Evaluation of Pretrained and Instruction-Based Pretrained Models for Emotion Detection in Arabic Social Media Text

Md. Rafiul Biswas
Hamad Bin Khalifa University/Qatar
mbiswas@hbku.edu.qa

Shimaa Ibrahim, Mabrouka Bessghaier, Wajdi Zaghoauni
Northwestern University in Qatar
{shimaa.ibrahim, mabrouka.bessghaier, wajdi.zaghouani}@northwestern.edu

Abstract

This study evaluates three approaches—instruction prompting of large language models (LLMs), instruction fine-tuning of LLMs, and transformer-based pretrained models on emotion detection in Arabic social media text. We compare pretrained transformer models like AraBERT, CaMelBERT, and XLM-RoBERTa with instruction prompting with advanced LLMs like GPT-4o, Gemini, Deepseek, and Fanar, and instruction fine-tuning approaches with LLMs like Llama 3.1, Mistral, and Phi. With a highly preprocessed dataset of 10,000 labeled Arabic tweets with overlapping emotional labels, our findings reveal that transformer-based pretrained models outperform instruction prompting and instruction fine-tuning approaches. Instruction prompts leverage general linguistic skills with maximum efficiency but fall short in detecting subtle emotional contexts. Instruction fine-tuning is more specific but trails behind pretrained transformer models. Our findings establish the need for optimized instruction-based approaches and underscore the important role played by domain-specific transformer architectures in accurate Arabic emotion detection.

1 Introduction

In recent years, emotion analysis has gained significant attention due to its critical role in understanding human emotions across applications such as social media monitoring, sentiment analysis, and user experience research. Sentiment analysis (SA), or opinion mining, is a core task in Natural Language Processing (NLP) involving detecting, extracting, and classifying opinions and emotions expressed in text (Marreddy and Mamidi, 2023; Hussein, 2018). However, traditional SA primarily focuses on polarity detection (positive, negative, neutral) (Singh et al., 2013), often overlooking the complexity and intensity of human emotions. Emotions are inherently ambiguous, and a single text frequently conveys multiple emotional states, necessitating a nuanced multilabel classification approach (Hong et al., 2025).

Text-based emotion recognition has evolved through feature engineering and deep learning techniques (Bharti et al., 2022). Nevertheless, existing research predominantly focuses on single-label emotion detection, limiting the ability to capture the intricacies of multilabel emotional expressions. This issue is particularly pronounced in Arabic due to limited availability and imbalance in labeled datasets, restricting advancements in Arabic emotion analysis (Alqahtani and Alothaim, 2022a). To address this gap, this study leverages a multilabel subset of an existing Arabic emotion dataset, aiming to provide a crucial resource for advancing Arabic emotion classification.

Recent advancements in large language models (LLMs) have demonstrated exceptional capabilities in comprehending, interpreting, and generating human-like text (Santoso et al., 2024). Beyond linguistic comprehension, these models incorporate emotional and social intelligence, significantly enhancing human-AI interactions (Huang et al., 2019). Instruction tuning—fine-tuning LLMs with natural language instructions and task-specific responses—has emerged as a promising method for enhancing performance across NLP tasks (Ouyang et al., 2022; Mishra et al., 2022). Unlike traditional models, instruction-tuned LLMs demonstrate improved generalization to new scenarios without extensive retraining, making them particularly beneficial for underrepresented languages like Arabic (Chouikhi et al., 2024).

In this study, we explore the effectiveness of instruction tuning for Arabic emotion analysis, comparing instruction-tuned large language models (LLMs) with fine-tuned transformer models in a

Proceedings of Recent Advances in Natural Language Processing,pages 158–165
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-019

multilabel emotion classification setting. We explicitly identify the use of fine-tuned transformers and evaluate model performance using both micro and macro F1 scores to account for dataset imbalance across emotion classes. Furthermore, we provide detailed justifications for preprocessing choices, present our prompt templates for reproducibility, and discuss the limitations of our approach. This research offers valuable insights into the challenges and strategies involved in Arabic multilabel emotion classification, supporting future progress in Arabic NLP applications.

2 Related Work

Recent advancements in Arabic emotion analysis have been driven by labeled datasets with varying annotation methodologies. ArPanEmo offers 11,128 manually labeled social media posts focusing on the Saudi dialect during COVID-19 (Althobaiti, 2023a), while SemEval-2018 Arabic Emotion provides 4,381 multi-label tweets across 11 emotions (Mohammad et al., 2018). ExaACE extends this with 20,050 posts supporting multi-label annotation. However, issues like class imbalance, dialectal variation, and subjective interpretation persist, limiting effectiveness (Aslam et al., 2024).

