. Similarly, the expression "احمرَ وجهُه" (*his face turned red*) may signal embarrassment or anger without explicitly mentioning any emotional state. Emotion-laden words, on the other hand, evoke affective responses through semantic, cultural, and experiential associations. For instance, the term "العيد" (*Eid/the festival*) often evokes feelings of joy within Arabic communities despite lacking an explicit emotion label. Similarly, words such as "جنازة" (*funeral*), "احتلال" (occupation), and "مرض" (*illness*) can evoke sadness, anger, or fear through lived experiences rather than direct emotional meaning.

These distinctions highlight one of the central challenges of IER: emotional meaning is often distributed across actions, events, semantic associations, and contextual knowledge rather than encoded in isolated emotion words. Consequently, accurate recognition of implicit emotions requires models capable of deep contextual understanding and semantic reasoning beyond simple lexical matching.

2.3.4 *Linguistic Cues of Implicit Emotion*

Implicit emotional expression is realized through a diverse range of linguistic indicators, including:

Metaphor and Figurative Language constitute one of the most widespread means of expressing implicit emotion in various languages [20, 21]. Conceptual metaphors, as theorized by Lakoff and Johnson [61], map emotional experiences onto concrete domains, allowing speakers to express emotion indirectly. In Arabic, figurative expression is deeply embedded in the linguistic and cultural tradition, with classical and colloquial Arabic both exhibiting rich metaphorical repertoires for expressing emotional states. For example, the expression "قلبي يحترق على أمي" (My heart is burning for my mother) implicitly conveys intense grief or longing without explicitly naming the emotion.

Irony and Sarcasm represent another major source of implicit emotion, in which the surface meaning of an utterance is systematically inverted or distorted to convey a contrary affective stance [58–60]. The semantic mechanism underlying irony is one of deliberate semantic incongruence, when the speaker exploits the gap between the literal meaning of their words and the contextually intended meaning [62]. Expressions such as "ما أروع هذا اليوم" (What a wonderful day), uttered in a context of misfortune, communicate negative emotion through positive surface lexis.

Event and Situation Description the most semantically complex form of implicit emotion expression is the description of events, actions, or situations whose emotional significance is recoverable not from emotional vocabulary but from the reader's commonsense knowledge and appraisal of the described situation [16, 18]. A sentence such as "اليوم ربحت السيارة" (Today I won the car) contains no emotion word, yet its emotional content is readily inferable through commonsense reasoning about the

affective consequences of the described event, whereby the situation itself produces the emotion as its natural outcome. Conversely, a sentence such as "أغلق الباب بعنف وخرج" (He slammed the door and left) contains no emotion word either, yet its emotional content is readily inferable through commonsense reasoning about the affective causes of the described behavior, whereby the action itself betrays a prior emotional state that must have driven it.

Rhetorical Devices are frequently used to convey implicit emotions, particularly in literary texts, user-generated reviews, and social media discourse [63, 64]. For example, a rhetorical question such as "كيف يفعلون هذا؟" (How could they do this?) may implicitly express anger, disbelief, or surprise without explicitly naming the underlying emotion.

2.4 COMPUTATIONAL APPROACHES IN NLP

2.4.1 *Text Representations*

Text representations constitute a fundamental aspect to NLP systems. As computational models cannot process raw textual data, input text must be transformed into numerical representations that capture syntactic , semantic, and contextual information in a machine-interpretable form. Over the years, text representation methods have evolved from sparse statistical representations to dense distributed embeddings and, more recently, to contextualized representations generated by transformer-based architectures. These representations play a critical role in emotion recognition tasks, as the quality of the encoded textual features directly influences the model's ability to capture affective meaning.

2.4.1.1 *Classical Representations*

Classical representation methods treat documents as collections of discrete linguistic units and represent them as high-dimensional sparse vectors based on the presence, frequency, or statistical significance of these units. These techniques include:

BAG OF WORDS (BOW) One of the earliest and most widely used text representation techniques in NLP. It represents a document as an unordered collection of words, where each dimension corresponds to a vocabulary term and its associated frequency within the text. Word order, grammatical structure, and syntactic relationships are ignored, and only the occurrence or frequency of words is considered.

The simplicity and computational efficiency of BoW made it highly popular in early emotion classification systems. However, the model suffers from several important limitations, including high dimensionality, sparsity, and the inability to capture semantic relationships or contextual dependencies between words.

TERM FREQUENCY–INVERSE DOCUMENT FREQUENCY (TF-IDF) It is a weighting scheme used to reflect the importance of a word in a document compared to a collection of documents. The main idea behind TF-IDF is to assess the importance of a word in a document by considering how often it appears in the document (Term Frequency, TF) and how rare it is across the entire collection of documents (Inverse Document Frequency, IDF). Mathematically, the TF-IDF weight of a term increases proportionally with its frequency in a document but decreases if the term appears in many documents across the corpus. Consequently, emotionally informative words receive higher weights than common functional words.

N-GRAMS To incorporate limited contextual information, classical NLP systems frequently employ n-gram representations, which extend the BoW model by representing documents using contiguous sequences of n tokens rather than isolated words alone. In this representation, unigrams correspond to individual words, bigrams represent sequences of two consecutive tokens, and trigrams capture sequences of three consecutive tokens.

Unlike the BoW model, *n*-gram representations partially preserve local word order and short contextual dependencies. For example, bigrams such as "لا أشعر" (I do not feel) or "يموت من" (dying from) retain local co-occurrence patterns associated with emotional expressions, while trigrams such as "لا أشعر بالسعادة" (I do not feel happy) capture negation-emotion constructions that are often misinterpreted by unigram-based models. Consequently, *n*-gram features are particularly useful for modeling negation, intensification, and short contextual patterns in sentiment and emotion analysis. In this thesis, TF-IDF, BoW, and *n*-grams representations are used in Chapter 4.

Classical text representations offer several advantages, including simplicity, interpretability, and low computational cost. They are particularly effective when training data are limited and often serve as strong baselines for text classification tasks.

However, these approaches exhibit substantial limitations for emotion recognition, especially for implicit emotion detection. Because they rely primarily on surface lexical patterns, they struggle to capture semantic similarity, contextual meaning, and long-range dependencies. In addition, sparse representations suffer from high dimensionality and limited generalization capability. To address these limitations, more advanced techniques have been developed, such as distributed word embedding approaches, which are discussed in the following section.

2.4.1.2 *Static Word Embeddings*

Word embedding methods address the fundamental limitations of classical sparse representations by mapping words to dense, low-dimensional continuous vector representations of words that capture semantic and syntactic relationships. Unlike sparse representations, embeddings encode semantic relationships learned from large unlabeled corpora, enabling generalization across lexically varied expressions and providing a compact, information-rich representation of word meaning that is far better suited to downstream emotion recognition tasks. In this space, words with similar meanings are located close to each other, enabling models to generalize better and capture nuanced relationships between words. For example, in a well-trained embedding space, the vectors for emotion-related words

such as 'happy' and 'joyful' or 'sad' and 'depressed' are positioned close to one another, reflecting their semantic and affective similarity.

The most widely used word embedding methods include Word2Vec and FastText. Each of these methods learns embeddings by analyzing large corpora of text, but they differ in their approaches and the types of relationships they capture.

WORD2VEC Word2Vec, introduced by Mikolov et al. [65], is one of the most influential embedding models in NLP. It learns vector representations by predicting contextual relationships between words using shallow neural networks on a large corpus of text. The model is commonly implemented using two architectures:

- **Continuous Bag-of-Words (CBoW)**: Predicts a target word based on its surrounding context words.For example, given the context words "أنا لا أستطيع التوقف عن" and "منذ وفاته" (I cannot stop ... since his death), CBoW predicts the target emotion-bearing word "البكاء" (crying), learning in the process that this contextual pattern is strongly associated with the semantic field of grief and sadness.

- **Skip-gram**: Predicts context words given a target word. For example, given the target emotion word "اكتئاب" (depression), Skip-gram predicts its surrounding context words such as يعاني (suffers), وحيد (lonely), الحزن (sadness), and البكاء (crying), thereby learning that these words frequently co-occur with depression and encoding their semantic and affective proximity in the embedding space.

Despite its superior performance on previous word representation techniques and computational efficiency, Word2Vec generates a single static representation for each word regardless of context. Consequently, polysemous words receive identical embeddings in different semantic situations, meaning that polysemous words and context-dependent emotional expressions receive the same representation in all contexts. The word "بارد" (cold), for instance, receives a single embedding that cannot simultaneously capture its literal me-

teorological meaning and its colloquial usage to describe an emotionally detached or indifferent person. This context-insensitivity is a critical limitation for IER, where the affective meaning of a word or expression is fundamentally dependent on its surrounding context. Additionally, Word2Vec cannot generate embeddings for words not seen during training, limiting its ability to handle rare or new words (the problem of Out-of-vocabulary).

FASTTEXT FastText, proposed by Bojanowski et al. [66], extends Word2Vec by representing each word not as a single unit but as a bag of character-level n-grams, from which the word embedding is composed as the sum of its constituent n-gram embeddings. This sub-word representation strategy confers two significant advantages over Word2Vec. First, it provides robustness to morphological variation, as morphologically related word forms (which share character-level n-gram subsequences) receive similar embeddings even when they are distinct surface forms. Second, it enables the construction of embeddings for out-of-vocabulary words by composing their representations from the embeddings of their constituent character n-grams.

Pre-trained FastText embeddings for Arabic have been made available by the Facebook AI Research team (Grave et al. [67]), trained on Common Crawl and Wikipedia data for 157 languages, including Arabic. For emotion recognition specifically, FastText's robustness to morphological variation and out-of-vocabulary words makes it particularly well-suited for social media and dialectal data, where non-standard orthography and morphological creativity are pervasive features of the text.

However, similar to Word2Vec, FastText produces only a single vector representation per word, which does not account for its varying meanings across different contexts. Additionally, the use of subword units increases the complexity of the model, as it requires storing and processing embeddings for a large number of n-grams. FastText embeddings are used in this thesis in Chapter 4 and 5.

2.4.1.3 *Contextual embeddings*

Contextual embeddings were introduced to overcome the limitations of static word embeddings, particularly their inability to capture the context-dependent nature of linguistic and emotional meaning. Unlike Word2Vec and FastText, which assign a single fixed vector

to each word regardless of context, contextual embedding models generate dynamic representations based on the surrounding context of each token. Consequently, the same word may receive different vector representations depending on its semantic usage within a sentence. These contextualized representations enable models to capture nuanced semantic, syntactic, and affective information, making them especially effective for tasks such as IER, where emotional meaning often depends heavily on contextual interpretation.

TRANSFORMERS The foundational technology underlying modern contextual embedding models is the Transformer architecture, introduced by Vaswani et al. [68] in the seminal paper *Attention Is All You Need*. Unlike earlier sequence modeling approaches based on recurrent or convolutional neural networks, the transformer entirely eliminates recurrence and convolutions in favor of a purely attention-based architecture. Recurrent models process text sequentially, handling one token at a time, which limits parallelization and makes capturing long-range dependencies more difficult. In contrast, transformers process all tokens simultaneously through a self-attention mechanism that models relationships between every token in the sequence in parallel. This architectural innovation enables the construction of rich contextual representations capable of capturing complex semantic, syntactic, and long-range dependencies across the entire text.

The transformer architecture consists of two principal components: an **encoder** and a **decoder**, each composed of a stack of identical layers incorporating two core mechanisms: multi-head self-attention and position-wise feed-forward networks. For the purposes of contextual word representation, the encoder component is of primary relevance, as it is responsible for transforming the input token sequence into a sequence of rich contextualized representations. In contrast, the decoder generates the output sequence by attending to both the encoder's output and its own previous predictions. An overview of the Transformer architecture is illustrated in Figure 2.1.

The self-attention mechanism is the defining innovation of the transformer architecture and the primary source of its contextual representation capability. Unlike recurrent architectures, which process tokens sequentially and propagate contextual information through a hidden state that is subject to vanishing gradient problems over long sequences, self-attention computes the representation of each token as a weighted sum of all other

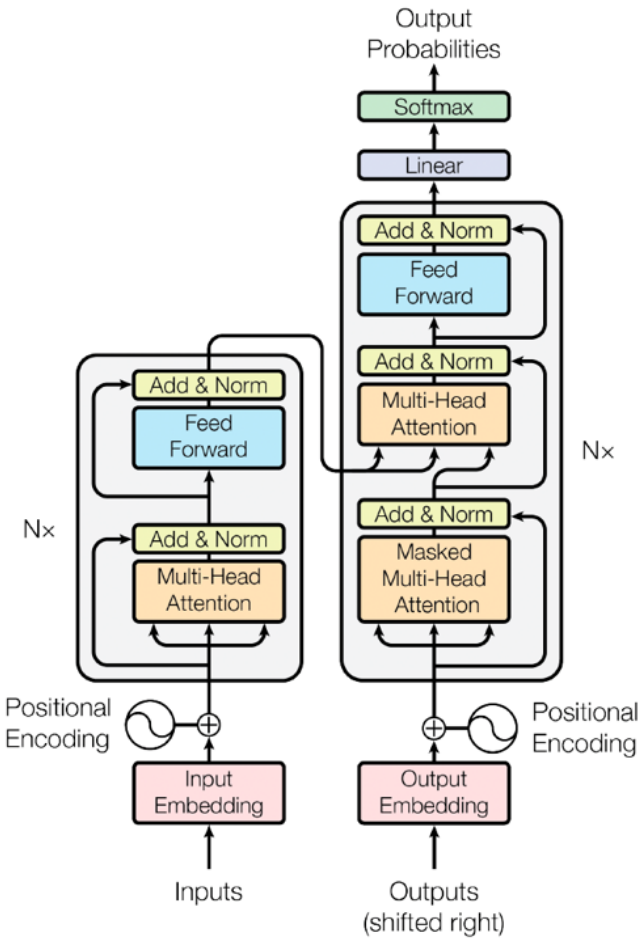


Figure 2.1: The Transformer architecture, consisting of an encoder and a decoder (Vaswani et al. [68].

tokens in the sequence, with weights determined by the degree of semantic and syntactic relevance between token pairs. This mechanism enables every token to directly attend to every other token in the sequence regardless of their positional distance, capturing long-range contextual dependencies that recurrent models struggle to model effectively.

Contextual embeddings generated by transformer-based models have substantially advanced NLP research, particularly in tasks requiring deep semantic understanding, such as sentiment analysis, sarcasm detection, and emotion recognition. Consequently, contextual embeddings have become the foundation of modern retrained language models and large language models, such as Bidirectional Encoder Representations from Transformers (BERT) [10] (Section 2.4.4.1) and Generative Pre-trained Transformer (GPT) [69] (Section 2.4.4.2).

2.4.2 *Classical Machine Learning Methods*

Machine Learning (ML), a major branch of artificial intelligence, focuses on developing computational models capable of learning patterns from data and making predictions or decisions without being explicitly programmed for specific tasks. In Natural Language Processing (NLP), ML techniques have been extensively applied to various text analysis tasks, including sentiment analysis and emotion recognition. Traditional ML approaches rely on handcrafted feature extraction and statistical text representations to transform textual data into numerical vectors that can be processed by classification algorithms.

Early methods in text emotion recognition relied heavily on traditional ML classifiers. In these approaches, textual data is first converted into numerical features using representations such as BoW and TF-IDF. This section explores the most widely used classical ML algorithms for recognizing emotions in text, namely Support Vector Machine (SVM), K-Nearest Neighbors (KNN), eXtreme Gradient Boosting (XGBoost), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), and Multinomial Naïve Bayes (MNB). ML algorithms are used in this thesis in Chapter 4.

2.4.2.1 *Support Vector Machine (SVM)*

SVM [70] is one of the most widely adopted classifiers for text emotion recognition, as it aims to identify the optimal hyperplane that maximally separates data points of different classes in a high-dimensional feature space. Its effectiveness stems from its ability to handle high-dimensional sparse text representations, such as TF-IDF vectors, and its use of kernel functions to capture non-linear decision boundaries between emotion categories.

2.4.2.2 *K-Nearest Neighbors (KNN)*

KNN [71] is a non-parametric instance-based classifier that assigns a label to a new text by identifying its k most similar instances in the training set and adopting the majority label among those neighbors, with similarity typically measured using cosine distance over TF-IDF feature vectors. While intuitive and requiring no explicit training phase, KNN

suffers from the curse of dimensionality in high-dimensional text feature spaces and scales poorly to large corpora.

2.4.2.3 *Decision Tree (DT)*

DT [72] classifies text instances by recursively partitioning the feature space according to the feature that maximizes information gain or minimizes Gini impurity at each node, producing a hierarchical tree structure in which each leaf corresponds to an emotion label. Decision trees offer high interpretability, as the classification path from root to leaf provides a human-readable explanation of the emotion prediction in terms of specific linguistic features. However, they are prone to overfitting on high-dimensional text feature spaces.

2.4.2.4 *Random Forest (RF)*

RF [73] is an ensemble method that constructs a large number of decision trees on bootstrap samples of the training data, each trained on a random subset of features, and assigns the label predicted by the majority of trees. This double randomization produces a diverse and robust ensemble that is substantially more accurate than any individual tree and provides useful feature importance estimates that identify the most discriminative vocabulary.

2.4.2.5 *EXtreme Gradient Boosting (XGBoost)*

XGBoost [74] is a highly efficient ensemble method that constructs a strong classifier by iteratively training shallow decision trees, where each successive tree corrects the residual errors of the preceding ensemble. Its built-in regularization, robustness to irrelevant features, and ability to capture non-linear feature interactions make it well-suited to the heterogeneous feature spaces typical of emotion recognition systems.

2.4.2.6 *Logistic Regression (LR)*

LR is a supervised classifier used to predict the probability of a categorical outcome based on a set of input features. Unlike linear regression, which is designed for continuous numerical outcomes, logistic regression is specifically suited for classification tasks where the dependent variable is categorical. Logistic regression exists in two main forms: binary

logistic regression, which is used when the target variable contains two classes, and multinomial logistic regression, which is employed when the target variable consists of more than two classes, such as emotion recognition problems involving emotions like joy, sadness, anger, fear, surprise, and disgust.

2.4.2.7 *Naïve Bayes (NP)*

Naïve Bayes (NB) is a probabilistic supervised classifier based on Bayes' theorem and assumes that features are conditionally independent given the target class. In the multinomial variant, the classifier models the frequency distribution of words within documents, making it particularly suitable for discrete textual features. Despite its simplicity, MNB demonstrates surprisingly competitive performance on emotion classification tasks and is particularly valued for its computational efficiency, fast training, and strong performance with small amounts of annotated data.

2.4.3 *Traditional Deep Learning Approaches*

DL is a subfield of ML that is primarily based on neural networks [75]. These neural networks are inspired by the structure and function of the human brain, consisting of interconnected nodes that process and transmit information. Focusing on training multilayer artificial neural networks, DL can automatically learn hierarchical representations from raw data, thereby reducing the need for manual feature engineering [76]. This capability enables DL models to capture complex, non-linear patterns and long-range dependencies in text, making them particularly suitable for emotion classification tasks. Traditional DL approaches are used in this thesis in Chapter 4 and 5.

For DL, many other types of network architectures are developed, and we will briefly describe some Traditional ones here.

2.4.3.1 *Convolutional Neural Networks (CNN)*

Originally developed for image recognition by Lecun et al. [77], Convolutional Neural Network (CNN) was successfully adapted for text classification by [78], who demonstrated that convolutional filters applied over word embedding sequences can effectively capture

local features and semantic patterns useful for text classification tasks such as emotion recognition. CNN applies multiple filters of varying window sizes over the input word embedding matrix, extracting local emotional features such as emotion-bearing phrases and affective collocations, followed by max-pooling to retain the most salient features across the sequence. CNN has been applied to Arabic emotion recognition with strong results, demonstrating its ability to capture morphologically rich emotional expressions through character- and word-level convolutional feature extraction [79].

2.4.3.2 *Recurrent Neural Networks (RNNs)*

Recurrent Neural Network (RNN) processes text sequentially, maintaining a hidden state that accumulates contextual information from preceding tokens and is updated at each time step, making it naturally suited to modeling the sequential and temporal structure of language. However, standard RNNs suffer from the vanishing gradient problem, which causes them to lose information from distant tokens and limits their ability to capture long-range emotional dependencies. To address these issues, advanced architectures such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) have been developed.

- **LSTM** [80]: is a specialized type of RNN designed to effectively capture long-term dependencies in sequential data. It was proposed as a solution to the vanishing gradient problem that plagues traditional RNNs. LSTM address this issue through a gated memory cell architecture that controls the flow of information over long sequences via three learned gates: the input gate, forget gate, and output gate.

- **BiLSTM** [81]: is an extension of the standard LSTM model that processes the input sequence in both forward and backward directions simultaneously, producing at each position a contextual representation that integrates information from both preceding and following context. This is a very valuable property for emotion recognition, where the emotional significance of a word may be determined by both its left and right context.

- **GRU** [82]: is a simplified variant of the LSTM that consolidates the input and forget gates into a single update gate and eliminates the separate memory cell, producing a more parameter-efficient architecture that achieves comparable performance to LSTM on most sequence modeling tasks with reduced computational cost. GRU has been

applied to emotion recognition tasks as a computationally lighter alternative to LSTM, demonstrating strong performance particularly in settings where training data is limited and model complexity must be controlled to prevent overfitting.

- **BiGRU** [83]: A Bidirectional GRU consists of two independent GRUs: the forward GRU, which captures contextual information from the sequence in a forward direction, and the backward GRU, which processes it in reverse. This bidirectional architecture enables the model to leverage information from both past and future states, improving its ability to better capture complex patterns and dependencies within the text.

2.4.3.3 *Attention Mechanism*

The concept of attention in machine learning is inspired by biological attention, where humans selectively focus on relevant information rather than processing all available data. The attention mechanism was first introduced by Bahdanau et al. [84] as an extension to recurrent architectures that enables the model to focus on the most relevant parts of the input sequence when producing a classification decision, rather than relying solely on the final hidden state of the recurrent network as a compressed representation of the entire input. Instead of treating all words equally, attention assigns different weights to them based on their relevance to the classification task, helping the model capture context and nuances more effectively and improving text classification accuracy.

For emotion recognition, the self-attention mechanism provides a powerful tool for capturing the contextual dependencies that determine affective meaning. Consider the Arabic sentence "ابتسم رغم كل ما مرّ به" (He smiled despite everything he had been through), the emotional significance of "ابتسم" (smiled) as an expression of resilience rather than genuine joy is entirely determined by its relationship to "رغم كل ما مرّ به" (despite everything he had been through), a long-range dependency that self-attention captures directly by allowing the representation of "ابتسم" to attend strongly to the contextualizing phrase, producing a contextual embedding that encodes the implicit emotional meaning of the full construction rather than the literal meaning of the smile word alone.

In this way, the attention mechanism mimics how we focus on the most relevant parts of a sentence. By assigning higher weights to key words, the model can prioritize the most important information, enabling it to accurately interpret the emotion.

2.4.4 *Transformers-Based Models*

Building on the transformer architecture introduced in Section 2.4.1.3, PLMs have revolutionized NLP by learning rich linguistic representations from massive unlabeled text corpora via self-supervised objectives, which can then be efficiently fine-tuned on downstream tasks with relatively small amounts of labeled data. This approach of pre-training and fine-tuning has established new SOTA results across a wide range of NLP tasks, including emotion recognition.

2.4.4.1 *Encoder-based Models*

Encoder-based models adopt the transformer encoder stack to produce rich bidirectional contextual representations of input text, making them particularly well-suited to classification tasks such as emotion recognition where the goal is to map an entire input sequence to a discrete label.

BERT: BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS

Introduced in 2018 by Devlin et al. [10], BERT showed that a single model could achieve SOTA results on eleven of the most common NLP tasks. It is the foundational encoder-based

PLM, pre-trained on English Wikipedia and BooksCorpus through two self-supervised objectives: Masked Language Modeling (MLM), which randomly masks 15% of input tokens and trains the model to predict them from their bidirectional context, enabling the learning of deep contextual representations that integrate both left and right context simultaneously; and Next Sentence Prediction (NSP), which trains the model to determine whether two sentences appear consecutively in the original text, encouraging the learning of inter-sentence relationships and discourse-level dependencies. Together, these two objectives enable BERT to develop rich, bidirectional contextual representations that capture complex semantic and syntactic relationships across the full input sequence — properties that are of critical importance for emotion recognition, where the affective meaning of a text is distributed across its full contextual structure rather than localized in individual lexical items.

BERT is trained on vast amounts of data in an unsupervised way, allowing the model to learn a rich representation of language that can be reused in a variety of downstream tasks. A notable feature of BERT is its ability to achieve strong results in specific domains with minimal architectural modifications. The model can be fine-tuned using small amounts of domain-specific labeled data, requiring only the addition of a single output layer for task-specific adaptation. Due to these capabilities, BERT-based models deliver SOTA performance across a wide range of NLP tasks, including sentiment analysis, hate speech detection, language inference, question answering, and more.

ROBERTA: ROBUSTLY OPTIMIZED BERT APPROACH Liu et al. [85] investigated ways of optimizing BERT's pretraining approach and found that they could further improve BERT. The new version, RoBERTa, obtained new SOTA results on multiple benchmarks and outperformed the original BERT. They pre-trained on a corpus ten times larger with fewer epochs on the data and used a batch size approximately eight times larger. They also removed the NSP training objective and found slightly improved performance. These modifications produce substantially stronger contextual representations than the original BERT, and RoBERTa has consistently outperformed BERT on downstream emotion classification benchmarks, establishing itself as one of the strongest encoder-based baselines in affective computing research.

ARABIC PRETRAINED LANGUAGE MODELS Arabic PLMs adapt transformer-based architectures to the linguistic characteristics of Arabic by leveraging large-scale Arabic corpora. These models are trained exclusively on Arabic texts, including texts in both MSA and Arabic dialects, and consistently produce stronger contextual representations for Arabic NLP tasks by capturing the morphological, syntactic, and cultural properties of Arabic more effectively. Such models have demonstrated strong performance in emotion recognition tasks. Arabic PLMs are used in this thesis in Chapter 4.

Notable examples include:

- **AraBERT** [34]: a BERT-based model specifically designed for the Arabic language to improve the SOTA in several Arabic Natural Language Understanding (NLU) tasks. The model was trained on about 70 million sentences, covering Arabic news from different Arab regions

- **CAMeLBERT** [35]: a collection of eight pre-trained BERT-based models designed for the Arabic language, encompassing both MSA and dialect varieties. The primary objective of these models is to explore the interplay of size, language variation, and fine-tuning task type.

- **MARBERT and ArBERT** [86]: MARBERT was pre-trained on 1 billion tweets, encompassing text in both MSA and Arabic dialects. It follows the same architecture as BERT but excludes the NSP task. A subsequent version, MARBERTv2, was introduced, which utilizes twice the number of tokens. In contrast, ArBERT was trained on 61GB of MSA text (approximately 6.5 billion tokens), sourced from Arabic books, news articles, and Arabic Wikipedia dumps.

- **QARiB** [87]: an Arabic BERT-based model developed by researchers from the Qatar Computing Research Institute for the Arabic language. The model was trained on a large corpus comprising 14 billion words, covering both MSA and Arabic dialects. Several versions of the model were introduced, including QARiB10, QARiB25, QARiB25_mix, QARiB60_mix, and QARiB25_mix_far.

- **AraELECTRA** [88]: is the Arabic ELECTRA-based model. It was pretrained on the same dataset as ARABERTv0.2, employing identical pre-processing techniques. ELECTRA's training process utilizes two transformers: a masked language generator

(G) and a BERT-based discriminator (D). The generator is trained to predict masked tokens, while the discriminator is trained to distinguish between original and replaced tokens generated by the generator.

Dialectal Arabic exhibits substantial lexical, morphological, and syntactic variation across regions, which often leads to performance degradation when using MSA-focused models. Arabic dialectal PLMs address this limitation by incorporating large-scale dialectal corpora during pretraining, enabling improved representation of colloquial expressions and region-specific linguistic phenomena. Arabic Dialectal PLMs are used in this thesis in Chapter 6.

Representative models include:

- **DziriBERT** [89]: a BERT-based model specifically pre-trained on large-scale Algerian dialectal text. It covers both Arabic script and Arabizi, capturing local linguistic features and cultural expressions, thus serving as a strong baseline for Algerian dialect NLP tasks.
- **AraBERT-Algerian** [90]: a variant of the AraBERT model, fine-tuned on Algerian dialect data. It leverages the robust architecture of AraBERT while adapting to local linguistic variations, making it a valuable model for emotion classification tasks in the Algerian context.

Multilingual PLMs are designed to support multiple languages within a single architecture, enabling cross-lingual transfer learning. While they provide broad linguistic coverage, their performance may lag behind Arabic models on fine-grained emotion recognition tasks. Prominent multilingual models include:

- **XLM-RoBERTa** [91]: is a large, multilingual, pre-trained language model designed to process text in multiple languages, based on the RoBERTa architecture. The model was trained on 2.5 TB of Wikipedia and CommonCrawl data spanning 100 languages, including Arabic.

BERT-based models are used in this thesis in Chapters 4, 5, and 6. In the scope of this thesis, we do not focus on designing novel neural network architectures but rather on

leveraging and adapting existing pretrained models as powerful contextual representation functions, exploiting their capacity to capture deep semantic and contextual dependencies in text.

2.4.4.2 *Decoder-based models*

Decoder-based models adopt the transformer decoder stack and are pre-trained through autoregressive language modeling, where they predict each token from its left context, producing powerful generative representations that encode extensive world and commonsense knowledge accumulated from massive text corpora.

GENERATIVE PRE-TRAINED TRANSFORMERS (GPT) Generative Pre-trained Transformers (GPT) are a family of transformer-based language models developed by OpenAI [69]. Unlike BERT, which relies on bidirectional encoding, GPT models employ a unidirectional transformer decoder architecture that predicts the next token in a sequence based on previously observed tokens. Through this autoregressive pre-training strategy, GPT models learn rich contextual representations that capture linguistic structure, semantic relationships, and broad world knowledge acquired from large-scale and diverse text corpora.