Early efforts used traditional methods such as SVMs, Naïve Bayes, and Decision Trees, often with emotion lexicons (Aljwari, 2022), but struggled with Arabic's morphology and dialects (Alqahtani and Alothaim, 2022b). Deep learning introduced CNNs and RNNs, with models like BiLSTM and GRU improving results using pre-trained embeddings (Abdelgwad et al., 2022; Daraghmi et al., 2024; Samara and Abandah, 2021; Al-Qerem et al., 2024). Hybrid approaches combined handcrafted features with deep networks, but challenges remain in colloquial and low-resource contexts (Aljwari, 2022).

Transformer models such as AraBERT and MARBERT significantly advanced Arabic emotion classification (Abdul-Mageed et al., 2021). Fine-tuning these models led to strong gains in multi-label classification. Ensemble techniques and stacked embeddings further improved results (Nfaoui and Elfaik, 2024; Aslam et al., 2024), though class imbalance and underrepresented emotions remain challenging.

Instruction tuning has gained traction for improving generalization and intent adherence in NLP (Zhang et al., 2023; Shi et al., 2024). While models like FLAN (Longpre et al., 2023) and Alpaca (Taori et al., 2023) have succeeded in English, Arabic remains underrepresented, with many resources relying on culturally limited translations. Recent monolingual instruction datasets show promise, but instruction tuning for Arabic emotion remains underexplored (Alyafeai et al., 2024).

Previous work has often failed to capture emotion co-occurrence, relying on single-label classification. Multi-label learning offers a better representation of emotional complexity but poses challenges in label correlation and fine-grained differentiation. The morphological complexity, slang, and informality of Arabic further hinder detection.

This study addresses these gaps by applying instruction tuning and LLMs for Arabic multi-label emotion analysis, aiming to better capture nuanced emotional expressions and overcome data scarcity through label-aware training and augmentation strategies

3 Methodology

This section discusses the dataset collection process and methodology applied to classify emotion in Arabic text.

3.1 Corpus description

There are a good number of emotion datasets in the Arabic text (Almahdawi and Teahan, 2019; Althobaiti, 2023b; Abdullah et al., 2020). However, all of them contain a single label for each text. We selected (Zaghouani et al., 2024) corpus, which can be used as multiperspective dataset such as emotion, emotion intensity, sentiment, offensive, hate speech, fact-checking, spam, vulnerability, humor, violence, and sarcasm. The corpus was incubated from Twitter data between August 2020 and October 2020. The corpus is annotated by multiple annotators. We selected the emotion category for the experimental evaluation of LLM performance. We randomly selected a sample of 10000 tweets labeled with emotion from the original corpus for this analysis. There are a total 12 labels: neutral, anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust. It includes a diverse range of emotional labels, with many instances containing multiple emotions. It captures the complexity of human emotions by including combinations such as Disgust with Trust, Sadness with Disgust, and other combinations like Love and Fear. The presence of overlapping emo-

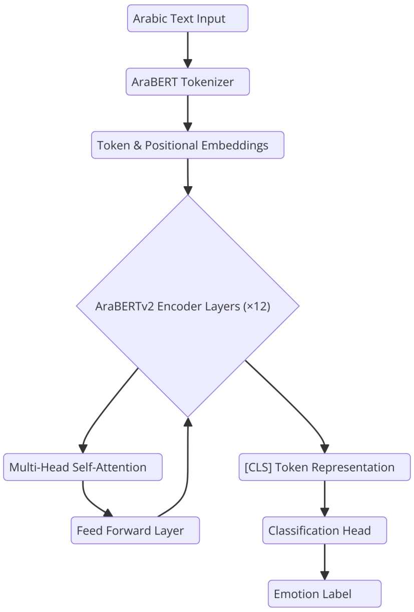

Figure 1: Architecture of Transformer for Emotion detection

tions throughout the dataset illustrates the multi-dimensional nature of emotional expression, where individuals may experience and express more than one emotion simultaneously. This diversity in emotional labels adds complexity to the emotion detection task, requiring models to identify and differentiate between multiple emotions in a single instance.

3.2 Transformer Model

Transformer models offer several compelling advantages for emotion classification tasks, especially with Arabic text. The Transformer architecture utilizes self-attention mechanisms to process sequential data, eliminating the need for recurrent layers. It primarily consists of two components: an encoder and a decoder. For text classification tasks such as emotion detection, typically only the encoder component is utilized. Transformers effectively capture context and long-range dependencies within text.

Several transformer-based models are available specifically tailored for Arabic text classification, such as AraBERT (Antoun et al., 2020), CaMelBERT (Inoue et al., 2021), and multilingual models like XLM-RoBERTa (Conneau et al., 2019). Figure 1 illustrates the AraBERT transformer architecture adopted for this analysis.

We selected a transformer-based model because it effectively captures deep contextual relationships between words, crucial for recognizing subtle emotional nuances in Arabic text. Unlike traditional models, which may overlook critical contextual information, transformers can differentiate between seemingly similar phrases that convey distinct emotions depending on their context.