GPT models are trained using self-supervised learning on massive unlabeled datasets and can later be adapted to downstream NLP tasks through fine-tuning or prompting techniques. Owing to their strong generative capabilities and contextual understanding, GPT models have achieved remarkable performance across numerous tasks, including text generation, summarization, question answering, translation, sentiment analysis, and emotion recognition.

2.4.4.3 *Large Language Models and Generative AI*

LLMs present the recent evolution of transformer-based architectures, characterized by massive training on diverse, large-scale corpora, typically comprising billions of tokens and parameters with sizes significantly surpassing pretrained language models such as BERT and RoBERTa. LLMs exhibit strong generalization capabilities across a wide range of NLP tasks, often without task-specific fine-tuning, through prompt-based interaction [92].

From an architectural perspective, LLMs are predominantly built upon deep Transformer stacks with self-attention mechanisms, enabling them to model long-range dependencies and complex contextual relationships. Unlike earlier PLMs, which are optimized for discriminative downstream tasks via fine-tuning, LLMs are primarily trained using autoregressive or sequence-to-sequence objectives that support text generation, reasoning, and instruction following [68].

This capability opens new possibilities for emotion recognition in low-resource settings such as Arabic implicit emotion, where labeled training data is scarce and fine-tuning is constrained by annotation costs. LLMs are used in this thesis in Chapters 6

CHARACTERISTICS OF LLMS LLMs are distinguished from conventional PLMs by several defining characteristics that collectively represent a qualitative shift in the capabilities of language models:

- **Scale and Emergent Abilities:** LLMs contain hundreds of millions to trillions of parameters and are trained on web-scale corpora, enabling emergent capabilities such as complex reasoning and nuanced affective understanding that do not manifest in smaller models.
- **In-Context Learning:** Rather than requiring fine-tuning on labeled data, LLMs perform tasks through natural language prompting in zero-shot or few-shot settings, without updating any model parameters [92].
- **Generative Capability:** Unlike encoder-based PLMs, LLMs are fundamentally generative, producing producing coherent and context-aware text that can articulate not only emotion predictions but also the inferential reasoning behind them through chain-of-thought explanations [93].
- **Task Generality:** A single LLM can support diverse tasks, including classification, summarization, translation, question answering, and emotion analysis, through prompt reformulation.

CLOSED VS. OPEN-WEIGHT LLMS LLMs can be broadly categorized into closed (proprietary) and open-weight LLMs based on accessibility and usage constraints.

Closed LLMs are accessible only through controlled APIs, with proprietary training data and restricted model internals. While such models often demonstrate outstanding performance, their closed nature limits transparency, reproducibility, and fine-grained adaptation for research purposes. Prominent closed LLMs include:

- **ChatGPT-4**: an advanced LLM developed by OpenAI as part of the GPT series [94]. Based on the GPT-4 architecture, it was released in March 2023 with 175 billion parameters. ChatGPT-4 has demonstrated significant advancements in NLU, reasoning, and contextual awareness across a wide range of domains.

- **Gemini**: a multi-modal LLM released by Google DeepMind in December 2024, designed to handle text, images, and audio inputs [95]. It leverages extensive multilingual pre-training and advanced reasoning abilities, making it suitable for emotion detection tasks that require detecting subtle linguistic and cultural cues.

- **DeepSeek**: developed by DeepSeek AI [96], and trained with a focus on efficiency and reasoning accuracy. It combines large-scale multilingual data with fine-tuning strategies that enhance contextual consistency, allowing effective handling of dialectal variations in Arabic.

Open-weight LLMs provide full public access to model weights and architectures, enabling researchers to fine-tune, modify, and deploy models locally without API cost constraints or reproducibility concerns. Representative open-weight LLMs include:

- **Llama 3**: an open-source LLM from the Meta AI series launched in April 2024, is available in both 8-billion and 70-billion-parameter versions [97]. The model was trained on diverse publicly available data sources, including CommonCrawl, GitHub, and Wikipedia.

- **Gemma-3**: a family of open-weight LLMs developed by Google DeepMind and released in March 2024, available in four parameter sizes: 1B, 4B, 12B, and 27B, built on the same research foundations as the Gemini model family [98]. Gemma-3 is pretrained on a diverse multilingual corpus and supports a 128K token context window, enabling the processing of long documents and extended conversational contexts.

- **Phi-4**: a 14-billion-parameter open transformer-based model released by Microsoft in December 2024, is the latest member of the Phi family of small language models

[99]. It demonstrates strong performance in complex reasoning tasks, particularly in mathematics and conventional language processing.

- **Command R7B Arabic**: the most recent release of Cohere's lightweight Command R7B model, introduced in February 2024, is specifically tailored to address the complexities of the Arabic language [100]. Command R7B Arabic exhibits improved performance in capturing both linguistic and cultural nuances of Arabic.

PROMPT ENGINEERING AND INFERENCE PARADIGMS Prompt engineering constitutes a central operational mechanism for controlling LLMs behavior at inference time. It involves the deliberate design of natural language prompts that specify task instructions, contextual information, and expected output formats. Since LLMs rely on prompts rather than fine-tuning or weight updates for task adaptation, prompt design directly influences performance, robustness, and reproducibility [101].

Prompting strategies can be broadly categorized into three main paradigms, each reflecting a different level of task guidance and contextual supervision:

- **Zero-Shot Learning:** In zero-shot prompting, the model is provided only with a task description and an input instance, without any labeled examples. The model relies entirely on its pretrained knowledge to infer the desired output [102]. This strategy is attractive for low-resource scenarios and rapid experimentation; however, performance may be unstable for fine-grained tasks such as IER.

- **Few-Shot Prompting:** Few-shot prompting augments the task description with a small number of labeled input–output examples embedded directly within the prompt. These examples act as task demonstrations, guiding the model toward the desired reasoning pattern. Few-shot prompting has been shown to significantly improve performance on classification tasks, including emotion recognition, by reducing ambiguity in task interpretation [92].

- **Chain-of-Thought Prompting:** Chain-of-Thought (CoT) prompting encourages the model to generate intermediate reasoning steps before producing a final answer. By explicitly prompting the model to reason step by step, CoT prompting enhances interpretability and improves performance on tasks requiring contextual inference and commonsense reasoning [93].

2.4.5 *Evaluation Metrics for Emotion Classification*

Emotion classification models are evaluated using quantitative measures derived from the comparison between predicted labels and true labels. The selection of appropriate performance measures is crucial to ensuring a rigorous and reliable assessment of model behavior. Commonly adopted measures include accuracy, precision, recall, and the F1-score. These metrics are defined as shown in (2.1), (2.2), (2.3), and (2.4), respectively.

2.4.5.1 *Accuracy*

Measures the proportion of correctly classified instances among all evaluated samples. It provides a global indication of the model's predictive performance and is defined as follows:

$$Accuracy = \frac{\sum_{c \in C} TP_c}{N} \tag{2.1}$$

where N is the total number of samples in the dataset.

2.4.5.2 *Precision, Recall, and F1-Score*

For C classes, these metrics are usually calculated per class first and then averaged:

- **Precision**: Measures how many of the items predicted for a specific class actually belong to it.
- **Recall**: Measures how many items of a specific class were correctly identified by the model.
- **F1-score**: The harmonic mean of Precision and Recall, providing a balance between the two.

For a given class c, we define:

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \tag{2.2}$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \tag{2.3}$$

$$F1_c = 2 \cdot \frac{Precision_c \cdot Recall_c}{Precision_c + Recall_c} \tag{2.4}$$

where:

- TP_c denotes the number of correctly predicted instances of class c,
- FP_c denotes the number of instances incorrectly predicted as class c,
- FN_c denotes the number of instances of class c incorrectly predicted as another class.

2.4.5.3 *Averaging Techniques*

To combine per-class scores into one global metric, three main strategies are used:

- **Macro-average**: Computes the metric independently for each class and then averages them, giving equal importance to all classes.
- **Micro-average**: Aggregates the contributions of all classes before computing the metric, favoring frequent classes.
- **Weighted-average**: Computes the average weighted by the number of true instances for each class.

2.4.5.4 *Multi-label Performance Measures*

In multi-label emotion classification, where an instance may be associated with multiple emotional states simultaneously, standard multiclass metrics are insufficient to capture the degree of overlap between predicted and ground truth label sets. To provide a more granular assessment, we incorporate the following specialized measures:

- **Exact-Match Accuracy**: Also known as subset accuracy, this is the most stringent metric. It calculates the percentage of samples whose predicted labels are the same as their gold labels. This measure doesn't consider the partially correct, regarding them as incorrect.
- **Jaccard Index (Multi-label Accuracy)**: This measures the average overlap between the predicted and true label sets. It is defined as the size of the intersection divided by the size of the union of the label sets.

- **Hamming Loss**: Calculates the average of how many times a label of an instance is classified incorrectly.

For a dataset of N instances and L labels, let T_i represent the set of true labels and P_i represent the set of predicted labels for instance i. These metrics are defined as follows:

$$Exact\text{-}Match = \frac{1}{N}\sum_{i=1}^{N} I(T_i = P_i) \tag{2.5}$$

$$Jaccard\ Index = \frac{1}{N}\sum_{i=1}^{N} \frac{|T_i \cap P_i|}{|T_i \cup P_i|} \tag{2.6}$$

$$Hamming\ Loss = \frac{1}{N \cdot L}\sum_{i=1}^{N}\sum_{j=1}^{L} I(t_{i,j} \neq p_{i,j}) \tag{2.7}$$

where $I(\cdot)$ denotes the indicator function, $t_{i,j}$ is the ground truth for label j in instance i, and $p_{i,j}$ is the corresponding model prediction. It is important to note that while higher values are desirable for the Jaccard Index and Exact-Match Accuracy, a lower value for Hamming Loss indicates superior model performance.

2.5 GAPS AND CHALLENGES IN IMPLICIT EMOTION RECOGNITION

Compared to explicit emotion recognition, IER presents a substantially complex and multifaceted set of challenges that significantly hinder the development of accurate computational tools.

- One of the primary challenges lies in the limited effectiveness of traditional lexicon-based approaches, which have been the dominant approach in sentiment and emotion analysis. Since implicit emotional expressions often lack explicit affective vocabulary, matching text against emotion lexicons such as the NRC Emotion Lexicon [103] or ArSenL [104] frequently provides little or no useful emotional signal. As a result, many conventional NLP techniques designed for emotion recognition become inadequate for implicit emotion analysis.

- Another major challenge concerns data annotation and resource construction. Although supervised ML models can be trained using annotated datasets, the cre-

ation of high-quality implicit emotion corpora remains both time-consuming and resource-intensive. Annotators must infer emotions from contextual and situational cues rather than rely on explicit emotional markers, which often requires substantial linguistic expertise and interpretative judgment.

- Furthermore, emotion is subjective and context-dependent by its very nature. The same textual expression may evoke different emotions depending on the reader's personal experiences, cultural background, social norms, or situational interpretation. This subjectivity complicates both annotation and evaluation processes, often resulting in lower inter-annotator agreement compared to explicit emotion recognition tasks.

- IER also requires deep contextual understanding that goes beyond surface syntactic and semantic analysis. Computational models must be able to interpret different linguistic signals and integrate world knowledge in order to infer the emotional meaning of a text. This demands representations that capture long-range contextual dependencies and cultural background knowledge capabilities, which are only beginning to be approximated by modern deep learning architectures.

- In the Arabic language, these challenges become even more pronounced due to its language-specific characteristics. The morphological complexity complicates lexical generalization and semantic representation. In addition, linguistic variation across Arabic varieties further increases the difficulty of the task, as emotional expressions and implicit affective cues differ substantially between MSA and regional Arabic dialects. Consequently, models trained on a particular dialect or on MSA often demonstrate limited generalization capability when applied to other dialectal contexts.

2.6 CONCLUSION

This chapter has outlined the theoretical foundations of emotion recognition in text, covering psychological theories, models, types of emotion expression, and computational approaches. It also highlighted the unique challenges posed by the Arabic language for IER. These foundations provide the necessary background for understanding the methodologies and contributions presented in the subsequent chapters.

3

LITERATURE REVIEW

This chapter reviews prior research relevant to IER and situates the present work within the broader landscape of affective computing and emotion-aware NLP. We begin by surveying existing Arabic emotion datasets, highlighting the resources currently available and the limitations that motivate the need for new annotated corpora. We then examine current datasets for IER, which have played a central role in shaping current research directions and evaluation benchmarks. Next, we review the state-of-the-art in IER, covering the evolution of approaches from rule-based and classical machine learning methods to deep learning, and transformer-based architectures. Finally, we outline the general considerations that guide this thesis..

3.1 RESOURCES AND DATASETS

The availability of annotated corpora is a fundamental prerequisite for supervised emotion recognition research, yet Arabic remains considerably under-resourced in this regard. An overview of Arabic emotion recognition resources identified only a small number of labeled datasets available in the literature for Arabic, a stark contrast to the volume of annotated corpora available for English. This section surveys the most commonly used datasets, examining their construction methods, annotation strategies, and the limitations that constrain their utility for broader research.

3.1.1 *Arabic Emotion Datasets*

The overwhelming majority of existing Arabic emotion datasets have been assembled through scraping social media platforms, most notably Twitter. One of the earliest dedicated resources for Arabic emotion analysis is the DINA dataset, introduced by Abdul-Mageed et al. [105], which comprises 3,000 tweets collected from Twitter and written across various Arabic dialects, manually annotated according to Ekman's six basic emotion categories: anger, disgust, fear, happiness, sadness, and surprise. Building directly on DINA, Alhuzali et al. [106] introduced the LAMA dataset, which contains 7,268 tweets covering Plutchik's eight basic emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. The expansion to Plutchik's richer taxonomy represents a theoretical advancement over Ekman's six-category scheme. Al-Khatib and El-Beltagy [107] presented the Arabic Emotions Twitter Dataset (AETD), primarily consisting of Egyptian dialect tweets. AETD includes 10,065 examples, each labeled with one of seven emotion tags (anger, fear, happiness, love, sadness, surprise, or sympathy) or marked as neutral in the absence of emotional content. One of the most widely benchmarks for Arabic emotion recognition was constructed as part of the SemEval-2018 Task 1: Affect in Tweets, which produced a labeled dataset of 4,381 tweets drawn from multiple Arabic dialects and annotated across eleven emotion categories: anger, anticipation, disgust, fear, happiness, love, optimism, pessimism, sadness, surprise, and trust [33]. Similarly, the SenWave dataset consists of 10,000 Arabic tweets annotated into ten categories, collected during the early months of the COVID-19 pandemic and spanning a variety of Arabic dialects [108].

Some efforts have expanded beyond Twitter. The ArPanEmo dataset, comprising 11,128 posts manually labeled for ten emotion categories, was collected from Twitter, YouTube, and online newspaper comments spanning March 2020 to March 2022, offering a somewhat broader textual landscape [109]. The Iraqi Arabic Emotion Dataset (IAEDS) drew its 1,365 examples from Facebook posts, labeled according to Ekman's six basic emotion categories [110]. While social media provides a convenient and large-scale source of naturally occurring emotional text, this data collection strategy introduces significant constraints: instances are inherently short and contextually sparse. Additionally, these broader collections remain anchored to social media ecosystems and fail to represent the full spectrum

of Arabic textual registers, including formal writing, literary expression, and instructional or legal discourse.

Datasets have been annotated through a range of methodologies, each carrying distinct trade-offs between quality, scalability, and consistency. Manual annotation by human experts or crowdsourced workers remains the gold standard for reliability, but it is labor-intensive. The ArPanEmo dataset employed a semi-automatic process in which posts were first filtered using a lexicon of emotion-related terms to separate emotional from neutral content, and then subjected to fine-grained manual labeling, a hybrid approach designed to manage scale while preserving annotation quality, achieving a Fleiss' kappa of 0.71 [109]. In contrast, the IAEDS was evaluated by four external judges, yielding an average inter-annotator agreement of 0.751 [110].

Beyond data collection and annotation, existing Arabic datasets are also limited in the type of emotional expression they capture. Most resources focus exclusively on explicit emotion recognition, targeting emotions that are directly expressed through affective words and phrases. In contrast, implicit emotion recognition remains largely underexplored in Arabic NLP, partly due to the language's complex linguistic and morphological characteristics. This limitation is particularly important, as a large proportion of emotional communication in Arabic, especially in culturally rich contexts, relies on indirect expression and metaphor rather than overt emotional vocabulary.

3.1.2 *Implicit Emotion Datasets*

Recently, the IER task has garnered significant attention in the field of NLP. However, resources for this task remain limited and are mainly available for high-resource languages such as English and Chinese.

Emotions are often expressed explicitly through the use of direct emotional words such as happy, sad, or angry. For explicit emotion recognition, numerous datasets and corpora have been compiled and made available for English, such as Alm's Affect dataset [111], which comprises stories annotated for affect at the sentence level, Aman [112] which includes emotion-rich sentences sourced from blogs and labeled with the six Ekman emotion categories, SemEval-2019 Task 3 [113], GoEmotions [114], and EmoBank [115]. However, emotions can also be conveyed implicitly, without the use of such explicit emotional terms.

This makes IER in text an extremely challenging task. Despite its importance, only a limited number of datasets have been specifically designed for this purpose. A comprehensive analysis was conducted to explore well-known datasets, with a particular focus on annotation methods, the presence of emotional words, and the cause or stimulus of the emotion.

One of the first widely used resources for emotion recognition is the International Survey on Emotion Antecedents and Reactions (ISEAR) dataset [116]. The corpus comprise self-reported emotional events, where participants in a psychological study described situations associated with a specific emotion without explicitly stating the emotion itself. The dataset contains 7,665 instances, manually labeled by experts with one of seven emotion classes: joy, fear, anger, sadness, disgust, shame, and guilt. These reports were collected from approximately 3,000 students in 37 countries around the world.

In recent years, social networks have become essential for data acquisition through distant supervision, enabling a cheap and fast way to collect large amounts of automatically labeled data; this strategy was employed in the WASSA 2018 Implicit Emotions Shared Task (IEST) [41]. The IEST dataset is a large automatically labeled dataset consisting of 191,731 tweets annotated with Ekman's basic emotions (anger, fear, sadness, joy, surprise, and disgust). In this dataset, explicit emotional words were masked in the tweets. This strategy was used to challenge models to infer emotions only from contextual cues, particularly from the cause or stimulus that triggers the emotion.

For Chinese, the IER task has also received significant attention. The Chinese Implicit Emotion Analysis (CIEA) corpus was created by [117] based on the WASSA 2018 IEST [41]. In this corpus, emotional words are masked, and the data is used to predict implicit emotions. The corpus consists of 285,000 self-labeled microblogs collected from the Chinese microblogging website Sina Weibo, each annotated with one of Ekman's basic emotions.

Some datasets with explicit emotion have been used for IER, often with additional preprocessing. As an example, Udochukwu and He [118] used the SemEval-2007 [7] and Alm [8] datasets in their work. To ensure that implicit emotions could not be easily identified, the authors removed all sentences containing emotion words listed in the WordNet-Affect emotion lexicon. Table 3.1 provides an overview of the related resources for IER

Dataset	Source	Emotion Classes	Language	Size	Annotation Type	Cause	Masking
ISEAR [116]	Self-reports	Joy, Fear, Anger, Sadness, Disgust, Shame, Guilt	English	7,665	Expert annotation	✓	✗
IEST [41]	Twitter	Ekman's basic emotions	English	191,731	Distant supervision	✓	✓
CIEA [117]	Sina Weibo	Ekman's basic emotions	Chinese	285,000	Self-labeling	✗	✓

Table 3.1: Related resources for Implicit Emotion Recognition.

3.2 IMPLICIT EMOTION RECOGNITION: STATE OF THE ART

Approaches to IER can be categorized into rule-based, ML-based, DL-based, and hybrid methods. For more details, see [119]. Knowledge-based and rule-based techniques [16, 118, 120], as well as classical ML methods such as SVM, DT, and Naïve Bayes (NB) classifiers [16, 121, 122], have been proposed to address the challenge of recognizing implicit emotions.

Early work by [120] introduced EmotiNet, a commonsense knowledge base of concepts with associated affective values, designed to detect event-triggered emotions from text. The authors propose an approach to automatically detect emotions in contexts where no explicit emotional cues are present, leveraging commonsense knowledge. The proposed method achieved an accuracy of 42.24% in recognizing implicit emotions on the ISEAR dataset.

The same authors in [16] compared the classification performance of three methods: supervised learning, lexicon-based, and the EmotiNet knowledge base approaches. The results demonstrated that using EmotiNet as a resource for IER outperformed the other methods, achieving an F1-score of 45.27% on the ISEAR dataset.

A rule-based approach was proposed by [118], based on cognitive emotion theory. This method utilized a set of crafted rules to identify emotions associated with specific event types. In contrast to earlier methods that used lexicons and sentiment analysis, this method required measuring variables derived from the Ortony, Clore, and Collins (OCC) [123] theory and combining them with logical rules to infer emotion classes. On the ISEAR, SemEval-2007, and Alm datasets, the proposed approach outperformed lexicon matching and NB methods, achieving F1-scores of 51.3%, 57.3%, and 65.5%, respectively.

In [121], the authors combined an SVM classifier with a statistical method based on Vector Space Model (VSM) and a keyword-based approach to demonstrate its effectiveness in recognizing implicit emotions. Each subsystem analyzes input information from a different perspective and generates an emotion label as output. The information fusion component then aggregates these results to determine the final emotion label. Compared to the individual subsystems, the combined approach outperformed the SVM by 9.13%, VSM by 16.6%, and the keyword-based method by over 23% on the ISEAR dataset.

Similarly, the authors in [122] proposed DuFER, the Dual Framework for Emotion Recognition that combines emotional, categorical, and hierarchical models. They employed a modified Term Frequency-Inverse Document Frequency (TF-IDF) weighing and Maximum Likelihood Estimation weighing to train three classifiers: DT, NB, and SVM. Their results demonstrated that the model achieved an F1-score of 68.93% in recognizing implicit emotions on the IEST dataset.

Recently, Zhao and Zhang [124] proposed a ML–based framework for classifying emotions conveyed indirectly in narrative texts. Their approach applies the Boruta algorithm to identify the most relevant features and then evaluates multiple classifiers, including SVM, KNN, and Artificial Neural Networks (ANN). Experimental results show that SVM achieves the best performance, reaching an accuracy of 91%, outperforming both ANN and KNN. These findings highlight the effectiveness of combining feature selection with traditional ML methods for detecting implicit emotional cues in literary narratives.

This task has seen significant advancements in recent years, especially since the WASSA 2018 IEST. End-to-end DL models incorporating attention mechanisms have demonstrated superior performance and interpretability, as evidenced by studies such as [125–132]. Rozental et al. [125] introduced a novel use of a language model as an emotion classifier in their proposed system 'Amobee', in the WASSA 2018 shared task. The authors used a transformer-based decoder architecture for the language model. Based on a custom-built Twitter dataset, the model combines an LSTM-based network enhanced with a CNN-based attention mechanism, resulting in the best performance with a macro F1-score of 71.45%.

The second best performance in WASSA-2018 was achieved by a transformer-based model [126]. This model incorporated a pre-trained ELMo layer for word embeddings, a BiLSTM layer for contextual learning, and a max-pooling operation followed by a dense layer for final prediction. It obtained an F1-score of 71.05% on the IEST dataset.

Another strong system in the competition was NTUA-SLP, proposed by [127]. This approach employed an LSTM network with self-attention mechanisms and neural TL methods. First, a word2vec embedding model was pre-trained on a large corpus of unlabeled Twitter messages. Second, a pre-trained classifier originally developed for sentiment analysis was used to encode emotion information. Third, a language model was trained on a Twitter corpus. Finally, the weights from these models were transferred to specific layers. By combining the predictions of the individual TL models, the NTUA-SLP system achieved an F1-score of 70.29%.

Similarly, the authors in [129] employed BiLSTM and LSTM models with an attention mechanism for IER. By tuning different dropout rates and applying ensemble methods, they improved classification performance and achieved an F1-score of 68.48%.

A BiGRU model combined with capsule networks was proposed by [128]. The BiGRU was utilized for learning latent sentence representations, while the capsule network extracted high-level features to implicitly identify emotions. The model achieved an F1-score of 69.20%.

CNN architectures have also been applied to the IER task using ensemble classifier approaches. For example, the authors in [130] proposed a system that combined CNN-based and FastText classifiers, incorporating word-level Valence, Arousal, and Dominance (VAD) features alongside ELMo embeddings for emotion prediction. This approach achieved an F1-score of 66.15% on the IEST dataset.

Pecar et al. [131], highlighted the importance of multilevel pre-processing techniques in recognizing implicit emotions. The authors introduced a BiLSTM model that incorporated GloVe and ELMo layers for word embeddings, as well as the InferSent model and Universal Sentence Encoder (USE) for sentence embeddings. The proposed approach achieved a macro-average F1-score of 65.55%.

To model implicit emotion objectives as latent variables in classification, [132] proposed a novel implicit objective network for the IER task. The approach employed a variational autoencoder with a BiLSTM encoder enhanced by a multi-head attention mechanism. It achieved F1-scores of 75.2% on the ISEAR dataset and 65.8% on the IEST dataset.

Recent studies have demonstrated that pre-trained models such as BERT and RoBERTa achieve SOTA performance across a range of NLP tasks, including IER [133–135].

To analyze the text at both the word and sentence levels, Khoshnam et al. [133] proposed a transformer-based framework that employs Robustly optimized BERT approach (RoBERTa) to extract feature vectors for words and sentences. The framework was combined with an emotional vector generated using the EmoSyn Emotional Dictionary to determine the final emotional label of a sentence. By integrating emotional vectors with contextual features, the framework outperforms RoBERTa by 3%, achieving an F1-score of 70.57% on the IEST dataset.

Similarly, Qian et al. [134] employed an interactive capsule network that leverages BERT as an encoder to extract semantic features from both preceding and following contexts. The model achieved F1-scores of 73.1% and 76.5% on the CIEA and IEST datasets, respectively, outperforming CNN, GRU, BiGRU, Attention BiGRU, Capsule BiGRU, BERT, ALBERT, and RoBERTa models.

Recently, Feng et al. [135] proposed a RoBERTa-based fine-grained framework for detecting implicit sentiment in social media texts. Their approach first generates contextual word representations using RoBERTa, after which complementary semantic features are extracted through two parallel modules: a BiLSTM for capturing contextual dependencies and a TextCNN for modeling local n-gram patterns. To overcome the limitations of traditional multi-feature fusion strategies, the authors introduce an adaptive fusion mechanism based on a GRU that dynamically balances the contribution of contextual and local features. Evaluated on the IEST dataset, the proposed model improves accuracy by 1.45% to 72.93%, and the F1 score by 1.4% to 72.85%, compared to the best baseline model, Amobee [125].

Table 3.2 provides a comparative analysis of essential approaches employed to address the IER problem. The studies are compared based on the datasets used, feature representation methods, learning algorithms, and achieved results, offering a comprehensive resource for researchers in this field.

3.3 GENERAL CONSIDERATIONS IN THE THESIS

Despite the growing volume of resources for Arabic NLP, a critical examination of existing literature reveals a significant deficiency in corpora dedicated to IER. While section 3.1.1 highlighted several prominent datasets for Arabic explicit emotion detection, these resources predominantly rely on explicit linguistic cues. A vast majority of these datasets

Study	Dataset	Approach	Feature Representation	Learning Algorithm	Metrics	Result (%)
[120]	ISEAR	Commonsense knowledge-based	Ontology	Emotinet knowledge base	Accuracy	42.24
[16]	ISEAR	Commonsense knowledge-based	The presence of words in WordNet Affect and LIWC, n-grams, inter-example similarity	SVM SMO, lexicon knowledge-based method, Emotinet knowledge base	F1-score	45.27
[118]	ISEAR, SemEval-2007, Alm	Rule-Based approach	POS Tagging	Lexicon matching method, NB, Rule-based method	F1-score	51.30, 57.30, 65.50
[121]	ISEAR	Hybrid approach	n-grams (n = 1, 2), POS tags, Frequency of emotion words in WordNet-Affect, ANEW	Combinational System (SVM, VSM and Keyword-based module)	F1-score	68
[124]	Self-compiled literary corpus	ML approach	Boruta algorithm	SVM, KNN, and ANN	Accuracy	91%
[125]	5B Tweets, Semeval 2018 shared task 1, IEST	Ensemble approach	Universal Sentence Embedding, DeepMoji	LMs with LSTM-based networks containing a CNN attention mechanism	F1-score	71.45
[126]	IEST	DL approach	Pretrained ELMo	BiLSTM	F1-score	71.05
[127]	550M Twitter messages, SemEval17 Task4A, IEST	Transfer Learning	Word2vec	Ensemble TL + LSTM with self-attention mechanism	F1-score	70.29
[128]	IEST	Ensemble approach	Word2vec	Combination of BiGRU with capsule networks	F1-score	69.20
[129]	IEST	DL approach	Word2vec	BiLSTM + LSTM with attention mechanism	F1-score	68.48
[130]	IEST	DL approach	Pretrained ELMo + VAD features	Stacked CNN blocks	F1-score	66.15
[131]	IEST	DL approach	GloVe, ELMo, Universal Sentence Encoder, and InferSent	BiLSTM with dropout and Gaussian noise regularization	F1-score	65.55
[132]	ISEAR, IEST	DL approach	GloVe	BiLSTM-based VAE with multi-head attention mechanism	F1-score	75.20, 65.80
[122]	Test data of IEST	ML approach	Modified TF-IDF weighing, Maximum Likelihood Estimation	NB, SVM, DT	F1-score	68.93
[133]	Test data of IEST	Transformer-based	BERT and RoBERTa embeddings	Combining RoBERTa with Emotional vector	F1-score	70.57
[134]	IEST, CIEA	Transformer-based	BERT embedding	Interactive capsule network	F1-score	76.50, 73.10
[135]	IEST	Transformer-based	RoBERTa embedding	Combining RoBERTa embeddings + BiLSTM + TextCNN + GRU fusion	F1-score, Accuracy	72.85, 72.93

Table 3.2: A summary of previous works on IER in textual data.

are constructed using distant supervision techniques, where data is collected based on the presence of specific emotion keywords, emojis, or hashtags. While this method allows for the rapid accumulation of large-scale data, it introduces a fundamental bias: it was developed to recognize explicit emotions, leaving the phenomenon of implicit emotions largely unrepresented.

This gap represents a major challenge in the advancement of Arabic IER. Without a gold-standard dataset that isolates implicit instances, it is impossible to train models that genuinely "read between the lines" rather than merely spot keywords. This deficiency serves as the primary motivation for our first contribution. In Chapter 4, we address this scarcity by constructing and annotating a specialized dataset for MSA, specifically curated to exclude explicit emotion keywords, thereby providing the necessary foundation for training and evaluating the models developed in subsequent sections.

Beyond the resource gap, a second critical limitation concerns the modeling strategies employed in existing Arabic affective NLP research. The majority of proposed models treat implicit and explicit emotion recognition as equivalent classification tasks, applying identical modeling pipelines regardless of whether the emotional content is explicitly expressed or must be inferred from contextual cues. This architectural indifference fails to account for the fundamentally different cognitive and linguistic processes underlying implicit emotional interpretation and results in models that perform adequately on explicit instances while systematically underperforming on implicit ones. The need for modeling approaches specifically designed to capture latent emotional meaning motivates the second contribution of this thesis, presented in Chapter 5, where we propose a hybrid contextual-sequential architecture tailored to the demands of IER in Arabic text.

A third and equally significant gap concerns the evaluation of latest LLMs in the context of Arabic affective understanding. Despite the rapidly growing deployment of LLMs across NLP tasks, their capacity to detect implicit emotions remains poorly understood, particularly in low-resource and dialectal settings. Existing evaluations of LLMs emotion tasks have focused predominantly on high-resourced languages and on explicitly expressed emotions, leaving open the critical question of whether these models can generalize to the linguistic, cultural, and dialectal complexity that characterizes real-world Arabic emotional communication. This gap is especially pronounced for the Algerian Arabic dialect, which remains almost entirely absent from existing LLM benchmarks despite being spo-

ken by over 40 million people. The third contribution of this thesis, presented in Chapter 6, directly addresses this gap through a systematic cross-lingual evaluation of state-of-the-art LLMs on emotion recognition in both Algerian Arabic dialect and English, providing the first rigorous assessment of LLMs generalization capacity across these contrasting linguistic settings.

3.4 CONCLUSION

In summary, the literature reveals substantial progress in emotion recognition and affective NLP, particularly for English and other high-resource languages. Numerous datasets and modeling approaches have contributed to the development of increasingly sophisticated systems. However, the review also exposes several important gaps. Arabic remains underrepresented in both resource availability and methodological exploration, especially for implicit emotion recognition.

These limitations underscore the need for dedicated Arabic resources and systematic evaluation of modern modeling approaches. The following chapters build upon these observations by introducing a new dataset and investigating computational methods designed specifically to address the challenges of implicit emotion recognition in Arabic text.

4

BUILDING THE FOUNDATION: ANNOTATED DATASET AND BASELINE MODELS FOR ARABIC IMPLICIT EMOTION RECOGNITION

4.1 INTRODUCTION

This chapter presents the first core contribution to this thesis. It establishes a structured foundation for Arabic IER through the construction of a new emotional Arabic dataset and the evaluation of several modeling approaches.

While existing approaches have primarily focused on explicit emotion recognition, relatively little attention has been given to implicit emotion, especially in Arabic text. In contrast to many NLP tasks, the construction of annotated datasets for IER is particularly challenging, as it requires careful control over lexical emotion triggers and relies heavily on human inference and contextual understanding. Moreover, for Arabic, such resources remain extremely scarce due to its rich morphology and complex syntactic structures, even when restricted to MSA. In this chapter, we address this limitation by tackling a fundamental challenge for IER: the availability of high-quality annotated data and reliable benchmark models.

To this end, we introduce the AIER dataset, a publicly available emotional dataset consisting of real-world Arabic reviews written in MSA. AIER is annotated with emotions based on Ekman's basic emotions, in addition to emotional cues and causes. This multi-level annotation framework enables fine-grained analysis of emotional expression and constitutes a valuable resource for a wide range of Arabic classification tasks. The dataset is inspired

by the formulation of the WASSA 2018 IEST [41] for English but adapted to the linguistic characteristics of Arabic.

In addition to dataset construction, we benchmark a wide range of computational approaches. These include classical ML models, DL architectures, pretrained language models, and generative LLMs. The objective is to establish strong and reproducible baselines and to analyze the relative effectiveness of different modeling approaches in capturing latent emotional meaning beyond lexical cues.

CONTRIBUTIONS: Our contributions in this chapter are as follows:

1. We present a novel annotated dataset for Arabic IER, specifically designed to model emotional meaning inferred from context rather than explicit emotion-bearing words. The dataset is made publicly available at https://github.com/hananebtt/Arabic-Implicit-Emotion-Recognition-Dataset-AIER-.git.

2. We conduct a comprehensive evaluation of various classification approaches, including classical ML models, DL architectures, Arabic pretrained language models, and generative LLMs under a unified experimental setup. This evaluation establishes reproducible baselines for Arabic IER.

3. Our analysis demonstrates that contextualized transformer models outperform traditional approaches while also revealing persistent challenges in capturing nuanced implicit emotional cues. These findings highlight the difficulty of the task and motivate the development of more advanced modeling strategies, which will be explored in subsequent contributions to this thesis.

This work represents, to the best of our knowledge, the first effort to introduce an annotated Arabic dataset for implicit emotion, accompanied by an extensive baseline evaluation across multiple modeling approaches. The findings of this chapter provide a solid empirical foundation for the subsequent contributions of this thesis, which extend the investigation of IER to other Arabic varieties and modeling settings.

DECLARATION: This chapter is based on the following journal publication:

- H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani,"Enhancement of Implicit Emotion Recognition in Arabic Text: Annotated Dataset and Baseline Models," *IEEE Access*, vol. 13, pp. 165096–165116, 2025, doi: 10.1109/ACCESS.2025.3611337 [136].

The chapter is presented largely in alignment with the published manuscript. Referencing, notation, and structural organization have been adapted for consistency across the thesis, with minor stylistic and formatting changes introduced where necessary.

4.2 DATASETS IN EMOTION RECOGNITION

Datasets constitute a foundational component of emotion recognition research, as they directly shape the formulation of tasks, the design of computational models, and the validity of experimental evaluation. Although substantial progress has been made through the development of numerous emotion datasets across languages and domains, existing resources exhibit fundamental limitations in their ability to model implicit emotional expression, which is prevalent in natural language.

Chapter 3 specifically Section 3.1 reviews existing emotion datasets with a particular emphasis on their capacity to support IER. An investigation of the datasets shows that most available corpora are tailored to explicit emotion recognition, only a small number of datasets attempt to capture emotions inferred from context, events, or underlying causes.

Table 4.1 presents an overview of commonly used Arabic emotion datasets, such as DINA [105], AETD [107], SemEval-2018 Ec-Ar [33], and AraEmoCorpus [109]. These datasets have contributed significantly to the development of Arabic emotion recognition models; however, they largely assume that emotional states are directly expressed through affective words. In addition, we found that only a limited number of non-Arabic datasets explicitly address IER. Notable examples include the IEST [41] dataset in English and CIEA [117] in Chinese, which are specifically designed to infer emotions from context. The scarcity of such datasets, even in high-resource languages, highlights the difficulty of annotating implicit emotions and the lack of standardized benchmarks for this task.

Importantly, no publicly available Arabic dataset has been specifically constructed to support IER, resulting in a significant gap in resources necessary to advance contextual emotion understanding in Arabic. Motivated by this gap, our work introduces the AIER

dataset, which serves as a dedicated benchmark for IER in Arabic. By incorporating annotations of emotions and their roles (cues and causes), the proposed dataset aims to facilitate robust evaluation and promote research toward deeper, context-aware emotion modeling in Arabic text.

Arabic Datasets		**Non-Arabic Datasets**		
Dataset	**Task**	**Dataset**	**Task**	**Lang**
DINA [105]	EER	ISEAR [116]	EER	ENG
LAMA-DINA[106]	EER	Alm [8]	EER	ENG
AETD [107]	EER	SemEval-2007 [7]	EER	ENG
SemEval-2018 Ec-Ar [33]	EER	IEST [41]	IER	ENG
IAEDS [110]	EER	CIEA [117]	IER	CHI
AraEmoCorpus [109]	EER			
ArPanEmo [137]	EER	**Our Dataset**		
Arabic-Poem-Emotion [138]	EER	**AIER (this thesis)**	IER	ARA

Table 4.1: Overview of publicly available Arabic and non-Arabic datasets used in emotion recognition, indicating the corresponding task (EER: Explicit Emotion Recognition; IER: Implicit Emotion Recognition) and language.

4.3 THE AIER DATASET

The AIER dataset is a novel MSA corpus consisting of Arabic reviews collected from widely used Arabic sentiment analysis resources. We annotate the dataset with emotion labels based on the six basic emotion categories of Ekman [6], in addition to corresponding emotional cues and causes. This multi-level annotation framework enables fine-grained analysis of emotional expression and constitutes a valuable resource for a wide range of Arabic classification tasks.

The development of the dataset involved five key steps: collection, annotation, validation, annotator agreement, and implicit scenario simulation. Figure 4.1 illustrates the overall process of AIER dataset development.

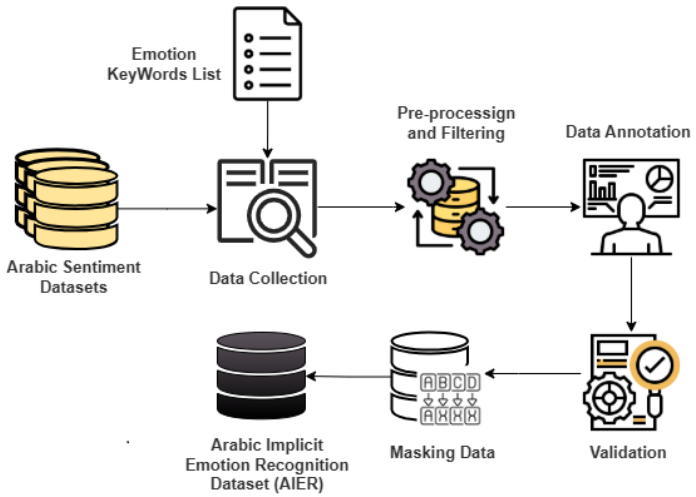


Figure 4.1: An overview of the AIER dataset building process

4.3.1 *Data Collection*

Creating a new dataset requires significant time and effort. For this project, a rich types of Arabic text samples was collected from existing Arabic sentiment resources. These include utterances from the Hotel Arabic Review Dataset (HARD) [139], the Book Reviews in Arabic Dataset (BRAD) [140], and reviews of movies, hotels, restaurants, and products from Elsahar and El-Beltagy [141] dataset. The original data consisted of approximately one million reviews, written in both dialectal Arabic and MSA, and annotated for polarity (positive, negative, or neutral). Arabic reviews were selected as the primary text data due to their abundance and diversity, which facilitates a comprehensive analysis of emotions across various contexts. This variety not only enriches the dataset but also enables the capture of subtle linguistic patterns.

The AIER dataset was carefully collected using the following criteria:

1. For each of Ekman's basic emotions (anger, fear, disgust, joy, surprise, and sadness), a set of commonly used synonyms was selected as query words based on the NRC Emotion Lexicon [103]. Table 4.2 presents the query words used for dataset extraction.

2. For each query word, a search was performed across the raw data, and all matching instances were saved.

3. Reviews written in Arabic dialects or containing multiple emotional words, including synonyms and antonyms, were excluded.

4. If an emotional word in a review was identified as a proper noun, the review was eliminated.

5. Reviews containing fewer than six tokens were also removed.

6. These steps were meticulously carried out by three PhD students to ensure consistency and accuracy.

Finally, light pre-processing was conducted, as the selected datasets had already been pre-processed by their original authors. During this phase, Arabic diacritics, non-Arabic characters, URLs, email addresses, emoticons, and punctuation marks were removed. Reviews containing mixed-language content or neutral sentiment were also excluded. To preserve the dataset's integrity, all null and duplicate entries were removed.

Emotion	**Query words**
الفرح	سعد، سعيد، سعادة، فرح، فرحة، سرور، بهجة
Joy	happy, joyful, happiness, joy, delight, pleasure, cheer
الخوف	خوف، خاف، خائف، فزع، رعب
Fear	fear, afraid, fearful, panic, terror
الحزن	حزن، حزين، كئيب، كآبة، تعاسة، تعيس
Sadness	sadness, sad, depressed, depression, misery, miserable
الغضب	غضب، غاضب، ساخط، سخط، غاظ، غيظ، حانق، حنق
Anger	anger, angry, resentful, resentment, irritate, fury, enraged, wrath
المفاجأة	فاجأ، مفاجئة، مفاجأة، فجأة، بغتة، بغت
Surprise	surprise, unexpected, sudden, abruptly
الاشمئزاز	مقزز، تقزز، اشمأز، اشمأزاز، قرف
Disgust	disgusting, disgust, nauseate, nausea

Table 4.2: Query words used for data collection.

4.3.2 *Annotation Process*

The dataset was annotated using the semi-automatic annotation tool dedicated to the analysis and annotation of Arabic texts introduced in [142]. This multi-level tool enables the lexical-morphological, syntactic, and semantic analysis of MSA. The system integrates these levels of analysis in a cascading manner, providing multiple layers of information, including the related emotion, its corresponding cue, and its cause [2, 142–144]. The annotation process is followed by a majority voting approach to determine the most appropriate classification.

In total, 12,700 instances were annotated with emotion, cue, and emotion cause using the annotation tool. Three undergraduate students majoring in linguistics carefully annotated these instances in accordance with annotation guidelines and a sample of pre-annotated data, both of which were provided to facilitate the annotation process. The guidelines were as follows:

- **Emotion selection**: Assign one of Ekman's basic emotions (anger, fear, disgust, joy, surprise, or sadness) to the review.
- **Target word selection**: Select the target word (cue) corresponding to the emotion class, and finally,
- **Emotion Cause Identification**: select the description of the cause that motivates the emotion, and flag reviews that lack a motivating cause.

After collecting all annotations, the annotation labels for each instance in the dataset were counted. The label for each review was determined using a majority voting approach. Ultimately, the label receiving the highest number of votes from the annotators was selected as the preliminary label. This annotation process ensured the enrichment of the dataset with subtle contextual cues essential for detecting implicit emotions.

4.3.3 *Validation*

After the initial annotations were completed, the dataset was validated using the semi-automated annotation tool by an academic expert. During this process, reviews lacking a

motivating cause were excluded from the dataset. The initial label assigned to a review by the annotators was accepted as final if it aligned with the expert's judgment. In cases of disagreement, the annotators were encouraged to discuss and reach a consensus on the remaining unresolved data. However, the final decision rested with the expert. Upon completing the validation process, the dataset comprised 5,266 verified instances, each accompanied by its corresponding emotion label, target word, and emotion cause.

4.3.4 *Annotator Agreement*

To evaluate the consistency and reliability of the annotation process, inter-annotator agreement was computed on a randomly selected 10% sample of the annotated dataset.

For emotion classification, two complementary measures, the Cohen's Kappa (k) coefficient [145] and the average percentage of overlap across all annotator pairs, were employed. The Cohen's Kappa (k) coefficient was calculated by (4.1):

$$k = \frac{P_o - P_e}{1 - P_e} \tag{4.1}$$

Where (k) denotes Cohen's kappa coefficient, P_o represents the observed agreement, which is the proportion of instances where the raters agree, and P_e represents the expected agreement, calculated as the proportion of agreement expected by chance.

As Table 4.3 shows, an average Cohen's kappa of 0.82 and a corresponding percentage agreement of 85% were achieved across all pairs of raters, indicating a higher level of agreement in emotion categorization.

Annotator Pair	**Cohen's kappa (k)**	**Agreement (%)**
R1 - R2	0.81	84%
R1 - R3	0.86	88%
R1 - R4	0.81	85%
R2 - R3	0.81	84%
R2 - R4	0.79	83%
R3 - R4	0.81	84%
Average	**0.82**	**85%**

Table 4.3: Agreement statistics for emotion classification.

Table 4.4 and Figure 4.2 present the observed pairwise agreement scores for each emotion category. Notably, the 'joy' and 'surprise' classes achieved significantly high agreement scores, indicating a strong consensus among annotators. These emotions are likely easier to identify due to their salient features and distinctive characteristics. In contrast, the 'fear' and 'anger' classes exhibited the lowest agreement scores, which may be attributed to their overlapping characteristics with other emotions, such as 'sadness' and 'disgust'.

Annotator Pair	Joy	Anger	Fear	Sadness	Surprise	Disgust
R1 - R2	0.939	0.737	0.750	0.846	0.914	0.838
R1 - R3	0.970	0.800	0.853	0.846	0.957	0.897
R1 - R4	0.980	0.737	0.765	0.846	0.886	0.853
R2 - R3	0.949	0.726	0.750	0.856	0.900	0.824
R2 - R4	0.939	0.747	0.765	0.779	0.886	0.853
R3 - R4	0.970	0.737	0.809	0.808	0.871	0.853
Average	**0.956**	**0.747**	**0.784**	**0.830**	**0.902**	**0.853**

Table 4.4: Pairwise observed agreement for each emotion class.

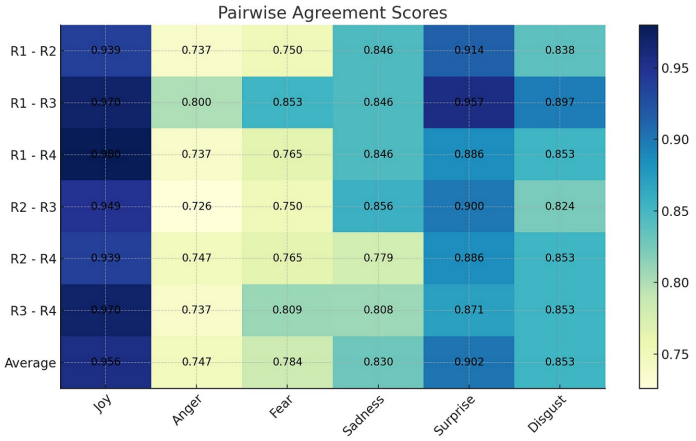


Figure 4.2: Observed agreement result for each emotion.

For emotion cues and cause annotations, where the task involves identifying text spans, inter-annotator agreement was calculated using the F1-score at the span level. This metric captures agreement based on the exact overlap of span boundaries across annotators, providing a robust evaluation of consistency in span-based annotation tasks.

The results indicated a fair level of agreement, with a span-based F1-score of 0.78 for emotion cue and 0.66 for emotion cause annotations. The lower F1-score for cause annotations was expected, given that we did not constrain data collection to instances containing

causal connectives. Their absence created a more challenging annotation setting. Without explicit connectives, annotators were required to rely on contextual cues and world knowledge to infer the underlying causes of emotions, which often led to divergent interpretations of span length. The difficulty of the cause detection task (whether by human annotators or automatic systems) has also been reported for English in [41], even though cause detection there was guided exclusively by causal connectives such as because, that, and when.

4.3.5 *Implicit Scenario Simulation*

Following the WASSA-2018 shared task [41] for English, we simulated implicit emotion scenarios by masking the explicit emotion word in each review and replacing it with a placeholder token [MASK]. This process eliminates direct lexical cues of emotion, directing the model to rely on contextual and linguistic information for inference. Training on such masked texts enables models to learn to recognize implicit emotions, focusing on the surrounding context rather than explicit emotional expressions. Table 4.5 presents some examples from the AIER dataset for implicit emotions.

In these examples, [MASK] refers to explicit emotional words in the utterance. Based on context analysis of the first example, it can be inferred that obtaining something desired leads to the emotion of happiness. In the second example, the daily rise of death is likely to provoke fear among people. Therefore, the emotion class in this case is identified as 'fear'. Similarly, the statement "the writer will never return to her country" suggests that the masked emotional word in the third example corresponds to 'sadness'. Although no explicit emotional words are present in the examples, the emotional classes can be deduced through careful analysis of contextual information.

4.3.6 *Dataset Description*

After the annotation and masking procedures, the AIER dataset comprises a total of 5,266 reviews, containing 62,272 words, of which 17,443 are unique. A detailed statistical summary of the dataset is provided in Table 4.6.

Emotion	Text
الفرح	أنا [MASK] جدا بحصولي على هذا الكتاب وقراءته.
Joy	I am so [MASK] to buy this book and read it.
الخوف	أنا [MASK] من الموت لأن أعداده تزيد في كل يوم.
Fear	I am [MASK] of death because its numbers are increasing every day.
الحزن	النهاية [MASK] فالكاتبة لن تعود أبدا إلى وطنها.
Sadness	The ending is [MASK] because the writer will never return to her country.
الغضب	أخي [MASK] عندما لا يستطيع تحقيق رغباته.
Anger	My brother becomes [MASK] when he can't achieve his desires.
المفاجأة	كان اتصال أخيه بعد غياب طويل [MASK] حقيقة
Surprise	His brother's call after a long absence was a real [MASK].
الاشمئزاز	شعرت ب [MASK] عندما وجدت ملاءات الفندق متسخة.
Disgust	I was [MASK] when I found the hotel sheets dirty.

Table 4.5: Selective examples from the AIER dataset.

Data Attributes	Attributes Value
Total data	5266
Total number of words	62272
Unique words	17443
Maximum sentence length	45
Minimum sentence length	6
Average sentence length	12

Table 4.6: General statistics of the AIER dataset.

As illustrated in Figure 4.3, the shortest text consists of 6 words, while the longest spans 45 words. On average, the texts are relatively brief, with an average length of 12 words.

The overall distribution of emotions within the dataset is relatively balanced, as shown below: Disgust: 15.2%, Surprise: 15.5%, Joy: 15.7%, Sadness: 17.2%, Anger: 18.0% , and Fear: 18.4%. This distribution ensures a comprehensive and fairly balanced representation of the six emotions, with no single emotion dominating significantly. Such balance indicates that the dataset is well suited for tasks requiring equal representation of various emotions, such as IER. Table 4.7 and Figure 4.4 show descriptive statistics of the AIER dataset.

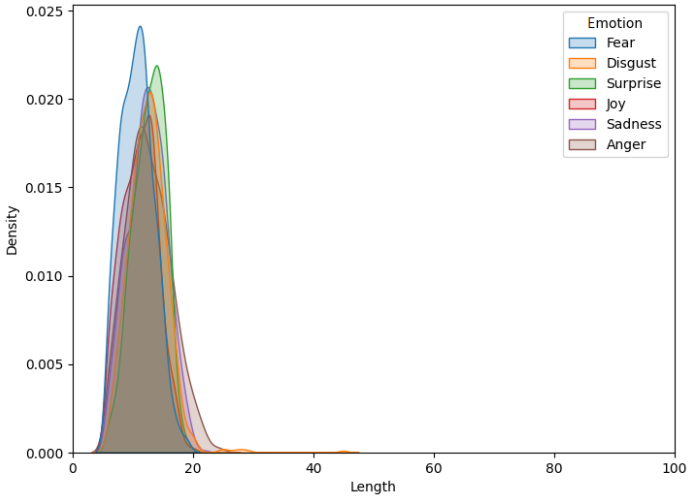


Figure 4.3: Text length distribution.

Class	No. of Sentences	No. of Words	No. of Unique Words
Joy	826	8980	3805
Anger	946	11662	5077
Fear	970	10171	4409
Sadness	904	11047	4499
Surprise	818	10377	4172
Disgust	802	10035	3698

Table 4.7: Dataset statistics across different emotion classes.

To support ongoing research on Arabic NLP, the dataset is publicly available at https://github.com/hananebtt/Arabic-Implicit-Emotion-Recognition-Dataset-AIER-.git.

4.4 IMPLICIT EMOTION MODELING

4.4.1 *BERT-based models*

To situate our work within the context of existing research in the field of NLP, experiments were conducted on the proposed dataset by adapting the most effective Arabic pre-trained language models. These models included AraBERT, CAMeLBERT, MARABERT, ARBERT, QARiB, AraELECTRA, and XLM-RoBERTa.

The experiments utilized different versions of Arabic pre-trained language models: AraBERTv01/v02, CAMeLBERT-MSA, MARBERT, MARBERTv2, ARBERT, and the QARiB base model, designed for the Arabic question answering task. Additionally, AraELECTRA

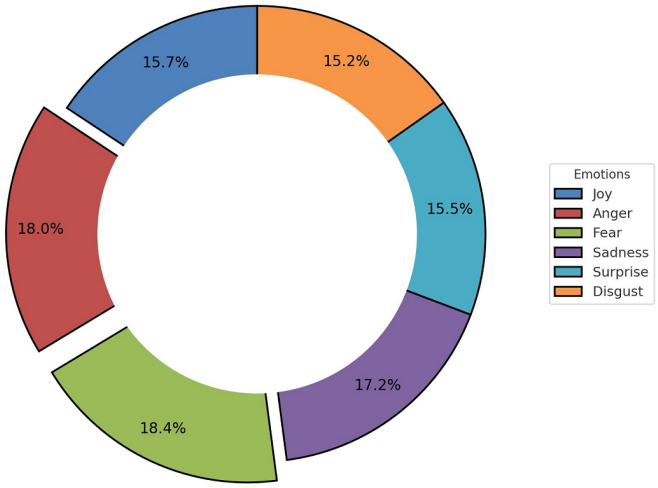


Figure 4.4: Emotion distribution within the AIER dataset.

was included with the 'araelectra-base-discriminator' version, recognized for its efficiency in pretraining. To evaluate the cross-lingual capabilities of Arabic pre-trained language models, XLM-RoBERTa, a robust multilingual model, was also tested.

Given the relatively balanced distribution across emotion categories, no balancing techniques were applied during model training. The dataset was randomly divided into training (70%), validation (15%), and testing (15%) subsets. The distribution of emotional labels across each subset is presented in Table 4.8.

Emotion	**Training**	**Validation**	**Testing**
Joy	569	130	127
Anger	679	143	123
Fear	671	151	146
Sadness	646	125	132
Surprise	577	123	128
Disgust	540	117	144

Table 4.8: Dataset Train/Validation/Test Split.

4.4.1.1 *Hyperparameter Settings*

Hyperparameters were carefully optimized to achieve the most accurate results. All pre-trained language models were fine-tuned using sequence lengths of 51, 64, and 128, with batch sizes of 8, 16, 32, and 64 for both training and evaluation. Learning rates of 2e-5, 3e-5,

and 5e-5 were also explored. Multiple configurations were systematically tested using the validation split, and the optimal values yielding the highest performance were selected.

The final hyperparameter settings were determined as follows: the maximum sequence length was set to 51, the learning rate to 2e-5, and the number of training epochs to 20, with an early stopping strategy employed to prevent overfitting and ensure robust performance. The models were optimized using the Adam optimizer and compiled with a categorical cross-entropy loss function. A summary of the best hyperparameter configuration is provided in Table 4.9.

Hyper-parameters	**Tested Values**	**Optimal Value**
Learning rate	2e-5, 3e-5, 5e-5	2e-5
Loss Function	Categorical Cross-entropy	Categorical Cross-entropy
Optimizer	Adam	Adam
Batch size	8, 16, 32, 64	16
Epochs	[5-20]	8
Max Length	51, 64,128	51

Table 4.9: Tested and optimal Hyperparameter values.

4.4.2 *Generative LLMs*

Recent advances in transformer-based architectures have led to the emergence of generative LLMs, such as ChatGPT, Gemini, and Llama, which have opened new avenues for text understanding and generation. These models demonstrate strong generalization capabilities that enable effective text comprehension for emotion recognition without the need for additional training or fine-tuning. In addition to the Arabic BERT-based models, five generative models were evaluated: ChatGPT-4 [94], Gemini 2.0 Pro [95], Llama 3 [97], Phi-4 [99], and Command R7B Arabic [100].

For experiments, a zero-shot prompting approach was adopted. The proprietary models, ChatGPT and Gemini, were accessed through their chatbot interfaces, with the GPT-4 Turbo version used for ChatGPT and the Gemini 2.0 Pro version used for Gemini. In contrast, the open-source models (Llama 3, Phi-4, and Command R7B) were accessed via the Ollama framework using their quantized versions: llama3:8b, phi4:14b, and command-r7b-arabic:7b, respectively. All experiments were performed in March 2025 to ensure that

the results reflect the model versions available at that time. To ensure a fair comparison with fine-tuned BERT-based models, all models were evaluated on the same test set (15% of the AIER dataset).

4.4.2.1 *Prompt Design*

Designing an appropriate prompt for a given task is crucial, as it directs the model to produce the intended output while ensuring better control over responses. During the experimental stage, several trials were performed to identify the prompts that achieved the most accurate and reliable results for IER in Arabic text.

The design process began with a simple and short instruction prompting the model to classify the emotion of a given text. However, this simple prompt frequently produced incomplete and inconsistent outputs. In several cases, the models generated extended explanations instead of a single label or provided multiple possible emotions, thereby reducing the clarity and reliability of the results.

To address these limitations, four fundamental components were incorporated into the prompt: role, instructions, context, and constraints. The design was iteratively refined by adding explicit task instructions, emphasizing the implicit nature of the emotion, defining the model's role as an expert in emotion recognition, and imposing strict output constraints. The final version required the model to select exactly one label from the predefined set of six emotions (Anger, Fear, Joy, Sadness, Disgust, Surprise), and to provide only the emotion without any additional explanation. These elements contributed to creating a structured and purposeful design that guided the models toward task-relevant responses.

The final prompt used in the evaluation was as follows:

> *You are an expert in emotion recognition.*
>
> *Your task is to detect the implicit emotion expressed in the given Arabic text.*
>
> *The text does not contain explicit emotion words; inference must rely on surrounding contextual cues.*
>
> *Select exactly one label from the following emotions: {Anger, Fear, Joy, Sadness, Disgust, Surprise}.*
>
> *Provide only the emotion label without any explanation.*

This prompt design contributed to maintaining consistency and fairness across all evaluated LLMs, reducing variability introduced by prompt formulation and allowing performance differences to be attributed to the models themselves.

4.4.3 *Classical ML models*

To establish a baseline IER system for further evaluation and comparisons across different systems, a comprehensive set of experiments was conducted, including classical ML models. The study employed the most widely used models to build an IER system, including SVM, KNN, XGBoost, RF, DT, LR, and MNB. During the data processing phase, text data was processed using the ISRI Arabic stemmer [146], with stop words and punctuation removed. No special handling for emotions was applied.

The hyperparameters for each algorithm were tuned through a manual search strategy informed by expert knowledge to determine the optimal combination of values that enhances the accuracy of the models and improves their performance. Table 4.10 outlines the selected hyperparameter configurations for each algorithm.

Model	**Hyperparameter**
SVM	dual = False, tol = 0.0001, C = 1
KNN	n_neighbors = 5, weights = 'distance', metric = 'cosine'
XGBoost	use_label_encoder = False, eval_metric = 'mlogloss'
RF	n_estimators = 100, max_depth = None, random_state = 27
DT	criterion = 'gini', max_depth = None, min_samples_split = 2, min_samples_leaf = 1, random_state = 27
LR	C = 10, solver = 'liblinear', max_iter = 1000
MNB	alpha = 1.0, fit_prior = True, class_prior = None

Table 4.10: Hyperparameter used for Classical ML models.

4.4.3.1 *Feature Extraction*

For feature extraction, two techniques were employed: BoW and TF-IDF, both combined with n-gram features. BoW, a simple approach, represents the text as a vector of word occurrences, ignoring grammar and word order. In contrast, TF-IDF considers the importance of words in the document relative to the entire corpus, thereby reducing the influence of

commonly used words. Both techniques were evaluated separately using different n-gram configurations, including unigrams, bigrams, trigrams, and their combinations. Based on the experimental results, the combination of unigrams, bigrams, and trigrams (unigram + bigram + trigram) was adopted, as it demonstrated superior performance.

4.4.4 *DL-based models*

To evaluate the performance of traditional DL-based models, we implemented five neural network architectures: CNN, LSTM, GRU, BiLSTM, and BiGRU. For all models, pre-trained FastText embeddings for Arabic [66] were used, with a vector dimensionality of 300.

The performance of DL algorithms depends heavily on careful hyperparameter tuning. However, as the number of hyperparameters increases and the parameter space becomes more complex, the computational cost becomes extremely high. In this study, the hyperparameter values for the classifiers were empirically determined based on manual experiments conducted using the developed dataset. During training, the models were trained for up to 20 epochs, with an early stopping strategy implemented to achieve the best results. The training process employed the categorical cross-entropy loss function and the Adam optimizer, initialized with a learning rate of 1e-3, which was found to be well-suited for this classification task.

In the CNN model, two convolutional layers were employed, with 32 and 64 filters, respectively. The kernel size was set to 2, and the ReLU activation function was applied to each layer. Subsequently, a MaxPooling layer was added to reduce the spatial dimensions. Finally, an output layer with six nodes and a softmax activation function was used to generate the predictive probability distribution for each of the six emotion classes.

Both LSTM and GRU models incorporate a recurrent layer, specifically, an LSTM layer in the LSTM model and a GRU layer in the GRU model, each configured with 128 hidden units. To mitigate overfitting, a Dropout layer with a rate of 0.3 is implemented following the recurrent layer. The output layer consists of six neurons and a softmax activation function, producing probability distributions for the six emotion classes.

The BiLSTM model features a 128 unit bidirectional LSTM layer and a 64 neuron Dense layer with ReLU activation to learn higher level features. The output layer consists of six neurons and a softmax activation function, producing probability distributions for the

six emotion classes. To avoid overfitting, the dropout method is applied to regularize the model with different dropout rates. The same architecture is used for the BiGRU model, but with a Bidirectional GRU layer instead of the Bidirectional LSTM layer.

4.4.5 *Evaluation*

In order to evaluate the performance of the selected models on the annotated dataset, accuracy, precision, recall, and F1-score were employed as the official evaluation metrics. These metrics are defined as shown in Section 2.4.5.

4.5 RESULTS AND DISCUSSION

This section reports the experimental results of the proposed models, aiming to evaluate the effectiveness of classical ML, DL, and transformer-based approaches for IER in Arabic text.

4.5.1 *Performance of ML classifiers*

Table 4.11 and Figure 4.5 provide an overview of the performance of classical ML classifiers on the AIER dataset with TF-IDF and BoW features.

Results revealed that SVM with BoW features achieved the highest F1-score (49.96%) among the ML approaches, surpassing MNB (48.91%), LR (48.51%), XGBoost (44.70%), RF (43.42%), DT (38.90%), and KNN (19.33%). On the other hand, ML models with TF-IDF features performed worse in terms of F1-score than BoW. The SVM classifier outperformed the remaining models by achieving the highest F1-score of 49.37%. In contrast, DT obtained the lowest F1-score, at only 28.75%, while the F1-scores of other classifiers (KNN, XGBoost, RF, LR, and MNB) ranged between 40.16% and 49.04%. In addition, SVM with BoW outperformed other ML models concerning the highest accuracy (49.43%), precision (50.67%), and recall (49.61%) measures.

Features	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
BoW	SVM	**49.43**	**50.67**	**49.61**	**49.96**
	KNN	25.14	40.32	23.03	19.33
	XGBoost	44.11	47.03	44.08	44.70
	RF	43.26	50.41	42.85	43.42
	DT	38.70	40.82	38.74	38.90
	LR	47.91	50.04	47.86	48.51
	MNB	48.38	49.55	49.08	48.91
TF-IDF	SVM	**48.86**	50.08	**48.97**	**49.37**
	KNN	42.12	42.63	41.83	41.54
	XGBoost	40.70	44.31	40.11	40.16
	RF	42.88	**50.96**	42.09	42.73
	DT	29.31	30.81	28.97	28.75
	LR	48.29	50.35	48.37	49.04
	MNB	48.19	50.38	48.74	48.92

Table 4.11: Performance comparison of classical ML-based classifiers on the AIER dataset.

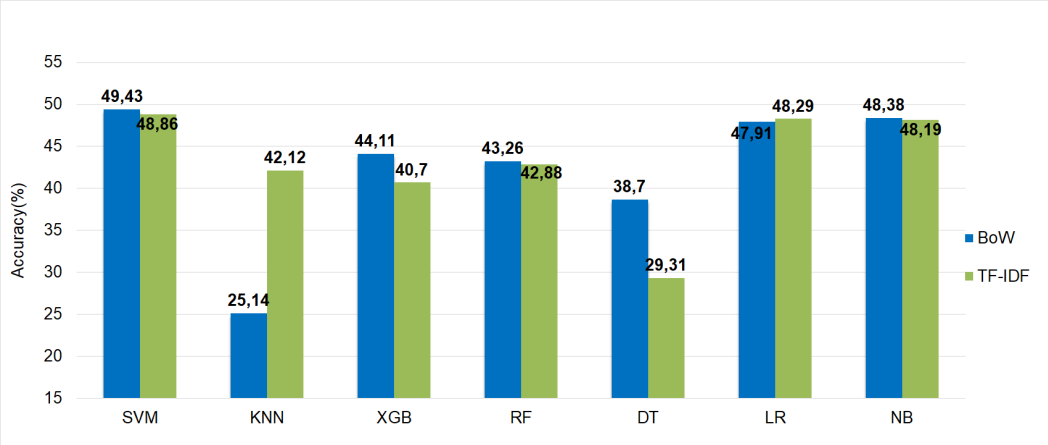


Figure 4.5: Classification results of classical ML-based models on AIER dataset, for BoW and TF-IDF feature extraction.

Regarding performance across individual emotion classes, as shown in Table 4.12, SVM and MNB using BoW features achieve the highest performance for the 'fear' class, with an F1-score of 63%, and demonstrate relatively consistent results for other emotions such as 'anger' and 'disgust'. The RF and XGBoost classifiers also show competitive performance in recognizing the fear and sadness emotions. LR, on the other hand, achieves results comparable to those of SVM and MNB, whereas DT performs slightly worse.

In summary, the performance of ML classifiers remains relatively modest, with the best results across all algorithms falling below 50% accuracy. This outcome is expected to some

Emotion	SVM	KNN	XGBoost	RF	DT	LR	MNB
Joy	0.41	0.08	0.37	0.38	0.30	0.36	**0.41**
Anger	**0.48**	0.32	0.45	0.44	0.41	**0.48**	**0.48**
Fear	**0.63**	0.23	0.61	0.59	0.50	0.61	**0.63**
Sadness	0.46	0.32	0.41	**0.59**	0.37	0.45	0.40
Surprise	**0.47**	0.14	0.38	0.34	0.32	0.46	**0.47**
Disgust	**0.54**	0.08	0.47	0.47	0.42	0.53	**0.54**

Table 4.12: Performance comparison of ML-based Approaches with BoW per emotion class.

extent, given that the dataset lacks explicit emotion-bearing words, which traditional feature representation methods and classifiers often rely on to classify emotions. These findings highlight the need for more advanced modeling techniques and richer word representations capable of capturing subtle semantic and contextual cues, thereby motivating the adoption of deep learning in subsequent experiments.

4.5.2 *Performance of DL classifiers*

When dealing with subtle linguistic nuances, DL algorithms also exhibit limitations in terms of accuracy. Table 4.13 and Figure 4.6 present the performance of various end-to-end DL models on the AIER dataset.

The CNN model demonstrated weak performance, achieving the lowest accuracy of 58.23%, followed by the GRU and LSTM models, which showed slightly better results with accuracies of 59.24% and 59.75%, respectively. The BiLSTM model emerged as one of the top performers, achieving an accuracy of 60.63%, a precision of 61.18%, a recall of 60.63%, and an F1-score of 60.60%. Finally, the BiGRU model outperformed the remaining DL-based models, obtaining the highest accuracy of 63.67%, which is approximately 14.24% higher than the best ML approach (i.e., SVM + BoW).

Table 4.14 summarizes the F1-score results across individual emotion classes. The BiGRU achieved the highest performance for the 'fear' class, with an F1-score of 76.92%, and demonstrated strong results for 'anger' (54.55%) and 'surprise' (65.43%). Similarly, the BiLSTM excels in recognizing 'disgust' and 'surprise' classes. The LSTM performed best for

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN	58.23	58.21	58.23	57.76
LSTM	59.75	61.35	59.75	60.34
GRU	59.24	60.61	59.24	58.70
BiLSTM	60.63	61.18	60.63	60.60
BiGRU	**63.67**	**64.36**	**63.67**	**63.57**

Table 4.13: Performance of DL-based classifiers on the AIER dataset.

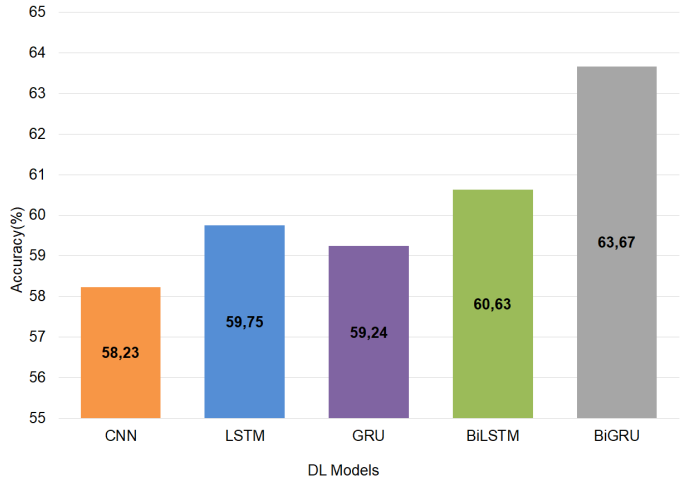


Figure 4.6: Classification results of DL-based models on AIER dataset.

'sadness' and 'joy', while the GRU and CNN yielded moderate performance overall, with the CNN performing slightly better for 'disgust' with an F1-score of 66.12%.

Emotion	CNN	LSTM	GRU	BiLSTM	BiGRU
Joy	50.81	54.05	51.08	53.79	**54.98**
Anger	50.61	48.46	45.41	48.91	**57.02**
Fear	69.80	70.83	70.07	70.99	**75.32**
Sadness	49.01	**59.48**	57.43	55.37	55.85
Surprise	60.54	62.15	62.57	**66.67**	65.38
Disgust	66.12	66.08	64.57	67.68	**73.24**

Table 4.14: Performance comparison of DL-based approaches per emotion class.

Overall, the bidirectional models (BiLSTM and BiGRU) generally outperform their unidirectional counterparts, indicating that modeling contextual information from both preceding and following tokens is particularly beneficial for IER. This improvement can be attributed to the ability of bidirectional architectures to capture richer contextual dependencies and latent emotional cues that are not explicitly expressed in text. However, although DL

models yield noticeable performance gains over traditional ML approaches, their results still indicate limitations in fully capturing complex and implicit emotional expressions. These limitations are further exacerbated by the morphological richness and dialectal variation of Arabic, as well as by the absence of explicit emotion-bearing words in the dataset.

4.5.3 *Performance of BERT-based Models*

Table 4.15 and Figure 4.7 summarize the results achieved by fine-tuned BERT-based models. The fine-tuned Arabic pre-trained language models significantly outperformed both traditional ML and DL algorithms, achieving scores exceeding 70% across all metrics on the AIER dataset. XLM-RoBERTa acquired the lowest result among all models, at only 70.99%. The AraELECTRA, on the other hand, has a nearly 2.19% higher F1-score (73.18%) than XLM-RoBERTa. AraBERTv02, which was only developed with a larger corpus size compared to AraBERTv01, outperformed the aforementioned models with an F1-score of 77.48%. QARiB and MARBERT also demonstrated high performance, achieving F1-scores of 78.52% and 77.25%, respectively. CAMeLBERT and ARBERT showed competitive results, though slightly lower than MARBERT. Among all models, the MARBERTv2 obtained the highest performance, with an F1-score of 79.83% and an accuracy of 79.62%.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
XLM-RoBERTa	71.14	70.94	71.14	70.99
AraELECTRA	73.04	73.42	73.21	73.18
AraBERTv01	74.18	74.53	74.42	74.38
AraBERTv02	77.34	77.66	77.38	77.48
QARiB	78.23	78.83	78.38	78.52
CAMeLBERT	75.82	75.89	75.89	75.84
ARBERT	75.57	76.36	75.50	75.83
MARBERT	77.22	77.46	77.26	77.25
MARBERTv2	**79.62**	**80.06**	**79.71**	**79.83**

Table 4.15: performance comparison of different transformer-based models, on the AIER dataset.

Table 4.16 presents the performance of BERT-based models across different emotion classes. The MARBERTv2 model consistently outperformed the other models, achieving

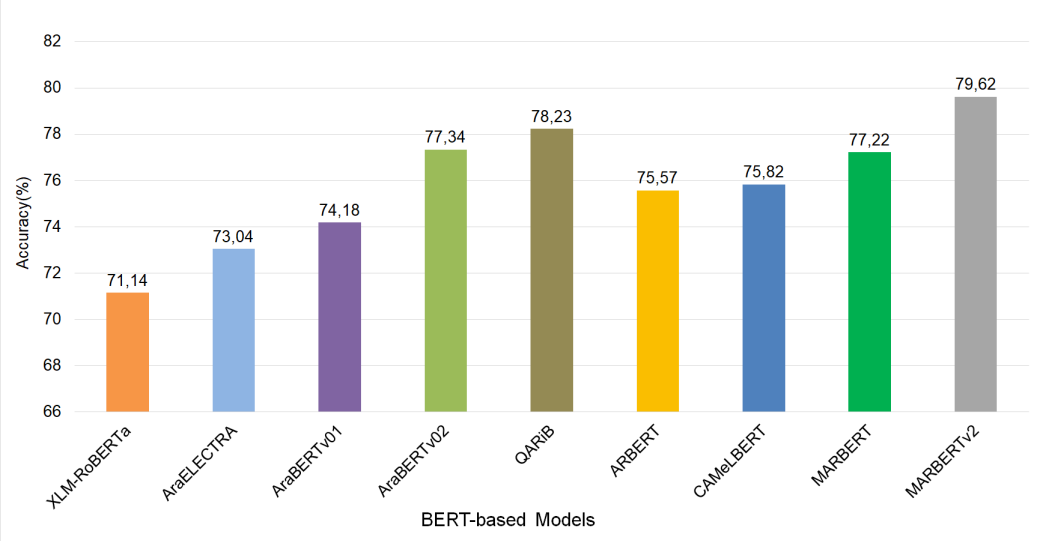


Figure 4.7: Classification results of implicit emotion using Fine-tuned pre-trained models.

the highest F1-scores in most classes, including 'joy' (75.69%), 'anger' (88.98%), 'fear' (88.07%), 'sadness' (70.68%), and 'surprise' (86.81%). However, other models, such as MARBERT, CAMeLBERT, and QARiB, also demonstrated strong performance, particularly in classifying the 'anger' and 'fear' classes. In contrast, XLM-RoBERTa and AraELECTRA exhibit relatively lower performance in classes such as 'disgust' and 'joy,' highlighting the varying capabilities of these architectures in capturing subtle emotional expressions.

Model	Joy	Anger	Fear	Sadness	Surprise	Disgust
XLM-RoBERTa	65.96	85.25	75.00	60.67	79.73	60.00
AraELECTRA	63.57	82.45	80.99	65.96	85.11	60.98
AraBERTv01	67.65	87.76	80.48	64.00	83.33	63.04
AraBERTv02	72.28	88.21	81.27	69.12	86.01	68.00
QARiB	72.03	87.45	86.07	67.66	86.41	**71.48**
CAMeLBERT	73.10	88.51	81.27	62.55	83.45	66.67
ARBERT	72.43	86.34	80.83	66.67	83.62	65.06
MARBERT	72.00	**88.98**	85.26	68.18	83.67	66.15
MARBERTv2	**75.69**	**88.98**	**88.07**	**70.68**	**86.81**	68.73

Table 4.16: Performance comparison of BERT-based models per emotion class.

Overall, these findings demonstrate the strong effectiveness of BERT-based models pre-trained on large-scale Arabic corpora in capturing nuanced linguistic patterns and implicit emotional cues. The superior performance of MARBERTv2 and other Arabic-specific pre-trained models can be attributed to their exposure to dialectal and informal Arabic during pre-training, which enables more robust contextual representations. These results under-

score the critical role of domain- and language-specific pre-training in achieving SOTA performance in Arabic NLP tasks. Moreover, the observed improvements further confirm that transformer-based architectures provide substantial advantages over traditional DL models for IER, motivating their adoption as the core modeling framework in this thesis.

4.5.4 *Performance of Generative LLMs*

Table 4.17 and Figure 4.8 summarize the results of generative models on the test data. We show a significant decrease in all scores regarding the generative LLMs when compared to the BERT-based models. Gemini 2.0 Pro achieved the best performance, with an accuracy of 66.33%, a precision of 67.04%, a recall of 66.33%, and an F1-score of 66.05%. ChatGPT-4 demonstrated comparable performance, achieving an accuracy of 64.17% and an F1-score of 64.50%. Phi-4 showed moderate performance with an accuracy of 53.67%, while Llama 3 performed poorly, achieving an accuracy of 44.83%. Although Command R7B Arabic was designed specifically for Arabic, it only marginally outperformed Llama 3, with an accuracy of 47.33%.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
ChatGPT-4	64.17	65.98	64.17	64.50
Gemini 2.0 Pro	**66.33**	**67.04**	**66.33**	**66.05**
Llama 3	44.83	49.79	44.83	43.17
Phi-4	53.67	54.83	53.67	53.03
Command R7B Arabic	47.33	51.91	47.33	47.29

Table 4.17: Performance comparison of different generative LLMs, on the AIER dataset.

Table 4.18 presents the F1-scores achieved by generative LLMs across different emotion classes. ChatGPT-4 achieved the highest scores in the 'joy' (85.86%), 'sadness' (59.45%), and 'surprise' (58.29%) classes, while Gemini 2.0 Pro led in 'anger' (54.00%), 'fear' (73.53%), and 'disgust' (67.57%). In contrast, Llama 3 and Phi-4 showed significantly lower performance, particularly in 'fear' and 'surprise' emotions, indicating difficulty in identifying less explicit emotional expressions. The Command R7B Arabic model performed better in detecting 'joy' and 'disgust' but struggled with 'fear' and 'surprise'.

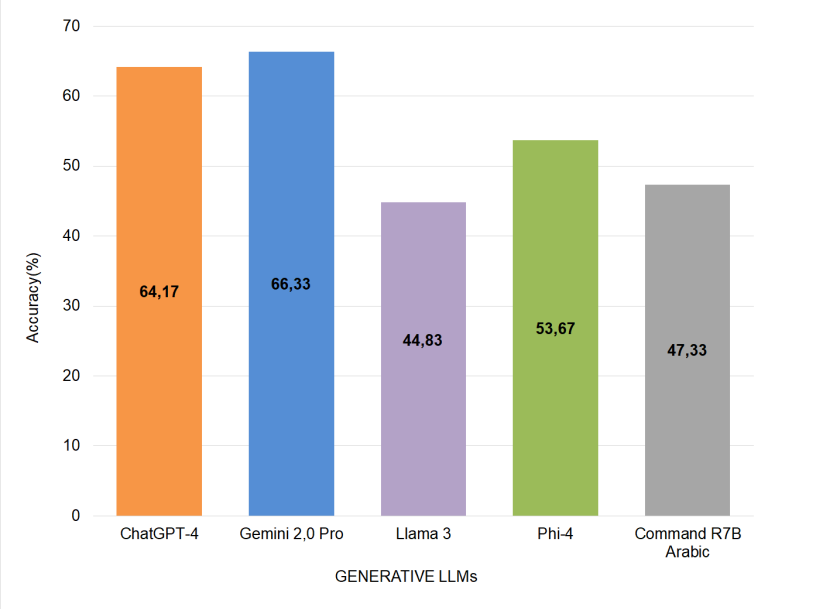


Figure 4.8: Classification results of generative LLMs on AIER dataset.

Emotion	ChatGPT-4	Gemini 2.0 Pro	Llama 3	Phi-4	Command R7B Arabic
Joy	**85.86**	85.57	67.98	80.20	77.55
Anger	53.46	**54.00**	35.42	42.00	40.21
Fear	66.27	**73.53**	22.76	41.29	29.58
Sadness	**59.45**	59.15	56.06	54.87	45.25
Surprise	**58.29**	56.44	36.02	47.13	34.83
Disgust	63.59	**67.57**	40.79	52.67	56.32

Table 4.18: Performance comparison of generative LLMs per emotion class.

In summary, the experimental results clearly highlight the performance disparity among the four classes of methods. As shown in Figure 4.9, classical ML algorithms performed worse than both traditional DL and transformer-based models. Specifically, the SVM + BoW approach achieved accuracy scores that were 12%, 16%, and 30% lower than DL, generative LLMs, and BERT-based models, respectively. BERT-based architectures demonstrated significant improvements in contextual understanding, outperforming DL-based algorithms and generative LLMs by 17% and 13%, respectively.

The strength of BERT-based models stems from their ability to learn from the surrounding context and capture complex contextual information through a bidirectional approach that considers words in both forward and backward contexts. Additionally, the BERT architecture leverages transfer learning, being pre-trained on a large amount of Arabic text data

to learn general language representations, which are then fine-tuned for the IER task using the developed dataset. Furthermore, the attention mechanism significantly improves the model's ability to disambiguate and capture implicit cues within complex morphological structures. Another characteristic that contributes to the effectiveness of BERT-based models is their masking strategy used during training. By randomly masking a percentage of input tokens, these models develop a deep understanding of word relationships, capturing subtle linguistic patterns and implicit emotional cues expressed through context.

For generative models, their performance was worse compared to Arabic pretrained models, which was, to some extent, anticipated, as BERT-based models underwent a fine-tuning process with the AIER dataset, whereas the generative models were not optimized for the data. As a result, their predictions relied solely on pre-trained knowledge, without any task-specific adaptation.

The differences between the two transformer architectures could also explain the observed disparity. By focusing on predicting the next word based only on past context using unidirectional training, generative models lack the ability to handle tasks like IER that require a more comprehensive understanding of both the preceding and succeeding contexts. In addition, this low performance may be attributed to limitations in prompt design. Despite the LLMs' general language understanding capabilities, generic prompts often fail to effectively guide the model to recognize these nuanced signals. We found that prompt engineering had a significant impact on performance; more structured and context-aware prompts slightly improved results but still did not match fine-tuned approaches.

4.5.5 *Cost-effectiveness Analysis*

To evaluate the cost-effectiveness of different method categories, their performance was analyzed based on three key metrics: training time, inference time, and model size (measured by the number of parameters required). All models were evaluated on the Google Colab platform configured with 12 GB of RAM and a Tesla T4 GPU. The implementation framework, employing TensorFlow, Keras, and Scikit-learn, calculated model accuracy, training and inference times, and model size for each method. Table 4.19 summarizes runtime and model size results.

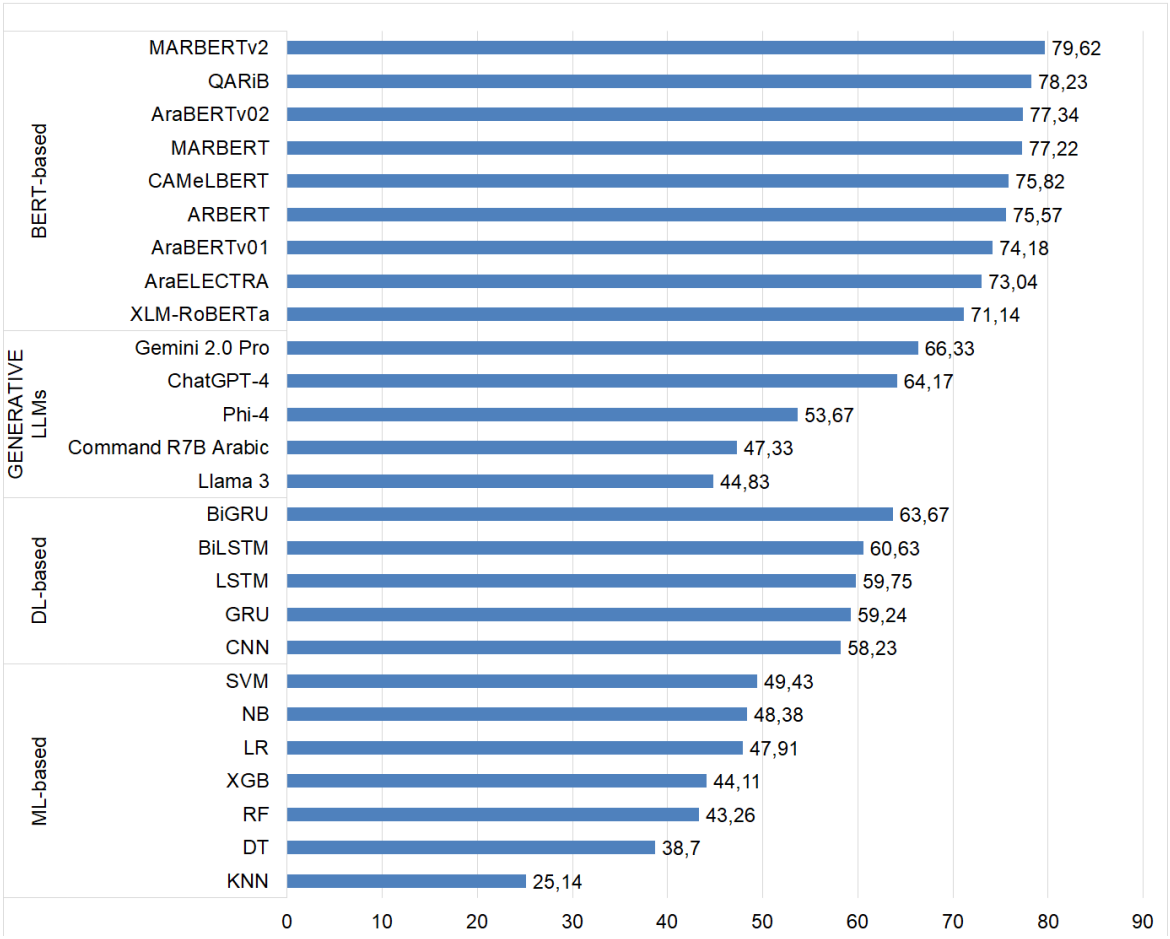


Figure 4.9: Performance comparison of different classes of classifiers.

Results showed that classical ML approaches have the lowest computational cost among the evaluated methods, with training times ranging from 0.002 seconds (KNN) to 8.326 seconds (RF). These models delivered near-instant predictions (0.001 to 0.246 seconds), as their inference complexity depended solely on the feature count. This efficiency was attributed to their simple architectures and compact model sizes (4,231 to 441,210 parameters).

In contrast, DL models exhibit moderate training times, ranging from 31.86 seconds (CNN) to 59.96 seconds (BiGRU). Recurrent architectures were slower than convolutional ones due to their sequential processing, and bidirectional variants further increased latency by doubling the computational pathways. Despite this, their inference times remain efficient (0.21–0.69 seconds), with CNN being the fastest. The model sizes (5M parameters) are significantly larger than classical ML.

The fine-tuned BERT-based approaches exhibited significantly higher computational demands compared to other methods, with execution times ranging from 275 seconds (CAMeLBERT, 109M parameters) to 391 seconds (XLM-RoBERTa, 279M parameters).

Generative LLMs, despite their flexibility, are hampered by substantial inference time and memory usage issues, particularly due to their massive parameter counts (7B to 1.8T). Their inference time depends on the method used (web interface or local deployment via Ollama) and the policies of the provider, which limit the number of allowed requests.

Concerning the performance-cost balance, the four types of algorithms yielded different trade-offs between accuracy and computational cost, depending on their computational complexity. Classical ML models showed lower accuracy but remain computationally efficient, requiring minimal resources while delivering rapid inference times. Their computational complexity increases linearly with the input size, ranging from $O(n)$ to $O(n^3)$ depending on the algorithm. This efficiency makes them suitable for real-time applications but less effective for capturing implicit emotions. In contrast, DL models demonstrated moderate accuracy with a corresponding increase in the computational cost. Their computational complexity rises due to recurrent structures and matrix multiplications, resulting in slower inference compared to ML models. BERT-based models further enhance accuracy due to their superior contextual understanding. However, they require higher computational costs and longer inference times than traditional DL models. The quadratic scaling of self-attention mechanisms with sequence length introduced inherent computational complexity, making 'BERT-based models resource-intensive yet highly effective. Finally, generative LLMs provide variable accuracy in zero-shot settings. However, they entail the highest computational cost and the slowest inference time, as they generate text sequentially with significantly large parameter sizes.

4.6 CONCLUSIONS

In this chapter, we established a structured foundation for IER in Arabic through the construction and analysis of a novel AIER dataset and the evaluation of different baseline models. By introducing the AIER dataset, annotated with Ekman's six basic emotions as well as emotional cues and causes, this work addresses a critical gap in Arabic affective com-

Approach	Models	Training Time (s)	Inference Time (s)	Size (Params)
ML-based	SVM	0.142	0.001	69.4K
	KNN	0.002	0.246	-
	XGBoost	6.952	0.019	12.7K
	RF	8.326	0.149	441K
	DT	2.617	0.002	4.2K
	LR	2.179	0.001	69.4K
	MNB	0.005	0.008	416K
DL-based	CNN	31.86	0.21	5.6M
	LSTM	42.08	0.37	5.4M
	GRU	50.48	0.42	5.3M
	BiLSTM	52.67	0.66	5.6M
	BiGRU	59.96	0.69	5.5M
BERT-based	XLM-RoBERTa	391.76	2.65	279M
	AraELECTRA	312.41	2.86	135M
	AraBERTv01	315.69	2.80	135M
	AraBERTv02	325.77	3.10	135M
	QARiB	283.10	2.44	135M
	CAMeLBERT	275.74	2.33	109M
	ARBERT	321.21	2.42	163M
	MARBERT	326.85	2.48	163M
	MARBERTv2	314.81	2.57	163M
Generative LLMs	ChatGPT-4	/	923	1.8T
	Gemini 2.0 pro	/	202	-
	Llama 3	/	1208	8B
	Phi-4	/	2674	14B
	Command R7B Arabic	/	2311	7B

Table 4.19: Cost-effectiveness comparison

puting. This multi-level annotation framework enables fine-grained analysis of emotional expression and constitutes a valuable resource for a wide range of Arabic classification tasks. Additionally, the proposed masking strategy further enables the simulation of real implicit emotion scenarios, allowing for a realistic and challenging evaluation setting.

Through a comprehensive experimental study, we evaluated classical ML models, DL architectures, transformer-based models, and generative LLMs. The results highlighted the limitations of traditional approaches when dealing with implicit emotional expressions and demonstrated the clear power of pretrained BERT-based models in capturing subtle

contextual cues. At the same time, the findings revealed persistent challenges related to implicit emotion modeling in Arabic, underscoring the need for more advanced architectures and hybrid modeling strategies.

The present study focused on MSA to establish a controlled experimental setting and reduce linguistic variability during the dataset construction and model development stages. Starting with MSA provided a rigorous assessment of the models' behavior in a relatively formal and linguistically stable environment.

Building on the insights and limitations identified in this chapter, the next phase of our research extends this work by proposing a hybrid architecture that integrates traditional deep learning architectures with contextualized representations derived from pretrained language models. The objective is to investigate whether refined hybrid architectures can further enhance IER in Arabic, particularly when dealing with the highly informal and noisy linguistic characteristics of Arabic tweets.

5

A TRANSFORMER-BASED HYBRID MODEL FOR IMPLICIT EMOTION RECOGNITION IN ARABIC TWEETS

5.1 INTRODUCTION

This chapter introduces AraBERT-BiGRU, a hybrid contextual–sequential architecture developed to address Implicit Emotion Recognition (IER) in Arabic social media text. The proposed framework integrates contextualized representations derived from PLMs with sequential deep learning architectures to effectively capture the linguistic variation, semantic subtleties, and contextual complexity characteristic of Arabic tweets. The chapter provides a comprehensive experimental evaluation of the proposed model under both implicit and explicit emotion recognition settings, enabling a thorough assessment of its robustness and effectiveness.

Emotional expression in real-world communication rarely follows standardized linguistic norms, particularly on social media platforms such as Twitter (now **X**). While the platform enables users from diverse backgrounds to share opinions and experiences, recognizing emotions in tweets remains highly challenging due to their characteristics and the complexity of emotional expression [52]. Tweets are short, informal, and often contain noisy linguistic elements such as dialectal variations, abbreviations, slang, misspellings, emojis, and hashtags, which complicate semantic understanding [53]. Moreover, emotions are not always expressed explicitly through emotion words; instead, they are often conveyed implicitly through figurative language such as metaphors, idiomatic expressions, similes, and contextual cues that require deeper interpretation [20]. Compared to MSA, tweets in-

troduce additional layers of lexical, morphological, and structural complexity, highlighting the need for context-aware and semantically rich models capable of effectively capturing implicit emotional cues.

To address these challenges, we propose a unified contextual–sequential framework that integrates transformer-based contextual encoding with recurrent sequential modeling. Specifically, the proposed architecture combines contextualized representations derived from the AraBERT model with a BiGRU neural network to further model sequential dependencies and refine contextual representations.

In this chapter, we detail the design, implementation, and evaluation of our proposed hybrid model. We describe the model architecture, including how contextualized representations are integrated with sequential modeling. To evaluate the effectiveness of the model, we conducted extensive experiments on multiple publicly available Arabic emotion datasets under both implicit and explicit recognition settings.

CONTRIBUTIONS: Our contributions in this chapter are as follows:

1. We introduce a hybrid contextual–sequential architecture that integrates AraBERT's deep bidirectional representations with the temporal modeling capabilities of BiGRU, enhancing the modeling of implicit emotions in Arabic tweets.

2. We provide an empirical evaluation of the proposed hybrid model on a newly compiled Arabic Twitter corpus, highlighting its effectiveness in capturing context-dependent and subtly expressed emotions.

3. We conduct a comparative analysis against traditional deep learning baselines, offering quantitative and qualitative insights into the benefits of combining transformer-based contextual embeddings with recurrent sequence modeling.

DECLARATION: This chapter is based on the following journal publication:

- H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, "A transformer-based hybrid model for implicit emotion recognition in arabic text", Engineering, Technology & Applied Science Research, vol. 15, no. 3, pp. 23834–23839, 2025. DOI: 10.48084/etasr.10261. [Online]. Available: https://etasr.com/index.php/ETASR/article/view/10261. [147].

The chapter is presented largely in alignment with the published manuscript. Referencing, notation, and structural organization have been adapted for consistency across the thesis, with minor stylistic and formatting changes introduced where necessary.

5.2 METHODOLOGY

In this section, we present the methodologies and models we used. We begin with the introduction of the baseline models, followed by a description of our novel hybrid framework. A detailed explanation of the architectures is provided, along with the hyperparameters and training procedures adopted to ensure consistent and robust evaluation across all experiments.

5.2.1 *Baselines*

To ensure a comprehensive comparison, we implemented several baseline models. These baselines include traditional deep learning models (CNN, LSTM, BiLSTM, GRU, and BiGRU) and BERT-based architectures (fine-tuned AraBERT). All models were evaluated following the same experimental settings to ensure a fair comparison with the proposed AraBERT–BiGRU model. All models were implemented following the same setup described in Section 4.4.4.

5.2.2 *Proposed Model*

In this section, we present the proposed hybrid architecture developed for IER. Given the linguistic variability and contextual intricacy characteristic of Arabic tweets, the model is designed to effectively capture both deep semantic representations and sequential emotional dependencies. Specifically, we introduce a hybrid framework that leverages the strengths of both transformer-based contextual encoding and recurrent sequential modeling, forming a unified contextual–sequential architecture. Figure 5.1 illustrates the architecture of our proposed model.

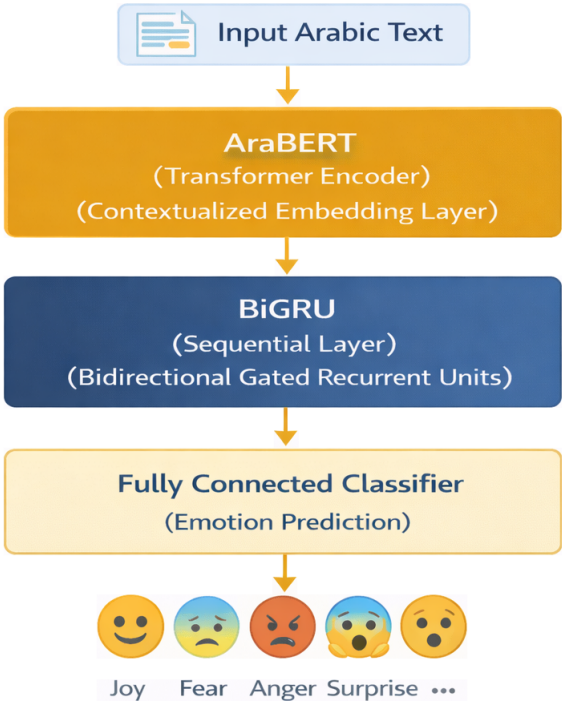


Figure 5.1: Architecture of the proposed Hybrid Model

The model combines a pretrained transformer encoder, AraBERT, with a BiGRU layer for sequential refinement. While AraBERT provides rich bidirectional contextual embeddings, the BiGRU layer further refines these representations by modeling token-level sequential dynamics that are particularly relevant for implicit emotional cues. Below, we describe the key components and their roles in the architecture.

AraBERT Pre-trained Language Model

AraBERT, pre-trained on diverse Arabic corpora, serves as the contextual encoder of the architecture. Unlike traditional word embedding methods, such as Word2Vec and GloVe, which produce static representations regardless of context, AraBERT generates contextualized embeddings that reflect a word's meaning based on its surrounding context, allowing the model to disambiguate lexical meaning and capture subtle semantic variations.

In addition to word embeddings, AraBERT is also well suited for generating sentence embeddings. Given an input sentence, AraBERT tokenizes the input text into a sequence of subword tokens. A special classification token [CLS] is added to the sequence, following the standard transformer-based encoding scheme. The input is then processed through multiple self-attention layers that compute bidirectional contextual representa-

tions. Rather than relying solely on static word-level embeddings, the hidden states generated by AraBERT encode rich syntactic and semantic dependencies across the entire sequence. In our architecture, the contextual token representations from the final transformer layer are forwarded to the sequential modeling component, allowing the BiGRU to operate over already contextualized embeddings.

Bidirectional Gated Recurrent Units (BiGRU)

The embedding representation produced by the AraBERT model is used as input for the BiGRU. The BiGRU effectively captures bidirectional contextual information by employing two independent GRUs: a forward GRU, which processes the sequence in its original order, and a backward GRU, which processes the sequence in reverse. Both GRUs generate forward and backward hidden states from the input word vectors. These states are then combined to form a comprehensive representation of each word within its full context, enabling the model to effectively capture long-range dependencies. Since emotional features are extracted through both forward and backward GRU networks, the BiGRU can better understand the global context of the text, which is essential for recognizing implicit emotions. We chose BiGRU over BiLSTM due to its simpler design and lower computational cost. This allows us to maintain high accuracy while improving efficiency.

Classification Layers

The refined sequence representation is subsequently passed to the classification module. First, a fully connected dense layer transforms the BiGRU output into a task-specific feature space, enabling the model to learn discriminative patterns associated with emotion classes. A dropout regularization mechanism is applied to reduce overfitting and enhance generalization.

Finally, a second dense layer produces the output logits corresponding to the predefined emotion categories. A softmax activation function converts these logits into a normalized probability distribution over the classes. The class with the highest probability is selected as the predicted label for the input sequence.

5.2.3 *Hyperparameter Settings*

Hyperparameter optimization was conducted using a grid search strategy on the validation set. The configuration yielding the highest F1-score was selected as the final setting. Table 5.1 presents the optimal hyperparameter values.

Hyper-parameters	**Value**
Learning Rate	2×10^{-5}
Loss Function	Categorical Cross-Entropy
Optimizer	Adam
Batch Size	32
Dropout Rate	0.25
GRU Units	256
Epochs	8

Table 5.1: The hyperparameters of our model

The architecture was implemented using the TensorFlow deep learning framework. Model optimization was conducted using a learning rate of 2×10^{-5}, which is consistent with commonly adopted fine-tuning settings for transformer-based language models. A dropout rate of 0.25 was applied to mitigate overfitting, and training was performed for a maximum of 8 epochs. An early stopping mechanism, monitored on validation performance, was employed to prevent over-training and ensure model generalization.

5.3 EXPERIMENTS

5.3.1 *Dataset*

Experiments were conducted using two publicly available Arabic emotion datasets: the SemEval-2018 Task 1 dataset [33] and the Arabic Emotion Tweets Dataset (AETD) [107].

- **SemEval-2018 Task 1 (Arabic Subtasks)**[1]: We use the Arabic Emotion Classification (Ar-Ec) and the Emotion Intensity Ordinal Classification (EI-oc) subsets. For the Ar-Ec only tweets annotated with a single emotion were retained. In the EI-oc subset,

1 https://competitions.codalab.org/competitions/17751

we focused exclusively on detecting emotions rather than predicting their intensity. Furthermore, tweets labeled as containing "no emotion" were excluded to ensure a clear focus on emotionally expressive content.

- **AETD Dataset**[2]: This dataset consists of 10,000 manually annotated tweets categorized into eight emotion classes: sympathy, joy, love, anger, sadness, fear, surprise, and none. The dataset provides balanced emotional coverage and reflects natural emotional expression in Arabic social media.

5.3.1.1 *Preprocessing and Filtering*

All tweets from both datasets were preprocessed using the most common preprocessing techniques to reduce noise and enhance linguistic consistency. The cleaning procedure included the removal of non-Arabic characters, URLs, mentions, retweets, numbers, repeated characters, punctuation marks, stop words, and Arabic diacritics. In addition, orthographic normalization was applied to unify variant letter forms and mitigate inconsistencies common in social media text. Emojis and emoticons were replaced with their corresponding textual emotion descriptors to preserve their affective information within the textual representation.

To ensure that emotional inference would depend primarily on contextual and semantic signals rather than explicit lexical indicators, all tweets containing explicit and direct emotion words were identified using the Arabic NRC Emotion Lexicon and then removed.

5.3.1.2 *Arabic Implicit Emotion Dataset (AIEmoD)*

By combining the previously described datasets, we constructed the Arabic Implicit Emotion Dataset (AIEmoD), a curated corpus comprising 7,684 Arabic tweets. This dataset was specifically designed to address the limited availability of high-quality annotated resources for IER in Arabic. AIEmoD focuses exclusively on implicitly expressed emotional content, providing a more challenging and realistic benchmark compared to traditional explicit emotion datasets. The distribution of emotion categories within AIEmoD is illustrated in Figure 5.2.

2 https://github.com/amrmalkhatib/Emotional-Tone

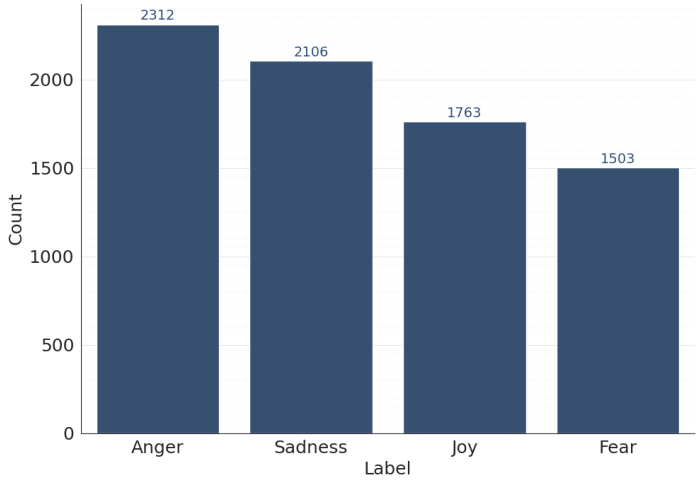


Figure 5.2: Emotions distribution within the AIEmoD dataset.

5.3.2 *Experimental Setup and Evaluation*

For the experimental evaluation, the AraBERTv0.2-Twitter variant of AraBERT was employed as the pretrained model. This configuration comprises 12 transformer layers, 768 hidden units per layer, 12 self-attention heads, and approximately 136 million trainable parameters. The Twitter-based version was specifically selected due to its pretraining on large-scale Arabic social media corpora, making it particularly suitable for modeling the linguistic characteristics of tweets.

For cross-dataset consistency, the study focused exclusively on the four shared emotion categories: anger, fear, joy, and sadness. After preprocessing and filtering, the resulting datasets were partitioned into training (70%), validation (15%), and testing (15%) subsets.

For the evaluation, we report precision, recall, and F1-score computed for each emotion class, along with their macro-averaged values to enable an overall comparison across models.

5.4 RESULTS AND DISCUSSION

This section outlines the results obtained. The model's performance is first compared with baselines. Subsequently, the results are examined in terms of explicit and implicit emotion recognition, as well as per-emotion performance. Finally, an error analysis is conducted, and the findings are compared with state-of-the-art approaches.

5.4.1 *Baseline Results*

Table 5.2 presents a comprehensive evaluation of our baselines with the proposed AraBERT–BiGRU across the SemEval-2018, AETD, and AIEmoD datasets.

Regarding the baselines, the comparative results demonstrate a clear performance gap between traditional deep learning baselines and the fine-tuned AraBERT model. Across all datasets, CNN, LSTM, GRU, and their bidirectional variants consistently show lower performance than AraBERT, underscoring the importance of large-scale pretrained contextual representations with deep bidirectional encoding and the attention mechanism integrated in AraBERT for detecting implicit emotion from Arabic tweets.

5.4.2 *Proposed Model vs. Baseline Results*

The proposed AraBERT–BiGRU model demonstrates consistent and robust performance across all datasets. On AIEmoD, it achieves an F1-score of 70.69%, with a recall of 70.83% and a precision of 70.93%. When compared to the strongest baseline, the proposed approach yields notable improvements across all evaluation metrics and datasets. Specifically, while AraBERT attains F1-scores of 54.25% on SemEval-2018, 77.07% on AETD, and 70.02% on AIEmoD, the AraBERT–BiGRU model surpasses these results, achieving 56.02%, 79.87%, and 70.67%, respectively.

Model	**SemEval-2018**			**AETD**			**AIEmoD**		
	Prec. (%)	**Rec. (%)**	**F1 (%)**	**Prec. (%)**	**Rec. (%)**	**F1 (%)**	**Prec. (%)**	**Rec. (%)**	**F1 (%)**
CNN	41.35	42.05	41.18	63.40	61.28	60.31	50.68	55.28	51.99
LSTM	46.40	49.24	45.92	65.96	63.56	63.51	58.07	58.66	56.53
BiLSTM	51.00	51.20	48.14	63.95	61.81	60.83	57.73	59.44	57.64
GRU	43.23	43.36	41.61	67.03	65.16	65.52	59.01	59.83	58.99
BiGRU	46.77	47.28	46.57	67.13	66.47	66.69	60.08	60.12	58.56
AraBERT	55.64	54.35	54.25	76.99	77.18	77.07	69.96	70.25	70.02
AraBERT-BiGRU	**56.88**	**56.09**	**56.02**	**79.97**	**80.14**	**79.87**	**70.93**	**70.83**	**70.67**

Table 5.2: Comparative results of the proposed model against baseline models on the implicit SemEval-2018, AETD, and AIEmoD datasets.

Notably, the proposed model achieves an improvement of approximately 8–10% in F1-score compared to neural network models, particularly on the SemEval-2018 and AIEmoD datasets. The observed performance is attributed to the complementary integration of

contextualized transformer-based representations with bidirectional sequential modeling. While AraBERT generates dynamic, context-sensitive representations that adapt to the surrounding linguistic structure, it enhances the quality of the learned embeddings. The subsequent BiGRU layer further refines these representations by effectively modeling long-range bidirectional dependencies within the text.

These findings collectively validate the effectiveness of the proposed hybrid architecture for IER, highlighting its ability to capture nuanced semantic and contextual information essential for this challenging task.

5.4.3 *Implicit vs. Explicit Affective Cues*

Table 5.3 compares the performance of the proposed AraBERT–BiGRU model in explicit and implicit emotion recognition across the SemEval-2018, AETD, and AIEmoD datasets. As expected, the model consistently achieves higher F1-scores in explicit emotion, reaching 62.14%, 85.96%, and 76.34%, respectively. This improvement is due to the presence of explicit emotional cues that facilitate the classification process. In contrast, implicit emotion, where such cues are absent, results in lower scores (56.02%, 79.87%, and 70.67%). Nevertheless, the relatively moderate performance gap indicates that the proposed hybrid architecture effectively captures contextual information and latent emotional cues even in the absence of explicit emotional markers.

Model	**SemEval-2018**		**AETD**		**AIEmoD**	
	Implicit	**Explicit**	**Implicit**	**Explicit**	**Implicit**	**Explicit**
AraBERT-BiGRU	56.02	62.14	79.87	85.96	70.67	76.34

Table 5.3: Explicit vs. Implicit performance of the proposed model on the SemEval-2018, AETD, and AIEmoD datasets

5.4.4 *Per-emotion Performance*

Table 5.4 provides a detailed per-class evaluation of the proposed model across the three datasets.

A clear pattern emerges in which *Joy* achieves the highest and most stable performance across all datasets, with F1-scores of 80.87% (SemEval-2018), 85.23% (AETD), and 88.35% (AIEmoD). This robustness suggests that joyful expressions, even when implicit, tend to exhibit relatively distinctive contextual signatures that are effectively captured by the hybrid architecture. In contrast, *Fear* emerged as the most challenging category, with comparatively low F1-scores across datasets, reflecting its semantic ambiguity and frequent overlap with the sadness emotion. *Anger* demonstrates substantial variability, achieving particularly strong results on the AETD dataset (84.38%), while showing moderate performance on SemEval-2018 and AIEmoD, indicating sensitivity to dataset composition and contextual distribution. *Sadness* presents intermediate performance, with noticeable improvement in AETD and AIEmoD compared to SemEval-2018.

Overall, the results confirm that while the proposed architecture effectively captures positive emotions, performance differences across categories reflect the complexity and contextual subtlety of implicitly expressed affect in Arabic tweets.

Emotion	**SemEval-2018**			**AETD**			**AIEmoD**		
	Prec. (%)	**Rec. (%)**	**F1 (%)**	**Prec. (%)**	**Rec. (%)**	**F1 (%)**	**Prec. (%)**	**Rec. (%)**	**F1 (%)**
Anger	49.39	61.83	54.92	82.89	85.91	84.38	65.84	76.44	70.74
Fear	43.53	46.84	45.12	55.17	53.33	54.24	47.31	47.31	43.35
Sadness	54.10	42.31	47.48	78.95	67.31	72.66	68.60	63.21	65.79
Joy	83.15	78.72	80.87	81.52	89.29	85.23	90.91	85.94	88.35

Table 5.4: Per-emotion evaluation results of the proposed model on the SemEval-2018, AETD, and AIEmoD datasets

5.4.5 *Error Analysis*

To understand the classification behavior of the proposed model, an error analysis was conducted on the misclassified tweets. Table 5.5 presents some examples belonging to the emotions of sadness and anger. We found that a considerable portion of the errors is

caused by sarcasm, where the literal wording of the tweet does not directly correspond to the underlying emotional state. For instance, sarcastic statements such as رضينا بالهم والهم مش راضي بينا implicitly convey sadness, making it difficult for the model to correctly infer the intended emotion. Ambiguity represents another reason for misclassification, as some expressions simultaneously reflect multiple emotional states, leading the model to confuse closely related emotions such as sadness and anger. Short tweets and dialectal expressions with limited contextual information are particularly challenging for the model, as they provide insufficient linguistic cues for reliable emotion inference.

In general, these observations indicate that most misclassifications are not model failures but rather reflect the intrinsic complexity of IER in Arabic. The overlap between emotional categories, the prevalence of sarcasm and indirect expression, and the noisy nature of social media text all contribute to the challenge. Nevertheless, the hybrid AraBERT–BiGRU architecture demonstrates a strong ability to capture contextual cues despite these linguistic complexities.

Tweet	**True Label**	**Predicted Label**	**Error Cause**
رضينا بالهم والهم مش راضي بينا	Sadness	Anger	Sarcasm
فى الحقيقة عندى كبت محتاج يخرج للنور	Anger	Sadness	Ambiguity
الواحد جاب اخره اهيه بجد	Anger	Sadness	Short Context
ادق عليه اقول ها وينك تاخرت عالعزيمه يقول خمس دقايق وانا عندك والكل يعرف ان هالخمس دقايق تقلب خمس ساعات	Anger	Sadness	Sarcasm
مش كل مره انا المفروض عليا احس وانا المستمع الكيل مال	Sadness	Anger	Ambiguity

Table 5.5: Examples of misclassified tweets and their corresponding error causes

5.4.6 *Comparison with state-of-the-art Methods*

The results reported in Table 5.6 provide a comparative analysis between the proposed AraBERT-BiGRU model and existing SOTA approaches on the AETD dataset for both explicit and implicit emotion classification. Previous studies focused on explicit emotion

recognition, where emotional expressions are directly stated in the text. Within this setting, the strongest baseline, AraBERT integrated with Named Entity Recognition (NER) and Sentiment Analysis (SA) [148], achieved an F1-score of 81.60%. Notably, when evaluated on explicit emotions, our model attains a superior F1-score of 85.96%, outperforming all reported methods.

More importantly, under the more challenging IER scenario, the proposed model achieves an F1-score of 79.87%, remaining highly competitive despite the increased inferential complexity. This demonstrates the robustness and generalization capability of the hybrid architecture, which effectively leverages contextualized transformer representations alongside bidirectional sequential modeling.

Study	**Method**	**Exp./Imp.**	**F1-score (%)**
[107]	Complement Naïve Bayes	Explicit	65.80
[149]	Human-engineered Feature-based and Deep Feature-based (HEF+DF) model	Explicit	71.80
[148]	Fine-tuned AraBERT	Explicit	78.93
[148]	AraBERT Integrated with NER and SA	Explicit	81.60
Our Model [147]	AraBERT-BiGRU	Implicit	**79.87**
Our Model	AraBERT-BiGRU	Explicit	**85.96**

Table 5.6: Comparison with the state-of-the-art methods on the AETD dataset.

5.5 CONCLUSION

In this part of our research, we proposed a hybrid model that leverages the contextual representation capabilities of AraBERT, together with the sequential modeling strengths of BiGRU, to address the challenging task of IER in Arabic tweets. Additionally, we introduced the AIEmoD dataset, a carefully curated resource designed to address the scarcity of annotated datasets for implicitly expressed emotions in Arabic tweets. The proposed model was trained and evaluated on three datasets: AIEmoD, AETD, and SemEval-2018.

The experimental results demonstrate that the AraBERT–BiGRU model consistently outperforms several baseline deep learning architectures. In particular, the model achieved F1-scores of 79.87% on the AETD dataset and 70.67% on AIEmoD, confirming its ability to effectively infer implicit emotional cues. Compared to existing SOTA models for explicit

emotion recognition, our model achieves superior performance even when applied to IER, underscoring its robustness and adaptability across both explicit and implicit contexts.

In addition to the quantitative evaluation, we also present a qualitative error analysis to understand the model's classification behavior. We observed that misclassification cases were primarily caused by sarcasm, emotional ambiguity, dialectal expressions, and tweets with limited contextual information. Such linguistic phenomena often obscure the direct expression of emotions, making it difficult for the model to accurately infer the intended emotional state. These findings emphasize the complexity of implicit emotion detection, particularly in Arabic, where figurative language and contextual nuance are frequently used in informal communication.

These observations highlight both the strengths and limitations of hybrid deep learning architectures when addressing the task of IER. The proposed hybrid architecture shows strong performance and achieves significant improvements over traditional baselines. However, it still relies on supervised learning and requires large annotated datasets to achieve optimal results. This dependency represents a notable challenge in many real-world scenarios, particularly in low-resource languages such as dialectal Arabic, where labeled data is scarce and costly to obtain.

Recently, LLMs have emerged as a transformative paradigm in NLP. These models demonstrate a remarkable capacity to capture complex semantic relationships and subtle contextual cues across different languages and domains. Motivated by these advances, the following chapter investigates the potential of LLMs-based approaches to recognizing implicit emotions across different linguistic and socio-cultural contexts, including dialectal Arabic and English texts.

6

DECODING HIDDEN EMOTIONS ACROSS LANGUAGES: LLMS-BASED INSIGHTS FROM ARABIC DIALECTS TO ENGLISH TEXTS

6.1 INTRODUCTION

In recent years, LLMs, with their generative architecture, have achieved remarkable performance across a wide spectrum of NLP tasks [150, 151]. Currently, the field of emotion recognition has also entered the era of LLMs, with increasing research efforts attempting to leverage them to understand, interpret, and identify human emotions in text [11, 152]. Trained on massive and diverse corpora, LLMs demonstrate strong capabilities in capturing linguistic patterns, contextual dependencies, and even some aspects of world knowledge, enabling them to infer meaning and emotional content from context in ways that were previously difficult to achieve [153]. However, the question remains as to whether these models truly understand the emotions expressed in texts or whether they only exploit superficial lexical cues without achieving a deeper semantic understanding of emotional meaning. To address this question, this chapter adopts a two-step exploration based on two case studies:

- **Study I: Multi-label Emotion Recognition in Algerian Dialect using LLMs**: first, we address the low-resource setting by evaluating whether state-of-the-art LLMs can effectively handle the morphological richness and cultural nuances of the Algerian Arabic dialect. This study examines the models' ability to detect co-occurring

emotions in a dialectal context where emotional states are rarely expressed through explicit lexical markers and must instead be inferred from contextual and semantic cues.

- **Study II: Benchmarking Implicit Reasoning in English**: second, we extend the analysis to a high-resource language, English, in order to isolate the models' capacity for implicit emotional reasoning. This study evaluates their ability to interpret emotional content and perform subtle reasoning when explicit emotion words and direct lexical cues are absent.

Together, these two case studies provide insight into the capabilities and limitations of LLMs-based approaches to emotion recognition across different languages and textual contexts. In particular, we investigate whether the apparent "emotional intelligence" of LLMs represents a generalizable capability or is largely influenced by the linguistic and cultural characteristics of the data on which these models were trained. These insights contribute to a deeper understanding of how modern language models can support emotion analysis tasks in both low-resource dialectal settings and more widely studied languages.

DECLARATION: This chapter is based on three peer-reviewed international conference publications:

- **Part I** (Section 6.2)
  - H. Boutouta, F. Senator, Z. Lakab, C. Mediani, and A. Lakhfif, "Multi-label emotion recognition in low-resource dialects: A case study on algerian arabic with large language models", Procedia Computer Science, vol. 275, pp. 817–824, 2026, 7th International Conference on AI in Computational Linguistics, issn: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2026.01.094. 2026.01.094. [Online] Available: https://www.sciencedirect.com/science/article/pii/S1877050926000943 [154].
- **Part II** (Section 6.3)
  - H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, "From context to emotion: Leveraging LLMs for recognizing implicit emotions", in Proceedings of

the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025), Southern Denmark University, Odense, Denmark: Association for Computational Linguistics, Aug. 2025, pp. 399–409. [Online]. Available: https://aclanthology.org/2025.icnlsp-1.39/ [155].

– H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, "Do LLMs Grasp Emotion or Just Keywords? A Benchmark for Implicit Emotion Recognition", In 2025 1st International Conference on Artificial Intelligence and Cyber Physical Systems (AICPS), 2025, pp. 104–109. DOI: https://doi.org/10.1109/AICPS66617.2025.11513483 [156].

6.2 MULTI-LABEL EMOTION RECOGNITION IN ALGERIAN DIALECT

6.2.1 *Introduction*

Emotion recognition in Arabic dialects has emerged as an important yet challenging task within NLP, driven by the growing volume of user-generated content on social media and online platforms. Unlike MSA, dialectal Arabic represents the language of everyday communication and thus provides a more realistic setting for studying emotional expression in natural contexts. However, this task is complicated by the interplay of linguistic diversity and limited computational resources.

Among these dialects, Algerian Arabic poses particular challenges due to its unique characteristics. Despite being spoken by more than 40 million people and widely used in informal communication, the Algerian dialect has received limited research attention and lacks adequate computational resources [157]. The dialect exhibits significant dialectal variation across different regions, with distinct morphological, phonetic, lexical, and syntactic features that are quite different from MSA and other Arabic dialects [50, 158]. This complexity is further amplified by frequent code-switching between Arabic, Berber, and French within the same sentence [159], resulting in highly heterogeneous and unpredictable linguistic patterns. Additional challenges arise from the lack of standardized orthography, which results in significant variation in how words expressing emotions are

written. This issue is further complicated by lexical ambiguity and polysemy, where the same word may convey different emotional meanings depending on the context [160].

In addition to these challenges, cultural and social diversity across dialect-speaking communities plays a crucial role in shaping emotional expression, which is often implicit and context-dependent, reflecting local norms and communication styles [51, 160]. Collectively, these factors make emotion recognition in the Algerian dialect a highly complex task that requires advanced models capable of capturing both linguistic variability and contextual emotional cues.

Recently, LLMs have demonstrated remarkable capabilities in understanding, interpreting, and recognizing human emotions [161, 162]. Compared to traditional transformer-based models, their contextual representation and reasoning abilities make them particularly well-suited for addressing the complexity of emotion recognition in low-resource and dialectal settings. Therefore, in this work, we investigate the effectiveness of state-of-the-art LLMs for multi-label emotion recognition in the Algerian dialect. Specifically, we evaluate a set of Arabic-focused PLMs alongside general-purpose LLMs, with the aim of assessing their ability to capture linguistic variability and infer context-dependent emotional cues.

6.2.2 *Experiments*

6.2.2.1 *Dataset*

Here, we used the BRIGHTER dataset introduced by Muhammad et al. [163] from the SemEval-2025 Task 11 [164]. It constitutes a large-scale, manually annotated resource for multi-label emotion recognition, covering more than thirty languages from seven language families spoken across Africa, Asia, Europe, and the Americas. See more details on their website [1].

For the purpose of this study, we focus specifically on the Algerian dialect subset (Arq) provided within the dataset. This subset comprises 1,903 gold-standard annotated instances, divided into 901 training samples, 100 development samples, and 902 testing samples. Each textual instance is annotated according to a multi-label emotion classification scheme, where six basic emotions (joy, anger, fear, surprise, sadness, and disgust) are

1 https://brighter-dataset.github.io/

assigned binary labels indicating the presence (1) or absence (0) of each emotion in the text.

The Algerian Arabic data originates from literary texts, specifically a manual translation of the novel La Grande Maison ("The Big House") by the Algerian writer Mohammed Dib. The novel was translated from French into Algerian Arabic by a bilingual native speaker, producing a dialectal version of the narrative. Subsequently, the translated text underwent further preprocessing to segment it into individual sentences, which were then annotated for emotional content by native speakers.

This literary source is particularly valuable for emotion analysis, as narrative texts typically contain rich emotional expressions and complex interactions between characters. Moreover, emotions in such narratives are frequently conveyed implicitly, emerging through contextual cues and narrative descriptions rather than through explicit emotion words. Consequently, this characteristic makes the dataset well suited for investigating the challenges of IER.

6.2.2.2 *Models Selection*

Our experiments below rely on a diverse set of models, including three Arabic PLMs and three general-purpose LLMs, which we briefly introduced in Sections 2.4.4.1 and 2.4.4.3. Such diversity allows for a comprehensive assessment of how differences in architecture, training objectives, and linguistic specialization affect performance on dialectal Arabic data.

PRETRAINED LANGUAGE MODELS We selected three dialectal Arabic PLMs, namely DziriBERT [89], AraBERT-Algerian [165], and MARBERTv2 [86]. We used the base versions from the Hugging Face transformers library. Following [10], we added a classification layer on top of the [CLS] token.

We fine-tuned these models on the Algerian Arabic dialect subset (Arq) training data for 5 epochs with an early stopping strategy. We applied a threshold tuning strategy to determine the optimal threshold that maximized the F1 score. Specifically, the threshold was searched within the range $[0.1, 0.9]$ in increments of 0.01, and the value that maximized the macro F1 on the development set was selected. The fine-tuning process employed a binary cross-entropy loss with sigmoid activation to handle multiple emotion labels per

instance. The AdamW optimizer was employed with a linear learning rate of 2e-5 and a training batch size of 8. The maximum sequence length is set to 180 tokens for all models to ensure consistent input formatting. The adopted hyperparameter setting was introduced in Table 6.1.

Hyperparameter	**Value**
Batch size	8
Learning rate	2e-5
Optimizer	AdamW
Epochs	5
Max sequence length	180

Table 6.1: Hyperparameters used for fine-tuning pretrained models.

GENERAL-PURPOSE LLMS We employ three general-purpose LLMs: namely, ChatGPT, Gemini, and DeepSeek. These models were accessed through their official chatbot interfaces using the GPT-5, Gemini 2.5 Flash, and DeepSeek-V2 versions, respectively. In order to evaluate their capability to infer emotions without task-specific adaptation, the models were assessed under a zero-shot experimental setup.

PROMPT DESIGN In our initial experiments, we tested a variety of prompts in both Arabic and English to evaluate their effectiveness in handling the linguistic and contextual nuances of dialectal emotion recognition. Our observations indicated that prompts formulated in English consistently outperformed those written in Arabic. Building on these findings, we developed a structured English prompt template to enhance zero-shot performance. This template guides the model by (1) assigning a specific *role* to the model, which is "Expert in multi-label emotion recognition"; (2) explicitly identifying the *task* to be performed, such as classifying emotions expressed in Algerian dialect text; (3) clarifying the nature of the task, including the possibility that a single text may convey more than one emotion simultaneously; and (4) providing the *set of emotion labels* to ensure precise and context-aware predictions. This structured approach enables the model to operate within clear and well-defined parameters.

The final prompt was selected after evaluating multiple variants on the development subset. For evaluation, we used the same Algerian Arabic dialect subset (Arq) test set for all models to ensure a fair and consistent comparison.

6.2.2.3 *Evaluation*

For the evaluation, precision, recall, and F1-score were computed for all models under each emotion class, with their macro-averaged values providing an overall performance comparison. Given the multi-label nature of the task, we further assessed model performance using three distinct metrics: Exact-match accuracy was used to evaluate the proportion of instances for which the predicted set of emotions exactly matched the ground truth labels; the Jaccard index (Multi-label Accuracy) was calculated to measure the average overlap between predicted and actual label sets; and Hamming loss was employed to quantify the fraction of incorrectly predicted label-instance pairs.

6.2.3 *Results and Discussion*

Table 6.2 presents the performance of different models. The fine-tuned PLMs show lower overall performance than the zero-shot LLMs. While the Gemini model achieved the highest macro F1-score among all models, AraBERT-Algerian obtained the highest macro F1-score among the BERT-based models, closely followed by MARBERTv2. DziriBERT exhibits the strongest recall but suffers from low precision, indicating that although it identifies more emotion instances, it also produces a high number of false positives. These findings demonstrate the stronger ability of zero-shot LLMs to detect emotions within the dialectal context.

However, it is important to acknowledge a potential limitation related to LLMs pretraining data. These models are trained on extremely large and diverse corpora, which may contain emotional expressions, dialectal Arabic content, or contextual patterns. Consequently, part of the observed zero-shot performance may stem from the models' prior exposure to similar linguistic or emotional structures rather than from genuine out-of-distribution generalization. This raises the possibility that the LLMs are leveraging memorized patterns or previously learned emotional cues instead of inferring emotions purely from contextual reasoning.

Model	Precision (%)	Recall (%)	F1-score (%)
Fine-tuned PLMs			
DziriBERT	53.80	**68.58**	59.54
AraBERT-Algerian	55.46	66.96	60.21
MARBERTv2	57.33	64.55	60.03
Zero-shot LLMs			
ChatGPT	63.96	55.26	57.73
Gemini	**67.50**	66.90	**66.53**
DeepSeek	65.82	56.04	60.03

Table 6.2: Performance results of fine-tuned PLMs and zero-shot LLMs on multi-label emotion classification.

6.2.3.1 *Multi-label Classification Performance*

From Table 6.3, it can be deduced that all models face difficulty in correctly predicting the complete set of co-occurring emotions. While the fine-tuned PLMs achieve competitive Jaccard accuracy (44–45%), their exact-match accuracy remains low (<17%). This disparity suggests that while these models correctly identify a significant portion of the relevant emotions, they rarely capture the entire subset perfectly. Interestingly, among the PLMs, MARBERTv2 achieves the highest exact-match accuracy (16.69%) and the lowest Hamming loss (0.269), indicating a more precise boundary for label prediction compared to the other PLMs.

Zero-shot LLMs generally outperform the fine-tuned models in exact-match accuracy (17–21%), with Gemini emerging as the top performer across all metrics (21.13% exact-match and 50.41% Jaccard). The lower Hamming loss observed in the LLMs category (0.20–0.24) suggests these models are more accurate compared to the PLMs. However, the overall difficulty in reconstructing complete emotion sets highlights persistent challenges in capturing label co-occurrence. This limitation is particularly pronounced in Algerian dialect texts, where emotions are often expressed implicitly or through overlapping linguistic cues that require deep reasoning to fully disentangle.

6.2.3.2 *Per-emotion Performance*

From Table 6.4, it can be seen that both model families offer distinct advantages for emotion detection in the Algerian dialect. Fine-tuned PLMs provide domain robustness and

Model	Correct	Exact-match (%)	Jaccard Acc. (%)	Hamming Loss
Fine-tuned PLMs				
DziriBERT	122	15.08	45.27	0.286
AraBERT-Algerian	130	16.07	44.96	0.276
MARBERTv2	**135**	**16.69**	44.91	0.269
Zero-shot LLMs				
ChatGPT	154	19.03	41.46	0.241
Gemini	**171**	**21.13**	**50.41**	**0.206**
DeepSeek	145	17.92	43.33	0.233

Table 6.3: Multi-label classification performance: Comparing Fully Correct Samples, Exact-match Accuracy, Jaccard Index, and Hamming Loss across models.

reliable discrimination between closely related emotions, while zero-shot LLMs exhibit superior flexibility, coverage, and contextual inference. The findings emphasize the growing capability of multilingual LLMs to handle dialectal variation. Nevertheless, overall performance remains limited by the quality and availability of annotated resources in the Algerian dialect. In particular, per-emotion F1-scores vary widely, with "disgust" and "fear" performing notably worse. These disparities could be attributed to several factors: the sparsity of emotion-specific terms, ambiguity in lexical items within the Algerian dialectal context, and the implicit nature of emotional expression, as Algerian speakers often convey their emotions indirectly without using explicit emotion words.

Model	Anger	Disgust	Fear	Joy	Sadness	Surprise
Fine-tuned PLMs						
DziriBERT	63.44	51.44	53.26	49.36	71.53	**68.19**
AraBERT-Algerian	63.21	50.04	55.67	57.64	69.76	65.00
MARBERTv2	62.33	49.05	56.65	58.00	70.68	63.42
Zero-shot LLMs						
ChatGPT	62.55	32.43	62.02	63.06	69.10	57.25
DeepSeek	62.96	44.21	**64.64**	61.31	69.60	57.48
Gemini	**71.40**	**51.58**	62.90	**69.70**	**78.09**	65.54

Table 6.4: Per-emotion F1-scores (%) across models.

6.2.3.3 *Error Analysis*

Table 6.5 presents cases in which Gemini encountered difficulties in detecting emotional expressions. In the first example, it predicted *fear* and *sadness* instead of *disgust*. The second example illustrates the model's failure to identify the appropriate emotion: in تهدر بلعقل وبلخف (talking quietly and cautiously), the behaviour reflects fear, yet the model interpreted it as enjoyment in social interactions.

The third and fourth examples show that the most recurrent errors occur when certain lexical items are misinterpreted in the Algerian dialectal context, such as اللقية ("disgusting") and ثانيت ("also").

Text	Type	Anger	Disgust	Fear	Joy	Sad	Surprise
الراوي: عيني عاودت هاذ الهدرة عشرة، مية، ألف خطرة.	Truth	0	1	0	0	0	0
The narrator: Aini repeated this sentence ten, a hundred, a thousand times.	Pred	0	0	1	0	1	0
الراوي: عيني كملت تهدر بلعقل و بلخف مع الجارة.	Truth	0	0	1	0	0	0
The narrator: Aini finished talking quietly and swiftly with the neighbor.	Pred	0	0	0	1	0	0
عمر خمم : كيفاش بقع هاذ اللقية قادر نحمل شوية ماكلة ؟	Truth	0	1	0	0	0	0
Omar thought: How on earth in this disgusting thing I even able to eat a little food ?	Pred	0	0	1	0	1	0
كامل عاودوا كيف كيف: لحم ثانيت.	Truth	0	0	0	1	0	1
All of them said simultaneously: Also the meat .	Pred	1	0	1	0	1	0

Table 6.5: Error analysis of texts misclassified by the Gemini model.

6.2.3.4 *Performance comparison with State-of-the-Art Systems*

Table 6.6 presents a comparative analysis between our evaluated models and the SemEval-2025 Task 11 baselines [164], including the Majority Vote ($M_{baseline}$) and RoBERTa ($R_{baseline}$), alongside the top-performing systems on the Algerian Arabic (Arq) dataset. The results

demonstrate that Gemini, even in a zero-shot setting, achieved a highly competitive macro F1-score that closely approaches those of the top-ranked system. Furthermore, the fine-tuned AraBERT-Algerian model demonstrated robust performance, significantly outperforming both the $M_{baseline}$ and $R_{baseline}$ by a wide margin.

System	Macro-F1 (%)
PAI Team [166]	66.8
Gemini (Evaluated)	**66.5**
JNLP Team [167]	64.1
AraBERT-Algerian (Evaluated)	**60.2**
$M_{baseline}$ [164]	44.5
$R_{baseline}$ [164]	41.4

Table 6.6: Performance comparison of baselines and top-ranked systems on SemEval-2025 Task 11 in the Algerian Arabic dataset (Arq).

6.2.4 *Conclusion*

This study provides a comprehensive evaluation of general-purpose LLMs and Arabic-specific PLMs models for multi-label emotion recognition in the Algerian dialect. The results demonstrate that general-purpose LLMs can achieve competitive performance in zero-shot settings, while fine-tuned Arabic-specific models remain highly effective in supervised scenarios. Overall, these findings highlight the strong potential of modern language models to address the linguistic variability and emotional complexity of dialectal Arabic, confirming that both zero-shot and fine-tuning strategies are viable for low-resource emotion recognition tasks.

6.3 BENCHMARKING IMPLICIT REASONING IN ENGLISH

6.3.1 *Introduction*

With the rapid growth of user-generated content on social media platforms, an increasing proportion of texts convey emotion implicitly [168], making IER an important and emerging topic in NLP [135]. However, this task remains challenging and relatively underexplored,

as accurately identifying emotions requires models to move beyond the lexical level and instead capture deeper semantic and structural relationships within the text [122]. In many cases, emotions must be inferred from contextual signals such as events, actions, causes, consequences, interactions between entities, and the broader situation being described [18, 41]. For example, consider the sentence *"She stared at the mailbox for hours"*—although no explicit emotion word appears, the text conveys a sense of sadness or anticipation. Moreover, implicit emotional expression encompasses a wide range of subtle linguistic phenomena, including sarcasm (e.g., *"Perfect, just what I needed"*) [58], ambiguous statements (e.g., *"She looked out the window as the train pulled away"*), which could imply sadness, longing, or even relief [22], and metaphorical language (e.g., *"A storm was brewing inside her"*), which implicitly expresses anger [21].

Early approaches to IER primarily relied on rule-based and knowledge-based methods [16, 118, 120], as well as classical ML-based models [16, 121, 122]. With the emergence of DL architectures, researchers began developing end-to-end models incorporating attention mechanisms to better capture contextual and semantic cues [125–127, 129, 132]. More recently, Transformer-based models such as BERT and RoBERTa have achieved strong performance across various NLP tasks, including IER [133–135].

With the advent of LLMs, a new era of research has begun to explore their potential for affective understanding, leveraging their extensive pretraining on diverse corpora and their capacity for in-context reasoning [153]. Despite their remarkable performance on a broad range of NLP benchmarks, their capability to perform implicit emotional reasoning in the absence of explicit emotional cues remains an open and insufficiently explored question.

To address this gap, this study benchmarks the implicit emotional reasoning capabilities of a set of state-of-the-art language models on the task of IER in English, a high-resource language in which these models have been extensively trained, thereby ensuring that any observed limitations can be attributed to deficiencies in affective reasoning rather than to linguistic unfamiliarity or data scarcity. Through this exercise, we aim to draw a clearer boundary between what LLMs effectively comprehend and what they merely approximate through learned co-occurrence patterns and to contribute a more grounded understanding of the current capabilities and limitations of LLMs in the domain of affective computing.

The main contributions of this study are summarized as follows:

- We conduct a comparative analysis of two paradigms for IER: medium-sized masked language models and prompting-based LLMs.
- We evaluate and compare the efficacy of zero-shot and few-shot learning for enhancing implicit emotional reasoning in LLMs.
- We test models on two distinct emotional scenarios: naturally implicit emotions and context-only inference with masked explicit emotion words, enabling a robust analysis of implicit emotional understanding.
- We provide a detailed analysis of the strengths and weaknesses of both model families in capturing implicit emotional cues, offering insights into their limitations.

6.3.2 *Experiments*

Our experiments rely on two categories of transformer-based models: pre-trained encoder-based models and generative decoder-based LLMs, across two implicit emotion scenarios and three different classification methods.

6.3.2.1 *Datasets Characteristics and Preprocessing*

The experiments were conducted using two publicly available benchmark datasets for emotion recognition: the WASSA-2018 Implicit Emotions Shared Task (IEST) dataset[2] and the ISEAR dataset [3].

IEST This dataset was developed for the WASSA-2018 Implicit Emotions Shared Task [41]. It is an automatically labeled dataset of 191,731 English tweets, split into 153,600 for training, 9,600 for validation, and 28,800 for testing. Each tweet is annotated with one of Ekman's six basic emotions: anger, disgust, fear, joy, sadness, or surprise.

To simulate implicit emotion scenarios, each tweet in the dataset has had its explicit emotion word masked and replaced with a placeholder token [#TARGETWORD#]. This design forces models to rely only on contextual cues to infer the underlying emotion, making it

2 https://implicitemotions.wassa2018.com/data/
3 https://github.com/sinmaniphel/py_isear_dataset

particularly suited for research on emotion understanding in indirect and implicit expressions. Due to computational constraints, the experiments were conducted exclusively on the test set using the officially released masked version. Some examples from the dataset are provided in Table 6.7.

Emotion	**Tweet**
Anger	I get impatient and [#TARGETWORD#] when I'm hungry.
Disgust	So many people looked at me just [#TARGETWORD#] when I said that mustaches are hot.
Fear	So [#TARGETWORD#] that I'm not good enough
Joy	you're gonna be [#TARGETWORD#] when you realize you deserve to be.
Sadness	Very [#TARGETWORD#] when he goes on these tirades
Surprise	They just jealous, they get [#TARGETWORD#] when she pull up.

Table 6.7: Example tweets from the IEST dataset, with the emotion word masked as [#TARGETWORD#].

ISEAR A manually labeled dataset was collected as part of a psychological study aimed at exploring emotional experiences across cultures [116]. The data were gathered from over 3,000 participants in 26 countries, all of whom had university-level education and were fluent in English. Each participant was asked to describe situations in which they had personally experienced one of seven emotions: joy, fear, anger, sadness, disgust, shame, and guilt. In total, the dataset contains approximately 7,666 instances, making it one of the most widely cited benchmarks for emotion recognition in psychology and affective computing tasks. Examples from the dataset are provided in Table 6.8.

The two datasets (IEST and ISEAR) differ significantly in terms of domain, genre, writing style, and emotional expression. A comparative overview is presented in Table 6.9. These differences introduce varying levels of linguistic complexity and contextual ambiguity, which may significantly influence models' performance and generalization across datasets.

Emotion	Example
Joy	An encounter with a man whom I love, after a very long separation.
Fear	After mischieviously ringing on the chemist's trade-entrance doorbell and getting caught by him.
Anger	At my Summer job, nobody looked after me in particular and I had to learn all on my own.
Sadness	After I had lived with my boyfriend in a foreign country for half a year, I saw that it was impossible for me to stay with him (for economic reasons). We separated although I loved him.
Disgust	A mother who shouts at her child for nothing.
Shame	During carnaval I danced for a few minutes normally I don't dance because I am rigid in my moving around during a dance, I stopped very soon.
Guilt	I speak harshly to my parents though they only mean my own good.

Table 6.8: Example from the ISEAR dataset for each emotion label.

Aspect	IEST	ISEAR
Emotional Expression	Implicit (emotion word masked)	Explicit (emotion word present)
Annotation	Automatically labeled	Manually labeled
Emotion Labels	Ekman's six basic emotions: anger, disgust, fear, joy, sadness, surprise	anger, disgust, fear, joy, sadness, shame, guilt
Genre	social media (Twitter)	Survey responses
Style	Informal, noisy, fragmented sentence	formal, structured, complete sentences
Text Length	Short (tweets, < 280 characters)	Medium (1–3 sentences per instance)
Size	191,731 instances	7,666 instances
Purpose	IER	ER
Contextual Clues	Sparse; relies on social and situational context	Rich descriptions of emotional experiences

Table 6.9: Key differences between the IEST and ISEAR datasets.

PRE-PROCESSING To align the ISEAR dataset with the IER task, we applied an additional filtering step: we ensured that none of the selected instances (for the testing set) contained explicit emotion words. This pre-processing step allows us to reuse the ISEAR as a proxy dataset for IER, focusing only on instances where emotions must be inferred from the described context rather than directly stated. Furthermore, the IEST dataset was subjected to standard pre-processing steps, including the removal of HTML tags, URLs, emojis, and extra whitespace, as well as the normalization of inconsistent punctuation. In

addition, user mentions and hashtags were either removed, and elongated words were reduced to their canonical forms.

After preprocessing, datasets were divided into training, validation, and testing sets in proportions of 80%, 10%, and 10%, respectively. The label distributions within the datasets are illustrated in Figure 6.1 and Figure 6.2.

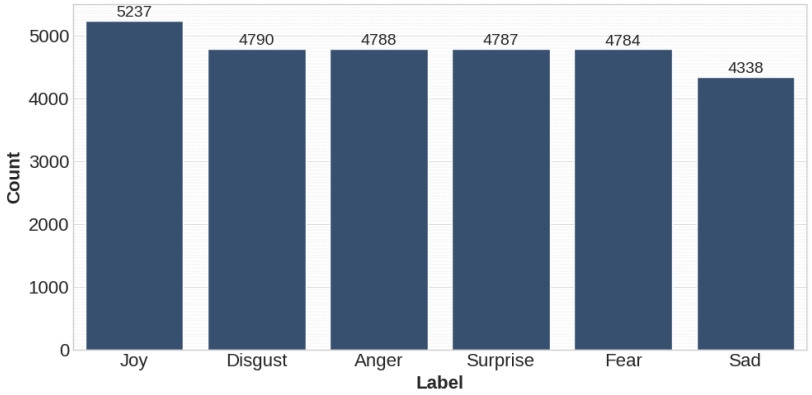


Figure 6.1: Distribution of emotions in the IEST dataset.

6.3.2.2 *Models Selection*

In this study, we investigate two different transformer-based architectures: (1) pre-trained encoder-based models and (2) generative decoder-based LLMs. The objective is to provide a comprehensive comparison between traditional supervised learning approaches and emerging prompt engineering methodologies in handling implicit emotional content.

ENCODER-BASED MODELS In the first approach, we employ encoder-based models, namely BERT and RoBERTa, which are fine-tuned on task-specific labeled datasets. These

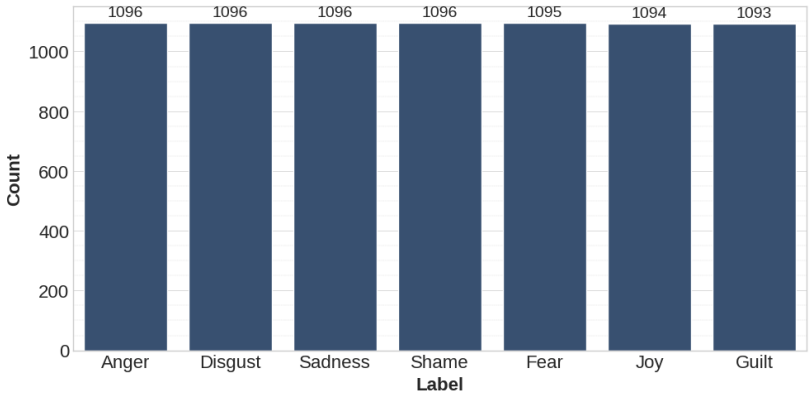


Figure 6.2: Distribution of emotion in the ISEAR dataset.

models are initialized with pre-trained weights and subsequently optimized using supervised learning to capture contextual semantic representations relevant to emotion classification. Their effectiveness in various NLP tasks makes them strong baselines for evaluating performance in IER.

HYPERPARAMETER SETTINGS We explored a range of hyperparameter configurations to optimize performance. Specifically, we experimented with different learning rates (1e-5, 2e-5, and 3e-5), batch sizes (16 and 32), training durations (ranging from 3 to 6 epochs), and maximum sequence lengths (64, 128, and 512). Additionally, we compared different model variants, including base and large versions. Each configuration was evaluated using the 10% development split, and the optimal setup was selected based on the macro F1-score. The chosen hyperparameters were validated across various random seeds to ensure robustness. Table 6.10 summarizes both the tested and optimal hyperparameter configurations.

Hyperparameter	**Tested Values**	**Optimal Value**
Learning Rate	1e-5, 2e-5, 3e-5	1e-5
Loss Function	Categorical Cross-Entropy	Categorical Cross-Entropy
Optimizer	Adam	Adam
Batch Size	16, 32	32
Epochs	3, 4, 5, 6	4
Max Length	64, 128, 512	128

Table 6.10: Hyperparameters tuning

We used the pre-trained "bert-base-uncased" and "roberta-base" models from the Hugging Face Transformers library. The models consist of 12 transformer layers, a hidden size of 768, and 12 attention heads. For both models, we appended a dense layer with a softmax activation function for classification. The models were trained for 4 epochs using a batch size of 32 and a maximum sequence length of 128. Training was performed using the Adam optimizer and categorical cross-entropy loss. We evaluated these models on the held-out test sets from the IEST and ISEAR datasets.

GENERATIVE LLMS In addition to supervised models, we evaluate the performance of generative decoder-based LLMs through various prompting strategies. Specifically, we

conduct experiments using a selection of SOTA models, including ChatGPT-5, Phi-4, LLaMA-3, and Gemma-3, under both zero-shot and few-shot configurations.

The closed-source model ChatGPT-5 was accessed via its chatbot interface, specifically using the GPT-5 variant. The open-source models, namely Phi-4, LLaMA-3, and Gemma-3, were evaluated through the Ollama framework utilizing their respective quantized configurations: Phi-4:14B, LLaMA-3:8B, and Gemma-3:12B. For evaluation, we used the same test sets from the IEST and ISEAR datasets, ensuring a fair and consistent comparison across all models.

PROMPTS DESIGN To evaluate the overall performance of the LLMs, we designed relatively consistent prompts that were applied across both datasets and models, rather than specific designs. Experiments were conducted under both zero-shot and few-shot settings.

- **Zero-Shot learning**: Early versions of the prompts contained only basic task instructions. The prompts were then iteratively refined to better align with the specific requirements of the task. We observed improved performance when the role of the model was explicitly defined, the implicit nature of the task was explicitly stated, and emotion labels were clearly specified. Consequently, the final prompts incorporated three essential components: the model role, a clear task definition, and a constrained output format. For instance, the model was consistently framed as an "expert in implicit emotion recognition," and formulations such as "the emotion is implied rather than stated" were included. Additionally, the model was instructed to return a single emotion selected from a predefined list.

  The final format was selected after testing several prompt versions on a development subset of the IEST and ISEAR datasets, evaluating performance manually and through agreement with gold-standard labels.

- **Few-Shot learning**: Here, models are shown a small number of exemplars to illustrate the relationship between text and emotion before being asked to predict the target instance. Each exemplar consists of a short input text followed by its corresponding emotion.

As shown in Figure 6.3, these examples are embedded directly within the prompt in a consistent structure, followed by the query instance. This setup aims to determine whether minimal supervision improves the ability of the LLMs to capture implicit emotional cues and generalize across unseen contexts.

Input

You are an expert in implicit emotion recognition.

The following sentence contains an emotion that is expressed implicitly. Based on context alone, identify the most likely emotion.

Choose only one from: [Emotion List].

Respond with the emotion without any explanation.

Example 1: "When I felt being treated unjustly by a teacher."

Output: Anger

Example 2: "When I could not do anything for my brother."

Output: Sadness

Now classify the following:

Text: [example text]

Output: Emotion

Figure 6.3: Prompt example for IER. The examples inside the dashed box are demonstrations of the few-shot setting and would be removed under the zero-shot setting.

6.3.2.3 *Evaluation Metrics*

For evaluation, precision, recall, and F1-score were computed across all emotion classes for each model. Overall performance was assessed using accuracy along with the macro-averaged metrics.

6.3.3 *Results and Discussion*

The experimental results are presented in Table 6.11. We report the evaluation results for the encoder-based and decoder-based models, using different approaches (fine-tuning, zero-shot, and few-shot prompting) across the IEST and ISEAR datasets.

6.3.3.1 *Results on Supervised Fine-tuning*

The fine-tuned encoder-based models demonstrate high performance with extreme stability across evaluation metrics. On the ISEAR dataset, RoBERTa achieved an accuracy of 70.84% and an F1-score of 70.34%. However, the true strength of this approach is observed in the IEST dataset, where RoBERTa maintained its robustness, achieving the highest accuracy (66.88%) and F1-score (66.67%). This highlights that task-specific optimization through fine-tuning allows encoders to effectively capture linguistic nuances and informal emotional cues.

6.3.3.2 *Generative Approach: Zero-shot vs. Few-shot*

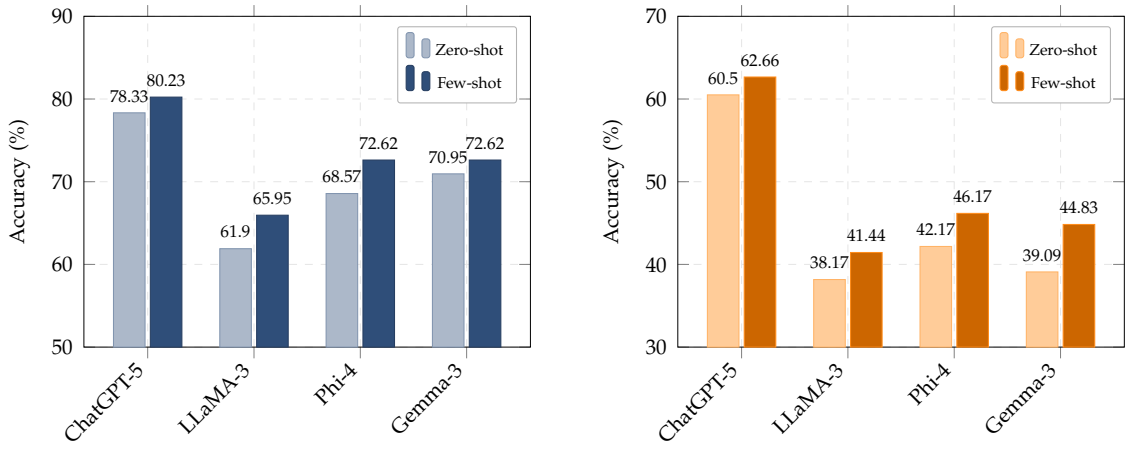
We can see that all evaluated LLMs consistently benefit from in-context examples (few-shot) compared to the zero-shot setting (see Fig. 6.4). Among them, ChatGPT achieves the best overall results, with its accuracy increasing from 78.33% to 80.23% on the ISEAR dataset and from 60.50% to 62.66% on the IEST dataset. Similarly, Phi-4 shows a notable improvement from 68.57% to 72.62%, while LLaMA-3 and Gemma-3 also exhibit consistent gains, improving from 61.90% to 65.95% and from 70.95% to 72.62% on ISEAR, respectively. These findings highlight the effectiveness of in-context learning, where providing a few labeled examples enhances emotion recognition performance by clarifying task expectations and emotional semantics.

A comparative analysis of the two approaches reveals a clear trade-off between the generalization and adaptation capabilities of the models in IER, largely influenced by model architecture, training strategy, and dataset characteristics. Generative LLMs demonstrate strong generalization by leveraging extensive contextual knowledge to infer emotions in well-structured and context-rich narratives (in the ISEAR dataset). In contrast,

encoder-based models benefit from masked language modeling objective combined with task-specific fine-tuning, which enhance their ability to capture fine-grained contextual dependencies (in the IEST dataset). Overall, these findings indicate that, while LLMs excel in generalization, supervised encoder models remain more reliable for domain-specific adaptation in challenging and constrained IER scenarios.

Approach	Model	ISEAR				IEST			
		Acc. (%)	F1. (%)	Prec. (%)	Rec. (%)	Acc. (%)	F1. (%)	Prec. (%)	Rec. (%)
Fine-tuned	BERT	70.36	69.66	70.33	70.66	62.02	61.52	61.83	62.00
	RoBERTa	70.84	70.34	70.79	71.00	**66.88**	**66.67**	**66.83**	**66.50**
Zero-shot	ChatGPT-5	78.33	77.98	78.55	78.33	60.50	61.05	61.61	60.50
	LLaMA-3	61.90	38.08	42.00	39.39	38.17	24.54	33.41	25.44
	Phi-4	68.57	47.81	49.28	48.00	42.17	42.32	47.55	42.17
	Gemma-3	70.95	71.08	75.84	70.95	39.09	38.73	48.71	39.08
Few-shot	ChatGPT-5	**80.23**	**79.89**	**80.49**	**80.24**	62.66	63.19	65.81	62.67
	LLaMA-3	65.95	41.62	44.44	41.97	41.44	40.98	47.24	41.31
	Phi-4	72.62	62.23	64.48	63.54	46.17	46.47	48.31	46.17
	Gemma-3	72.62	72.30	75.80	72.62	44.83	45.44	50.00	44.83

Table 6.11: Comparative results of different models on both ISEAR and IEST datasets. Acc, F1, Prec, and Rec denote accuracy, macro F1 score, precision, and recall, respectively.



(a) ISEAR Dataset

(b) IEST Dataset

Figure 6.4: Comparative analysis of Zero-Shot and Few-Shot Results across ISEAR and IEST datasets.

6.3.3.3 *Performance Across Datasets*

In Fig. 6.5, a consistent performance gap is observed between the two datasets, particularly for open-source LLMs. The ISEAR dataset comprises well-structured, self-reported

emotional experiences written in a coherent narrative style with relatively consistent grammar, thereby providing rich contextual cues that facilitate more accurate emotion inference across models.

In contrast, the IEST dataset represents a more challenging scenario. It is derived from social media (tweets), which are typically informal, noisy, and contextually ambiguous. In addition, tweets frequently contain slang, sarcasm, abbreviations, and culture-specific references that are difficult to interpret without broader context. This shift in genre introduces additional challenges for IER, as models must not only infer implicit emotions but also deal with less structured and noisier linguistic input.

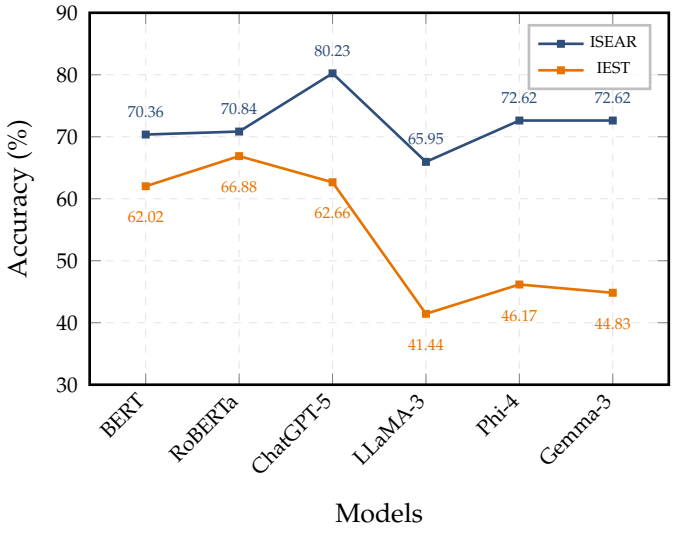


Figure 6.5: Performance Variance across the ISEAR and IEST Datasets (LLMs results are reported using the few-shot setting).

6.3.3.4 *Performance Variation Across implicit scenarios*

When examining implicit emotional scenarios across the two datasets, clear differences in the performance of the models were emerged. The ISEAR dataset contains naturally implicit emotions embedded within coherent, narrative-style descriptions, providing richer contextual cues that models can effectively leverage to infer the underlying emotions.

In contrast, the IEST dataset simulates implicit emotion by masking emotion words that were originally present in the tweets. This deliberate masking weakens contextual support and forces models to rely on incomplete and often ambiguous linguistic cues. This distinction highlights how the availability and quality of contextual information directly

affect task difficulty, with the IEST dataset posing a greater challenge due to its incomplete informative cues.

6.3.3.5 *Performance on Individual Emotions*

Table 6.12 presents a detailed comparison of the F1 scores for each emotion class across all models on the ISEAR dataset. The results show that ChatGPT-5 consistently achieves the highest F1 scores across most emotion categories, particularly Joy, Fear, and Disgust, highlighting its strong ability to recognize emotions when clear contextual cues are present. The few-shot setting generally yields modest performance improvements over the zero-shot and fine-tuning settings, with notable gains in more challenging categories such as Shame and Guilt. This finding suggests that minimal supervision helps the model better infer subtle emotional cues. Nevertheless, all models exhibit comparatively lower performance on Shame and Guilt than on other emotions, reflecting the inherent difficulty of capturing complex affective states and underscoring the critical role of contextual clarity in recognizing implicit emotions.

Approach	Model	Joy	Sadness	Anger	Fear	Disgust	Shame	Guilt
Fine-tuned	BERT	86.24	75.95	59.85	78.00	76.00	47.16	64.25
	RoBERTa	92.17	77.52	62.30	71.32	75.20	51.49	62.39
Zero-shot	ChatGPT-5	**96.70**	79.41	**75.86**	85.00	80.37	56.36	72.10
	LLaMA-3	84.62	63.69	62.40	76.03	71.43	38.83	21.92
	Phi-4	90.76	76.19	65.31	76.56	73.50	42.86	52.94
	Gemma-3	95.80	73.55	61.54	79.31	69.47	55.26	62.63
Few-shot	ChatGPT-5	**96.70**	**80.59**	**75.86**	**87.50**	**82.24**	**60.55**	**75.80**
	LLaMA-3	87.88	66.27	62.26	76.11	74.29	53.33	37.65
	Phi-4	95.87	75.68	67.89	77.94	75.41	42.86	62.18
	Gemma-3	95.87	75.17	70.59	80.65	73.08	57.53	53.19

Table 6.12: Per-emotion Result on the ISEAR dataset.

The results on the IEST dataset, presented in Table 6.13, reveal a different trend compared to that observed for the ISEAR dataset. In this case, RoBERTa achieves the best performance across most emotion categories, particularly Joy and Fear. Among the LLMs, ChatGPT-5 attains the highest performance across all emotions under the few-shot setting. In contrast, the open-source models exhibit substantial variability, indicating differential sensitivity to contextual cues. Overall, these findings highlight the robustness of ChatGPT-5 for IER, the

general effectiveness of few-shot prompting for inferring subtle emotional signals, and the current limitations of open-source LLMs when explicit lexical cues are absent.

Approach	Model	Joy	Sadness	Anger	Fear	Disgust	Surprise
Fine-tuned	BERT	70.21	56.28	51.52	71.11	59.18	60.30
	RoBERTa	**76.95**	**64.47**	**58.29**	**78.00**	**61.85**	59.98
Zero-shot	ChatGPT-5	73.50	55.90	50.78	70.50	58.63	56.97
	LLaMA-3	50.89	39.51	40.36	28.35	18.80	43.00
	Phi-4	52.17	43.48	40.66	48.32	32.32	36.94
	Gemma-3	47.73	45.07	36.90	44.30	30.67	27.69
Few-shot	ChatGPT-5	73.50	58.00	54.12	72.40	60.63	**60.46**
	LLaMA-3	49.15	39.39	39.39	41.79	34.29	41.84
	Phi-4	61.62	40.66	34.83	54.88	41.10	45.76
	Gemma-3	53.63	42.65	39.22	52.87	37.42	46.86

Table 6.13: Per-emotion Result on the IEST dataset.

6.3.4 *Conclusion*

In this study, we conducted a comprehensive evaluation of language models for the task of IER in english text, examining two different architectures: medium-sized masked language models, namely BERT and RoBERTa, and prompting-based LLMs, including ChatGPT-5, Phi-4, LLaMA-3, and Gemma-3, under zero-shot and few-shot settings. The models were assessed on two widely used datasets, IEST and ISEAR, which differ significantly in terms of domain, format, and the nature of emotional expression.

Our findings indicate that fine-tuned masked language models, particularly RoBERTa, remain highly competitive and excel at capturing implicit emotional cues when task-specific supervision is available. Among prompting-based models, ChatGPT-5 demonstrated the strongest reasoning capabilities, effectively inferring emotions from contextual cues, with few-shot prompting yielding consistent performance gains across both datasets. However, performance disparities between the two benchmarks underscore the critical role of text quality and structure: coherent narrative texts in ISEAR allowed for more accurate emotion inference, while the fragmented and noisy nature of social media content in the IEST dataset posed substantial challenges for all models.

Taken together, these results highlight both the promise and the current limitations of LLMs for IER. While prompting-based models show encouraging potential, particularly for positive emotions, they still fall short of fine-tuned models in complex and context-dependent cases. Progress in IER remains constrained by the scarcity of high-quality datasets, as building datasets that capture implicit nuance without relying on explicit markers is inherently challenging. Despite its limitations, the IEST dataset serves as a practical proxy by simulating implicitness through masked emotion words, offering a controlled evaluation setting.

6.4 CONCLUSION AND CROSS-LINGUISTIC INSIGHTS

This chapter presents a two-part investigation into the capacity of LLMs for implicit emotional reasoning, examined across two distinct linguistic settings. the study I addressed the low-resource Algerian Arabic dialect, evaluating both fine-tuned PLMs and zero-shot LLMs on the challenging task of multi-label emotion recognition in a dialectal context characterized by morphological richness and culturally embedded emotional expression. Study II extended the analysis to English, benchmarking a broader set of state-of-the-art LLMs under zero-shot and few-shot prompting approaches. The models were assessed on two widely used datasets, IEST and ISEAR, which differ significantly in terms of domain, format, and the nature of implicit emotional expression.

The results across both studies reveal several consistent and cross-linguistically stable findings. First, zero-shot LLMs demonstrated competitive and sometimes superior performance compared to fine-tuned models, particularly in low-resource settings and on well-structured datasets, confirming that large-scale pretraining endows these models with transferable reasoning capabilities. However, this advantage proved neither universal nor robust, as fine-tuned models such as RoBERTa retained a clear edge on noisier and more challenging text. Second, few-shot prompting emerged as a reliable strategy for improving implicit emotional reasoning across implicit emotional scenarios and model families, offering a practical direction for enhancing LLMs performance without task-specific fine-tuning. Third, and most importantly, both studies converge on a shared conclusion: the limitations of current LLMs in IER cannot be attributed only to data scarcity or linguistic unfamiliarity

but reflect a deeper and more fundamental deficit in emotional reasoning that persists across languages, model sizes, and prompting strategies.

These findings collectively suggest that while LLMs have made significant strides in natural language understanding, IER remains an open challenge. Bridging this gap will require not only larger and more capable models, but also richer and more nuanced training resources, more sophisticated prompting strategies, and evaluation frameworks that better capture the complexity of human emotional expression across languages and cultures.

7

GENERAL CONCLUSION

Implicit emotion recognition involves identifying subtle emotions through contextual cues rather than explicit emotional expressions. This task holds significant value for real-world applications, such as improving user experience, service quality, and providing appropriate responses to customer interactions. Despite notable advancements in emotion recognition, IER remains an underexplored area, particularly for low-resource languages such as Arabic.

This thesis has addressed the problem of Implicit Emotion Recognition (IER) in Arabic text by framing it as a multidimensional research problem spanning resource development, modeling paradigms, and linguistic variability. The study was motivated by two converging observations: first, that implicit emotional expression, conveyed not through explicit affective vocabulary but through contextual cues, figurative language, and cultural convention, remains a persistent and underexplored challenge in affective Natural Language Processing (NLP); and second, that Arabic, despite its global presence, continues to be underserved by emotion recognition research, particularly with respect to its rich dialectal variation and the complex linguistic characteristics that both shape implicit emotion expression and complicate its automatic recognition. Together, these observations defined a research framework pursued across three complementary scientific contributions.

The first contribution addressed the foundational challenge of resource scarcity by introducing the Arabic Implicit Emotion Recognition (AIER) dataset, a publicly available annotated benchmark composed of real-world Arabic reviews in Modern Standard Arabic (MSA), specifically designed to model emotional meaning inferred from contextual cues rather than explicit emotion-bearing words. The dataset was annotated with emotion roles, including emotion label, cues, and emotion cause, and was evaluated using four computational approaches, including classical Machine Learning (ML), Deep Learning (DL) architec-

tures, Pretrained Language Models (PLMs), and generative Large Language Models (LLMs). The experimental results demonstrated that contextualized transformer-based models consistently outperform traditional approaches, while also revealing persistent challenges in capturing subtle and nuanced implicit emotional cues, findings that underscored the difficulty of the task and motivated the modeling work pursued in the subsequent contribution.

The second contribution responded to these challenges by proposing a hybrid contextual–sequential architecture that integrates the deep bidirectional encoding capacity of AraBERT with the temporal sequential modeling capabilities of a Bidirectional Gated Recurrent Units (BiGRU) network. Evaluated on a newly compiled Arabic Twitter corpus under both implicit and explicit emotion recognition settings, the proposed model consistently outperformed established deep learning baselines, with quantitative and qualitative analyses confirming its effectiveness in capturing context-dependent and subtly expressed emotional cues.

The third contribution broadened the scope of the investigation by examining the emotional reasoning capabilities of state-of-the-art LLMs across two complementary settings. The first study evaluated LLMs performance on multi-label emotion recognition in the Algerian Arabic dialect, a critically low-resource variety characterized by morphological richness and near-total absence from existing NLP benchmarks. The second study extended the analysis to English text, benchmarking the implicit emotional reasoning of LLMs in a high-resource setting where direct lexical cues are deliberately absent, thereby isolating deeper semantic understanding from surface-level pattern exploitation. Together, the two studies provided a rigorous cross-lingual assessment of LLMs capabilities and limitations in affective understanding, revealing both the promising generalization potential and the persistent shortcomings of current models when confronted with linguistically diverse and emotionally complex textual contexts.

Taken together, the findings of this thesis lead to several important observations. First, IER in Arabic is a truly challenging but tractable problem, one that responds meaningfully to dedicated resource development and architecturally informed modeling strategies. The results consistently show that models which integrate contextual and sequential representations outperform those that rely on either dimension alone, suggesting that the nature of implicit emotional meaning requires equally layered computational treatment.

Second, the performance of LLMs on emotion recognition tasks is promising but uneven. While LLMs demonstrate a degree of generalization capacity across languages and dialects, their performance degrades considerably in low-resource and dialectal settings where cultural specificity, morphological complexity, and data scarcity combine to create conditions that current models are not fully equipped to handle. These findings suggest that LLMs, despite their impressive capabilities, cannot yet be relied upon as general-purpose instruments for affective understanding across the full linguistic diversity of Arabic.

Third, the scarcity of dedicated annotated resources remains one of the most significant bottlenecks in Arabic affective NLP. The construction of the AIER dataset represents a meaningful step toward addressing this gap, but the findings also highlight the need for broader, more diverse, and more linguistically representative benchmarks — particularly for dialectal varieties and for tasks that require pragmatic and cultural annotation beyond simple emotion labeling.

Despite these advances, the work presented in this thesis is not without limitations. The development of larger, more diverse, and more linguistically representative datasets for Arabic IER, including multiple dialects, domains, and annotation schemes, remains a priority. Future work should also explore multimodal approaches that integrate textual analysis with acoustic, visual, or contextual signals, which may provide additional cues for implicit emotion inference. The interpretability of LLMs deserves greater attention, as understanding why a model assigns a particular emotion label is as important as achieving high classification accuracy. Finally, the cross-lingual transfer of IER capabilities, from high-resource to low-resource languages and dialects, represents a rich and largely unexplored research frontier, which will enable important scientific and practical applications.

In conclusion, this thesis has demonstrated that implicit emotion recognition in Arabic is a problem that is both scientifically rich and practically significant, one that sits at the intersection of linguistics, culture, and computation, and that rewards rigorous, resource-conscious, and architecturally thoughtful research. By introducing dedicated resources, proposing effective modeling strategies, and critically evaluating the generalization capacity of state-of-the-art models, this work contributes a meaningful step forward in the broader effort to build NLP systems that are truly capable of understanding the full emotional depth of human language, across its many varieties, levels, and cultural expressions.

BIBLIOGRAPHY

[1] F. Senator, A. Lakhfif, I. Zenbout, H. Boutouta, and C. Mediani, "Leveraging chatgpt for enhancing arabic nlp: Application for semantic role labeling and cross-lingual annotation projection", *IEEE Access*, vol. 13, pp. 3707–3725, 2025. DOI: 10.1109/ACCESS.2025.3525493.

[2] F. Senator, H. Boutouta, A. Lakhfif, and C. Mediani, "Semantic role labeling of arabic emotional text in tweets", in *2023 International Conference on Advances in Electronics, Control and Communication Systems (ICAECCS)*, 2023, pp. 1–6. DOI: 10.1109/ICAECCS56710.2023.10104732.

[3] R. W. Picard, *Affective Computing* (Inteligencia artificial). MIT Press, 1997, ISBN: 9780262161701. [Online]. Available: https://books.google.dz/books?id=DpZ-ngEACAAJ.

[4] B. Pang and L. Lee, "Opinion mining and sentiment analysis", *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, Jul. 2008, ISSN: 1554-0669. DOI: 10.1561/1500000011. eprint: https://www.emerald.com/ftinr/article-pdf/2/1-2/1/10912242/1500000011en.pdf. [Online]. Available: https://doi.org/10.1561/1500000011.

[5] B. Liu, *Sentiment Analysis and Opinion Mining* (Synthesis digital library of engineering and computer science). Morgan & Claypool, 2012, ISBN: 9781608458844. [Online]. Available: https://books.google.dz/books?id=Gt8g72e6MuEC.

[6] P. Ekman, "An argument for basic emotions", *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[7] C. Strapparava and R. Mihalcea, "SemEval-2007 task 14: Affective text", in *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, E. Agirre, L. Màrquez, and R. Wicentowski, Eds., Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 70–74. [Online]. Available: https://aclanthology.org/S07-1013/.

[8] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: Machine learning for text-based emotion prediction", in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, R. Mooney, C. Brew, L.-F. Chien, and K. Kirchhoff, Eds., Vancouver, British Columbia, Canada: Association for Computational Linguistics, Oct. 2005, pp. 579–586. [Online]. Available: https://aclanthology.org/H05-1073/.

[9] M. Abdul-Mageed and L. Ungar, "EmoNet: Fine-grained emotion detection with gated recurrent neural networks", in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, R. Barzilay and M.-Y. Kan, Eds., Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 718–728. DOI: 10.18653/v1/P17-1067. [Online]. Available: https://aclanthology.org/P17-1067/.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://aclanthology.org/N19-1423/.

[11] X. Hong, Y. Gong, V. Sethu, and T. Dang, "Aer-llm: Ambiguity-aware emotion recognition leveraging large language models", in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10888198.

[12] S. Ghosh, S. Sahu, N. Ganguly, B. Mitra, and P. De, "Emokey: An emotion-aware smartphone keyboard for mental health monitoring", in *2019 11th international conference on communication systems & networks (COMSNETS)*, IEEE, 2019, pp. 496–499.

[13] A. Parkavi, T. B. N. Shetty, S. G. Shibu, R. Ismail, and R. A. Muthirakalayil, "Customer feedback analysis based on emotion detection using machine learning techniques with privacy preservation", in *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)*, 2023, pp. 1617–1624. DOI: 10.1109/ICCPCT58313.2023.10244876.

[14] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction", *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001. DOI: 10.1109/79.911197.

[15] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing", in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, L.-W. Ku and C.-T. Li, Eds., Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1–10. DOI: 10.18653/v1/W17-1101. [Online]. Available: https://aclanthology.org/W17-1101/.

[16] A. Balahur, J. M. Hermida, and A. Montoyo, "Detecting implicit expressions of emotion in text: A comparative analysis", *Decision support systems*, vol. 53, no. 4, pp. 742–753, 2012.

[17] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon", *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013. DOI: https://doi.org/10.1111/j.1467-8640.2012.00460.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8640.2012.00460.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8640.2012.00460.x.

[18] Y. P. H. Lau et al., "Decoding emotions in social media: A linguistic analysis of implicit emotions and events", Ph.D. dissertation, 2021.

[19] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Harnessing twitter "big data" for automatic emotion identification", in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 2012, pp. 587–592. DOI: 10.1109/SocialCom-PASSAT.2012.119.

[20] J. S. Y. Liew and H. R. Turtle, "Exploring fine-grained emotion detection in tweets", in *Proceedings of the NAACL Student Research Workshop*, J. Andreas, E. Choi, and A. Lazaridou, Eds., San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 73–80. DOI: 10.18653/v1/N16-2011. [Online]. Available: https://aclanthology.org/N16-2011/.

[21] A. Seyeditabari, S. Levens, C. D. Maestas, S. Shaikh, J. I. Walsh, W. Zadrozny, C. Danis, and O. P. Thompson, "Cross corpus emotion classification using survey data", *This paper was presented at AISB*, 2017.

[22] U. Orizu, "Implicit emotion detection in text", Ph.D. dissertation, Aston University, 2018.

[23] Z. Sadiqzade, "The linguistic expression of emotion: A cross-cultural analysis", *EuroGlobal Journal of Linguistics and Language Education*, vol. 2, no. 3, pp. 42–54, 2025. DOI: https://doi.org/10.69760/egjlle.2500195.

[24] E. Kim and R. Klinger, "Who feels what and why? annotation of a literature corpus with semantic roles of emotions", in *Proceedings of the 27th International Conference on Computational Linguistics*, E. M. Bender, L. Derczynski, and P. Isabelle, Eds., Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1345–1359. [Online]. Available: https://aclanthology.org/C18-1114/.

[25] D. Ghazi, D. Inkpen, and S. Szpakowicz, "Prior versus contextual emotion of a word in a sentence", in *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, A. Balahur, A. Montoyo, P. M. Barco, and E. Boldrini, Eds., Jeju, Korea: Association for Computational Linguistics, Jul. 2012, pp. 70–78. [Online]. Available: https://aclanthology.org/W12-3711/.

[26] Eberhard, David M. and Simons, Gary F. and Fennig, Charles D., *Ethnologue: Languages of the world — arabic*, Accessed: 2026-05-03, 2024. [Online]. Available: https://www.ethnologue.com.

[27] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions", *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 8, no. 4, pp. 1–22, 2009.

[28] N. Habash, *Introduction to Arabic Natural Language Processing* (Synthesis digital library of engineering and computer science). Morgan & Claypool Publishers, 2010, ISBN: 9781598297959. [Online]. Available: https://books.google.dz/books?id=kRIHCnC74BoC.

[29] C. A. Ferguson, "Diglossia", *WORD*, vol. 15, no. 2, pp. 325–340, 1959. DOI: 10.1080/00437956.1959.11659702. eprint: https://doi.org/10.1080/00437956.1959.11659702. [Online]. Available: https://doi.org/10.1080/00437956.1959.11659702.

[30] A. M. Alayba, "Arabic natural language processing (nlp): A comprehensive review of challenges, techniques, and emerging trends", *Computers*, vol. 14, no. 11, 2025, ISSN: 2073-431X. DOI: 10.3390/computers14110497. [Online]. Available: https://www.mdpi.com/2073-431X/14/11/497.

[31] K. Shaalan, "A survey of Arabic named entity recognition and classification", *Computational Linguistics*, vol. 40, no. 2, pp. 469–510, Jun. 2014. DOI: 10.1162/COLI_a_00178. [Online]. Available: https://aclanthology.org/J14-2008/.

[32] O. Oueslati, E. Cambria, M. B. HajHmida, and H. Ounelli, "A review of sentiment analysis research in arabic language", *Future Generation Computer Systems*, vol. 112, pp. 408–430, 2020.

[33] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 task 1: Affect in tweets", in *Proceedings of the 12th International Workshop on Semantic Evaluation*, M. Apidianaki, S. M. Mohammad, J. May, E. Shutova, S. Bethard, and M. Carpuat, Eds., New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1–17. DOI: 10.18653/v1/S18-1001. [Online]. Available: https://aclanthology.org/S18-1001/.

[34] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding", eng, in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, H. Al-Khalifa, W. Magdy, K. Darwish, T. Elsayed, and H. Mubarak, Eds., Marseille, France: European Language Resources Association, May 2020, pp. 9–15, ISBN: 979-10-95546-51-1. [Online]. Available: https://aclanthology.org/2020.osact-1.2/.

[35] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in Arabic pre-trained language models", in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, N. Habash, H. Bouamor, H. Hajj, W. Magdy, W. Zaghouani, F. Bougares, N. Tomeh, I. Abu Farha, and S. Touileb, Eds., Kyiv, Ukraine (Virtual): Association for Computational Linguistics, Apr. 2021, pp. 92–104. [Online]. Available: https://aclanthology.org/2021.wanlp-1.10/.

[36] K. R. Scherer, "What are emotions? and how can they be measured?", *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[37] J. J. Gross, "The emerging field of emotion regulation: An integrative review", *Review of General Psychology*, vol. 2, no. 3, pp. 271–299, 1998. DOI: 10.1037/1089-2680.2.3.271.

[38] R. Plutchik, *Emotion, a Psychoevolutionary Synthesis*. Harper & Row, 1980, ISBN: 9780060452353. [Online]. Available: https://books.google.dz/books?id=G5t9AAAAMAAJ.

[39] R. Lazarus, *Emotion and Adaptation*. Oxford University Press, 1991, ISBN: 9780195092660. [Online]. Available: https://books.google.dz/books?id=tTdIlwpxtWsC.

[40] S. M. Mohammad, "9 - sentiment analysis: Detecting valence, emotions, and other affectual states from text", in *Emotion Measurement*, H. L. Meiselman, Ed., Woodhead Publishing, 2016, pp. 201–237, ISBN: 978-0-08-100508-8. DOI: https://doi.org/10.1016/B978-0-08-100508-8.00009-6. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780081005088000096.

[41] R. Klinger, O. De Clercq, S. Mohammad, and A. Balahur, "IEST: WASSA-2018 implicit emotions shared task", in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, A. Balahur, S. M. Mohammad, V. Hoste, and R. Klinger, Eds., Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 31–42. DOI: 10.18653/v1/W18-6206. [Online]. Available: https://aclanthology.org/W18-6206/.

[42] P. EKMAN and W. V. FRIESEN, *The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding*, A. Kendon, Ed. Berlin, Boston: De Gruyter Mouton, 1981, pp. 57–106, ISBN: 9783110880021. DOI: doi:10.1515/9783110880021.57. [Online]. Available: https://doi.org/10.1515/9783110880021.57.

[43] R. Plutchik, "A psychoevolutionary theory of emotions", *Social Science Information*, vol. 21, no. 4-5, pp. 529–553, 1982. DOI: 10.1177/053901882021004003. eprint: https://doi.org/10.1177/053901882021004003. [Online]. Available: https://doi.org/10.1177/053901882021004003.

[44] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions", *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, 1977, ISSN: 0092-6566. DOI: https://doi.org/10.1016/0092-6566(77)90037-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/009265667790037X.

[45] K. R. Scherer, *Appraisal Theory*. John Wiley Sons, Ltd, 1999, ch. 30, pp. 637–663, ISBN: 9780470013496. DOI: https://doi.org/10.1002/0470013494.ch30. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470013494.ch30. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013494.ch30.

[46] R. S. Lazarus, "Progress on a cognitive-motivational-relational theory of emotion.", *The American psychologist*, vol. 46 8, pp. 819–34, 1991. [Online]. Available: https://api.semanticscholar.org/CorpusID:45157136.

[47] C. A. Smith and P. C. Ellsworth, "Patterns of cognitive appraisal in emotion.", *Journal of personality and social psychology*, vol. 48 4, pp. 813–38, 1985. [Online]. Available: https://api.semanticscholar.org/CorpusID:35092703.

[48] M. Abulaish and A. Bhardwaj, "Rolemo: A role-aware commonsense-augmented contrastive learning framework for emotion classification", *Machine Learning and Knowledge Extraction*, vol. 8, no. 3, 2026, ISSN: 2504-4990. DOI: 10.3390/make8030079. [Online]. Available: https://www.mdpi.com/2504-4990/8/3/79.

[49] W. Bouchelligua, R. Al-Dayil, and A. Algaith, "Effective data augmentation techniques for arabic speech emotion recognition using convolutional neural networks", *Applied Sciences*, vol. 15, no. 4, 2025, ISSN: 2076-3417. DOI: 10.3390/app15042114. [Online]. Available: https://www.mdpi.com/2076-3417/15/4/2114.

[50] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, "Arabic natural language processing: An overview", *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 5, pp. 497–507, 2021, ISSN: 1319-1578. DOI: https://doi.org/10.1016/j.jksuci.2019.02.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157818310553.

[51] A. Alsayat and N. Elmitwally, "A comprehensive study for arabic sentiment analysis (challenges and applications)", *Egyptian Informatics Journal*, vol. 21, no. 1, pp. 7–12, 2020, ISSN: 1110-8665. DOI: https://doi.org/10.1016/j.eij.2019.06.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110866519300945.

[52] J. Islam, R. E. Mercer, and L. Xiao, "Multi-channel convolutional neural network for Twitter emotion and sentiment recognition", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1355–1365. DOI: 10.18653/v1/N19-1137. [Online]. Available: https://aclanthology.org/N19-1137/.

[53] J. Eisenstein, "What to do about bad language on the internet", in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, L. Vanderwende, H. Daumé III, and K. Kirchhoff, Eds., Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 359–369. [Online]. Available: https://aclanthology.org/N13-1037/.

[54] A. Ortony, G. L. Clore, and M. A. Foss, "The referential structure of the affective lexicon", *Cognitive Science*, vol. 11, no. 3, pp. 341–364, 1987, ISSN: 0364-0213. DOI: https://doi.org/10.1016/S0364-0213(87)80010-1. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0364021387800101.

[55] A. Pavlenko, "Emotion and emotion-laden words in the bilingual lexicon", *Bilingualism: Language and Cognition*, vol. 11, no. 2, 147–164, 2008. DOI: 10.1017/S1366728908003283.

[56] E. A. H. Khalil, E. M. E. Houby, and H. K. Mohamed, "Deep learning for emotion analysis in arabic tweets", *Journal of Big Data*, vol. 8, no. 1, p. 136, 2021. DOI: https://doi.org/10.1186/s40537-021-00523-w.

[57] R. S. Alsulaiman and A. I. Alhojailan, "Emojis as tools for learning: Understanding their pragmatic functions in efl students' tweets.", *Theory & Practice in Language Studies (TPLS)*, vol. 14, no. 12, 2024.

[58] L. Zhu, J. Peng, and H. Zhang, "Text-based sarcasm detection with emoji contradictory clues assisting", in *Advanced Intelligent Computing Technology and Applications*, D.-S. Huang, Y. Pan, W. Chen, and B. Li, Eds., Singapore: Springer Nature Singapore, 2025, pp. 248–260, ISBN: 978-981-96-9986-5.

[59] A. Hassan, N. A. Helal, Y. M. Afify, and N. L. Badr, "Exploring emotional cues within conversational contexts for sarcasm detection", in *2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, 2024, pp. 127–130. DOI: 10.1109/NILES63360.2024.10753268.

[60] A. Rahma, S. S. Azab, and A. Mohammed, "A comprehensive survey on arabic sarcasm detection: Approaches, challenges and future trends", *IEEE Access*, vol. 11, pp. 18 261–18 280, 2023. DOI: 10.1109/ACCESS.2023.3247427.

[61] G. Lakoff and M. Johnson, *Metaphors We Live By*. University of Chicago Press, 2008, ISBN: 9780226470993. [Online]. Available: https://books.google.dz/books?id=r6nOYYtxzUoC.

[62] J. Karoui, F. Benamara, V. Moriceau, V. Patti, C. Bosco, and N. Aussenac-Gilles, "Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study", in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, M. Lapata, P. Blunsom, and A. Koller, Eds., Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 262–272. [Online]. Available: https://aclanthology.org/E17-1025/.

[63] X. Chen, Z. Hai, D. Li, S. Wang, and D. Wang, "Jointly identifying rhetoric and implicit emotions via multi-task learning", in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Online: Association for Computational Linguistics, Aug. 2021, pp. 1429–1434. DOI: 10.18653/v1/2021.findings-acl.123. [Online]. Available: https://aclanthology.org/2021.findings-acl.123/.

[64] X. Chen, S. Wang, X. Li, Z. Hai, Y. Li, D. Li, and J. Cai, "Identifying implicit emotions via hierarchical structure and rhetorical correlation", *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 11, pp. 3753–3764, 2023. DOI: https://doi.org/10.1007/s13042-023-01862-1.

[65] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", *arXiv preprint arXiv:1301.3781*, 2013.

[66] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information", *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, Jun. 2017, ISSN: 2307-387X. DOI: 10.1162/tacl_a_00051. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00051/1567442/tacl_a_00051.pdf. [Online]. Available: https://doi.org/10.1162/tacl_a_00051.

[67] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages", in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. Calzolari et al., Eds., Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://aclanthology.org/L18-1550/.

[68] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need", vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[69] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., "Improving language understanding by generative pre-training", 2018.

[70] C. Cortes and V. Vapnik, "Support-vector networks", *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995. DOI: https://doi.org/10.1007/BF00994018.

[71] T. Cover and P. Hart, "Nearest neighbor pattern classification", *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[72] J. R. Quinlan, "Induction of decision trees", *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986. DOI: https://doi.org/10.1007/BF00116251.

[73] L. Breiman, "Random forests", *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: https://doi.org/10.1023/A:1010933404324.

[74] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system", in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794. DOI: https://doi.org/10.1145/2939672.293978.

[75] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (Adaptive Computation and Machine Learning series). MIT Press, 2016, ISBN: 9780262035613. [Online]. Available: https://books.google.dz/books?id=Np9SDQAAQBAJ.

[76] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[77] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. DOI: 10.1109/5.726791.

[78] Y. Kim, "Convolutional neural networks for sentence classification", in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds., Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. DOI: 10.3115/v1/D14-1181. [Online]. Available: https://aclanthology.org/D14-1181/.

[79] M. Baali and N. Ghneim, "Emotion analysis of arabic tweets using deep learning approach", *Journal of Big Data*, vol. 6, no. 1, p. 89, 2019.

[80] A. Graves, "Long short-term memory", in *Supervised Sequence Labelling with Recurrent Neural Networks*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 37–45, ISBN: 978-3-642-24797-2. DOI: 10.1007/978-3-642-24797-2_4. [Online]. Available: https://doi.org/10.1007/978-3-642-24797-2_4.

[81] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures", *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[82] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model.", in *Interspeech*, Makuhari, vol. 2, 2010, pp. 1045–1048.

[83] K. Cho, "On the properties of neural machine translation: Encoder-decoder approaches", *arXiv preprint arXiv:1409.1259*, 2014.

[84] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate", *CoRR*, vol. abs/1409.0473, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:11212020.

[85] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach", *arXiv preprint arXiv:1907.11692*, 2019.

[86] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for Arabic", in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 7088–7105. DOI: 10.18653/v1/2021.acl-long.551. [Online]. Available: https://aclanthology.org/2021.acl-long.551.

[87] A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. Samih, "Pre-training bert on arabic tweets: Practical considerations", 2021. arXiv: 2102.10684 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2102.10684.

[88] W. Antoun, F. Baly, and H. Hajj, "AraELECTRA: Pre-training text discriminators for Arabic language understanding", in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Virtual): Association for Computational Linguistics, Apr. 2021, pp. 191–195. [Online]. Available: https://www.aclweb.org/anthology/2021.wanlp-1.20.

[89] A. Abdaoui, M. Berrimi, M. Oussalah, and A. Moussaoui, "Dziribert: A pre-trained language model for the algerian dialect", 2022. arXiv: 2109.12346 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2109.12346.

[90] Z. Benmounah, A. Boulesnane, A. Fadheli, and M. Khial, "Sentiment analysis on algerian dialect with transformers", *Applied Sciences*, vol. 13, no. 20, p. 11 157, 2023, ISSN: 2076-3417. DOI: 10.3390/app132011157. [Online]. Available: http://dx.doi.org/10.3390/app132011157.

[91] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale", *CoRR*, vol. abs/1911.02116, 2019. arXiv: 1911.02116. [Online]. Available: http://arxiv.org/abs/1911.02116.

[92] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language models are few-shot learners", *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[93] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., "Chain-of-thought prompting elicits reasoning in large language models", *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[94] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., "Gpt-4 technical report", *arXiv preprint arXiv:2303.08774*, 2023.

[95] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al., "Gemini: A family of highly capable multimodal models", *arXiv preprint arXiv:2312.11805*, 2023.

[96] A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Dengr, C. Ruan, D. Dai, D. Guo, et al., "Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model", *arXiv preprint arXiv:2405.04434*, 2024.

[97] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., "The llama 3 herd of models", *arXiv preprint arXiv:2407.21783*, 2024.

[98] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al., "Gemma: Open models based on gemini research and technology", *arXiv preprint arXiv:2403.08295*, 2024.

[99] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, et al., "Phi-4 technical report", *arXiv preprint arXiv:2412.08905*, 2024.

[100] Cohere. "Introducing command r7b arabic". [Online]. Available: https://cohere.com/blog/command-r7b-arabic.

[101] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing", *ACM computing surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[102] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners", *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[103] S. M. Mohammad and P. D. Turney, "Nrc emotion lexicon", *National Research Council, Canada*, vol. 2, p. 234, 2013.

[104] G. Badaro, R. Baly, H. Hajj, N. Habash, and W. El-Hajj, "A large scale Arabic sentiment lexicon for Arabic opinion mining", in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, N. Habash and S. Vogel, Eds., Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 165–173. DOI: 10.3115/v1/W14-3623. [Online]. Available: https://aclanthology.org/W14-3623/.

[105] M. Abdul-Mageed, H. AlHuzli, M. D. D. Elhija, M Diab, and M. Duaa'Abu Elhija, "Dina: A multidialect dataset for arabic emotion analysis", in *The 2nd workshop on Arabic corpora and processing tools*, 2016, p. 29.

[106] H. Alhuzali, M. Abdul-Mageed, and L. Ungar, "Enabling deep learning of emotion with first-person seed expressions", in *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, 2018, pp. 25–35.

[107] A. Al-Khatib and S. R. El-Beltagy, "Emotional tone detection in arabic tweets", in *Computational Linguistics and Intelligent Text Processing: 18th International Conference, CICLing 2017, Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part II 18*, Springer, 2018, pp. 105–114.

[108] Q. Yang, X. Chen, C. Ma, R. Yin, X. Gao, and X. Zhang, "Senwave: A fine-grained multi-language sentiment analysis dataset sourced from covid-19 tweets", *arXiv preprint arXiv:2510.08214*, 2025.

[109] A. Al-Laith and M. Alenezi, "Monitoring people's emotions and symptoms from arabic tweets during the covid-19 pandemic", *Information*, vol. 12, no. 2, p. 86, 2021.

[110] A. J. Almahdawi and W. J. Teahan, "A new arabic dataset for emotion recognition", in *Intelligent Computing: Proceedings of the 2019 Computing Conference, Volume 2*, Springer, 2019, pp. 200–216.

[111] E. C. O. Alm, *Affect in text and speech*. University of Illinois at Urbana-Champaign, 2008. [Online]. Available: https://books.google.dz/books?id=zYQyjwEACAAJ.

[112] S. Aman and S. Szpakowicz, "Identifying expressions of emotion in text", in *International conference on text, speech and dialogue*, Springer, 2007, pp. 196–205.

[113] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal, "SemEval-2019 task 3: EmoContext contextual emotion detection in text", in *Proceedings of the 13th International Workshop on Semantic Evaluation*, J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, and S. M. Mohammad, Eds., Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 39–48. DOI: 10.18653/v1/S19-2005. [Online]. Available: https://aclanthology.org/S19-2005/.

[114] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions", *arXiv preprint arXiv:2005.00547*, 2020.

[115] S. Buechel and U. Hahn, "Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis", *arXiv preprint arXiv:2205.01996*, 2022.

[116] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning.", *Journal of personality and social psychology*, vol. 66, no. 2, p. 310, 1994.

[117] D. Li, J. Wang, and X. Zhang, "Ciea: A corpus for chinese implicit emotion analysis", in *2019 International Conference on Asian Language Processing (IALP)*, IEEE, 2019, pp. 90–95.

[118] O. Udochukwu and Y. He, "A rule-based approach to implicit emotion detection in text", in *Natural Language Processing and Information Systems*, C. Biemann, S. Handschuh, A. Freitas, F. Meziane, and E. Métais, Eds., Cham: Springer International Publishing, 2015, pp. 197–203, ISBN: 978-3-319-19581-0.

[119] N. Alswaidan and M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text", *Knowledge and Information Systems*, vol. 62, no. 8, pp. 2937–2987, 2020.

[120] A. Balahur, J. M. Hermida, and A. Montoyo, "Detecting implicit expressions of sentiment in text based on commonsense knowledge", in *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis (WASSA 2.011)*, 2011, pp. 53–60.

[121] N. Riahi and P. Safari, "Implicit emotion detection from text with information fusion", *Journal of Advances in Computer Research*, vol. 7, no. 2, pp. 85–99, 2016.

[122] F. Khoshnam and A. Baraani-Dastjerdi, "A dual framework for implicit and explicit emotion recognition: An ensemble of language models and computational linguistics", *Expert Systems with Applications*, vol. 198, p. 116 686, 2022.

[123] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions* (Cambridge books online). Cambridge University Press, 1988, ISBN: 9780521353649. [Online]. Available: https://books.google.dz/books?id=Sp8FngEACAAJ.

[124] Y. Zhao and Y. Zhang, "Research on implicit emotion recognition and classification in literary works in the context of machine learning", *Alexandria Engineering Journal*, vol. 115, pp. 577–584, 2025, ISSN: 1110-0168. DOI: https://doi.org/10.1016/j.aej.2024.12.045. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110016824016430.

[125] A. Rozental, D. Fleischer, and Z. Kelrich, "Amobee at iest 2018: Transfer learning from language models", *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2018. DOI: 10.18653/v1/w18-6207. [Online]. Available: http://dx.doi.org/10.18653/v1/W18-6207.

[126] J. Balazs, E. Marrese-Taylor, and Y. Matsuo, "IIIDYT at IEST 2018: Implicit emotion classification with deep contextualized word representations", in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 50–56. DOI: 10.18653/v1/W18-6208. [Online]. Available: https://aclanthology.org/W18-6208.

[127] A. Chronopoulou, A. Margatina, C. Baziotis, and A. Potamianos, "NTUA-SLP at IEST 2018: Ensemble of neural transfer methods for implicit emotion classification", in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 57–64. DOI: 10.18653/v1/W18-6209. [Online]. Available: https://aclanthology.org/W18-6209.

[128] P. Rathnayaka, S. Abeysinghe, C. Samarajeewa, I. Manchanayake, and M. Walpola, "Sentylic at iest 2018: Gated recurrent neural network and capsule network-based approach for implicit emotion detection", *arXiv preprint arXiv:1809.01452*, 2018.

[129] Q. Zhou and H. Wu, "Nlp at iest 2018: Bilstm-attention and lstm-attention via soft voting in emotion classification", in *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 2018, pp. 189–194.

[130] W. Witon, P. Colombo, A. Modi, and M. Kapadia, "Disney at iest 2018: Predicting emotions using an ensemble", in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2018, pp. 248–253.

[131] S. Pecar, M. Farkas, M. Šimko, P. Lacko, and M. Bielikova, "Nl-fiit at iest-2018: Emotion recognition utilizing neural networks and multi-level preprocessing", in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2018, pp. 217–223.

[132] H. Fei, Y. Ren, and D. Ji, "Implicit objective network for emotion detection", in *CCF International Conference on Natural Language Processing and Chinese Computing*, Springer, 2019, pp. 647–659.

[133] F. Khoshnam, A. Baraani-Dastjerdi, and M. Liaghatdar, "Cefer: A four facets framework based on context and emotion embedded features for implicit and explicit emotion recognition", *arXiv preprint arXiv:2209.13999*, 2022.

[134] Y. Qian, J. Wang, D. Li, and X. Zhang, "Interactive capsule network for implicit sentiment analysis", *Applied Intelligence*, vol. 53, no. 3, pp. 3109–3123, 2023.

[135] X. Feng, Q. Sun, Y. Wang, and Y. Qiao, "Enhancing fine-grained sentiment analysis for implicit texts on social media", *Computer Speech Language*, vol. 96, p. 101 867, 2026, ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2025.101867. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230825000920.

[136] H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, "Enhancement of implicit emotion recognition in arabic text: Annotated dataset and baseline models", *IEEE Access*, vol. 13, pp. 165 096–165 116, 2025. DOI: 10.1109/ACCESS.2025.3611337.

[137] M. J. Althobaiti, "An open-source dataset for arabic fine-grained emotion recognition of online content amid covid-19 pandemic", *Data in Brief*, vol. 51, p. 109 745, 2023.

[138] S. Shahriar, N. Al Roken, and I. Zualkernan, "Classification of arabic poetry emotions using deep learning", *Computers*, vol. 12, no. 5, p. 89, 2023.

[139] A. Elnagar, Y. S. Khalifa, and A. Einea, "Hotel arabic-reviews dataset construction for sentiment analysis applications", in *Intelligent Natural Language Processing: Trends and Applications*, K. Shaalan, A. E. Hassanien, and F. Tolba, Eds. Cham: Springer International Publishing, 2018, pp. 35–52, ISBN: 978-3-319-67056-0. DOI: 10.1007/978-3-319-67056-0_3. [Online]. Available: https://doi.org/10.1007/978-3-319-67056-0_3.

[140] A. Elnagar and O. Einea, "Brad 1.0: Book reviews in arabic dataset", Nov. 2016, pp. 1–8. DOI: 10.1109/AICCSA.2016.7945800.

[141] H. Elsahar and S. El-Beltagy, "Building large arabic multi-domain resources for sentiment analysis", *Lecture Notes in Computer Science*, vol. 9042, pp. 23–34, Apr. 2015. DOI: 10.1007/978-3-319-18117-2_2.

[142] A. Lakhfif, M. T. Laskri, and E. Atwell, "Multi-level analysis and annotation of arabic corpora for text-to-sign language mt", *arXiv preprint arXiv:1605.07346*, 2016.

[143] A. Lakhfif and M. T. Laskri, "A frame-based approach for capturing semantics from arabic text for text-to-sign language mt", *International Journal of Speech Technology*, vol. 19, no. 2, pp. 203–228, 2016.

[144] A. Lakhfif and M. T. Laskri, "L'analyse et l'annotation à base de framenet: Contribution à l'étude contrastive des événements de mouvement en arabe et en anglais [framenet-based analysis and annotation: Contribution to the constrastive study of motion events in arabic and english]", *Traitement Automatique des Langues*, vol. 58, no. 3, pp. 67–95, 2017.

[145] J. Cohen, "A coefficient of agreement for nominal scales", *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[146] K. Taghva, R. Elkhoury, and J. Coombs, "Arabic stemming without a root dictionary", in *International Conference on Information Technology: Coding and Computing (ITCC'05)-Volume II*, IEEE, vol. 1, 2005, pp. 152–157.

[147] H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, "A transformer-based hybrid model for implicit emotion recognition in arabic text", *Engineering, Technology amp; Applied Science Research*, vol. 15, no. 3, 23834–23839, 2025. DOI: 10.48084/etasr.10261. [Online]. Available: https://etasr.com/index.php/ETASR/article/view/10261.

[148] M. J. Althobaiti, "Emotion recognition in arabic: A bert-based transfer learning approach leveraging semantic information of online comments", *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 9, pp. 3270–3282, 2023.

[149] N. Alswaidan and M. E. B. Menai, "Hybrid feature model for emotion recognition in arabic text", *IEEE Access*, vol. 8, pp. 37 843–37 854, 2020.

[150] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models", in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, Curran Associates, Inc., 2022, pp. 24 824–24 837. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[151] R OpenAI, "Gpt-4 technical report. arxiv 2303.08774", *View in Article*, vol. 2, no. 5, p. 1, 2023.

[152] Z. Liu, K. Yang, Q. Xie, T. Zhang, and S. Ananiadou, "Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis", in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 5487–5496. DOI: https://doi.org/10.1145/3637528.3671552.

[153] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., "A survey of large language models", *arXiv preprint arXiv:2303.18223*, vol. 1, no. 2, pp. 1–124, 2023.

[154] H. Boutouta, F. Senator, Z. Lakab, C. Mediani, and A. Lakhfif, "Multi-label emotion recognition in low-resource dialects: A case study on algerian arabic with large language models", *Procedia Computer Science*, vol. 275, pp. 817–824, 2026, 7th International Conference on AI in Computational Linguistics, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2026.01.094. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050926000943.

[155] H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, "From context to emotion: Leveraging LLMs for recognizing implicit emotions", in *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, Southern Denmark University, Odense, Denmark: Association for Computational Linguistics, Aug. 2025, pp. 399–409. [Online]. Available: https://aclanthology.org/2025.icnlsp-1.39/.

[156] H. Boutouta, A. Lakhfif, F. Senator, and C. Mediani, "Do llms grasp emotion or just keywords? a benchmark for implicit emotion recognition", in *2025 1st International Conference on Artificial Intelligence and Cyber Physical Systems (AICPS)*, 2025, pp. 104–109. DOI: 10.1109/AICPS66617.2025.11513483.

[157] F. Azouaou and I. Guellil, "Alg/fr: A step by step construction of a lexicon between algerian dialect and french", in *The 31st Pacific Asia Conference on Language, Information and Computation PACLIC*, vol. 31, 2017.

[158] J. C. Watson, *The phonology and morphology of Arabic* (Oxford linguistics). Oxford University Press, 2002, ISBN: 9780199257591. [Online]. Available: https://books.google.dz/books?id=lLQSDAAAQBAJ.

[159] H. Dahmani, H. Hussein, B. Meyer-Sickendiek, and O. Jokisch, "Natural arabic language resources for emotion recognition in algerian dialect", in *Arabic Language Processing: From Theory to Practice*, K. Smaïli, Ed., Cham: Springer International Publishing, 2019, pp. 18–33, ISBN: 978-3-030-32959-4.

[160] Z. H. Ali and H. J. Aleqabie, "Emotion detection in arabic text in social media: A brief survey", *Al-Furat Journal of Innovations in Electronics and Computer Engineering*, vol. 3, no. 2, pp. 412–421, 2024.

[161] O. Banimelhem and W. Amayreh, "The performance of chatgpt in emotion classification", in *2023 14th International Conference on Information and Communication Systems (ICICS)*, IEEE, 2023, pp. 1–4.

[162] S. J. Lee, H. T. Lee, and K. Lee, "Enhancing emotion detection through chatgpt-augmented text transformation in social media text", in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, IEEE, 2024, pp. 872–879.

[163] S. H. Muhammad et al., "BRIGHTER: BRIdging the gap in human-annotated textual emotion recognition datasets for 28 languages", in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 8895–8916, ISBN: 979-8-89176-251-0. DOI: 10.18653/v1/2025.acl-long.436. [Online]. Available: https://aclanthology.org/2025.acl-long.436/.

[164] S. H. Muhammad et al., "SemEval-2025 task 11: Bridging the gap in text-based emotion detection", in *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, S. Rosenthal, A. Rosá, D. Ghosh, and M. Zampieri, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 2558–2569, ISBN: 979-8-89176-273-2. [Online]. Available: https://aclanthology.org/2025.semeval-1.327/.

[165] Z. Benmounah, A. Boulesnane, A. Fadheli, and M. Khial, "Sentiment analysis on algerian dialect with transformers", *Applied Sciences*, vol. 13, no. 20, 2023, ISSN: 2076-3417. DOI: 10.3390/app132011157. [Online]. Available: https://www.mdpi.com/2076-3417/13/20/11157.

[166] Z. Ruan, R. You, K. Yang, J. Lin, W. Dai, M. Zhou, M. Jin, and X. Mei, "PAI at SemEval-2025 task 11: A large language model ensemble strategy for text-based emotion detection", in *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, S. Rosenthal, A. Rosá, D. Ghosh, and M. Zampieri, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 1136–1142, ISBN: 979-8-89176-273-2. [Online]. Available: https://aclanthology.org/2025.semeval-1.150/.

[167] J. Xue, P. Nguyen, M. Nguyen, and X. Liu, "JNLP at SemEval-2025 task 11: Cross-lingual multi-label emotion detection using generative models", in *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, S. Rosenthal, A. Rosá, D. Ghosh, and M. Zampieri, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 20–27, ISBN: 979-8-89176-273-2. [Online]. Available: https://aclanthology.org/2025.semeval-1.4/.

[168] N. B. Ocampo, E. Sviridova, E. Cabrio, and S. Villata, "An in-depth analysis of implicit and subtle hate speech messages", in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds., Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1997–2013. DOI: 10.18653/v1/2023.eacl-main.147. [Online]. Available: https://aclanthology.org/2023.eacl-main.147/.