3.3 Instruction Fine Tuning LLM

We performed instruction fine-tuning on three large language models: Llama 3.1, Mistral, and Phi, to identify emotional content in Arabic social media text. Using the Unsloth library (Daniel Han and team, 2023), models were fine-tuned with low-rank adaptation (LoRA) to enhance computational efficiency, employing 4-bit quantization for reduced memory usage and accelerated training. Instruction-based datasets were carefully prepared with input-output pairs specifying emotional annotations, formatted explicitly to align with each model's instruction-following capabilities. This fine-tuning enabled models to more accurately understand context-sensitive emotional nuances, significantly improving their performance in emotion classification tasks compared to baseline approaches.

3.4 Instruction Prompt Engineering

We applied instruction prompts to perform emotion recognition on Arabic social media text using advanced large language models, including GPT-4o, Gemini, Deepseek, and Fanar. Rather than conducting full instruction tuning, we utilized instruction-based prompts to leverage the pre-trained models' robust generalization capabilities, significantly reducing computational costs and complexity. This method allowed us to effectively harness each model's sophisticated understanding of language nuances, context, and semantics, ensuring accurate detection and classification of subtle emotional cues within Arabic text without requiring extensive fine-tuning efforts.

Prompt templates were generated using the `CreateInstructionSetForLLM` tool (Biswas, 2024), which automates instruction creation for multilabel tasks. We deployed GPT-4o via Azure OpenAI and used a structured prompt generation pipeline with a fixed system prompt and a user-defined instruction generation task.

4 Experiments

This section presents the experimental setup, the obtained results, and a detailed discussion of these results.

4.1 Dataset Preparation

The dataset contains Arabic tweets with emotion labels. To clean and preprocess the Arabic tweets dataset for emotion classification, we removed all non-Arabic characters, punctuation marks, and special symbols while preserving Arabic diacritics that can carry meaning. We performed text normalization by unifying various forms of Alef and Ya characters removed elongations (tatweel), and standardized common dialectal variations. Next, stopwords for both Arabic (Alrefaie, 2019) and English (NLTK library) were removed. URLs, usernames, and hashtags were either removed or replaced with placeholder tokens. Finally, we handled dialectal Arabic by mapping common dialectal words to their Modern Standard Arabic equivalents where possible, as Arabic tweets often contain a mix of formal and colloquial language.

4.2 Experimental setup

For the emotion detection task, we conducted three distinct types of experiments: transformer-based models, instruction prompting, and fine-tuning LLMs. For the transformer-based experiments, we used three pre-trained models namely, AraBERTv2, CamelBERT, and XLM-RoBERTa. AraBERTv2 was trained with a batch size of 4 for 5 epochs, while CamelBERT and XLM-RoBERTa were both trained using a batch size of 4, for 3 epochs each, incorporating a dropout rate of 0.01, a learning rate of 2e-5, and a sigmoid loss function. All transformer-based models followed a train-validation-test data split of 70:10:20. In the Instruction Prompting approach, we used chat completion models. In these experiments, models are designed to generate responses based on specific prompts. Essentially, the model is provided with an input instruction, and it responds to the prompt in a conversational manner. Chat completion models are often used in dialogue systems or conversational AI tasks, where the goal is to generate human-like responses based on the given context. The pre-trained models employed for this approach were OpenAI's GPT-4o, Deepseek-r1-distill-llama-8b, and Google's Gemini-2.0-Flash-001. Finally, the third type, instruction fine-tuning, involves fine-tuning pre-trained models explicitly with task-specific instructions. For these experiments, we selected Mistral-B-instruct-v0.3 and Llama3.1.

4.3 Descriptive Statics

The dataset contains 10,000 samples labeled with various emotion categories (see Table 1). The predominant emotion category is 'Disgust' with 5,883 occurrences, followed by 'No emotions' at 1,767 occurrences. There is significant diversity in emotion combinations, with several emotions appearing concurrently; for instance, combinations like 'Disgust' and 'Trust' (312 instances) or 'Sadness' and 'Disgust' (225 instances). Many emotions, however, appear very rarely, often in single-digit counts, such as 'Fear' (10), 'Pessimism' (13), and multiple complex emotion combinations occurring only once. Frequency less than 20 are not shown in the table 1. This indicates a heavily imbalanced dataset, primarily dominated by 'Disgust', potentially requiring specialized strategies to handle class imbalance in emotion classification tasks.

Table 1: Emotion counts in the annotated dataset

Emotions	Count
Disgust	5883
No emotions	1767
Disgust, Trust	312
Trust	242
Sadness, Disgust	225
Anger, Disgust	177
Surprise	164
Anticipation	157
Love, Disgust	138
Love	125
Disgust, Surprise	123
Sadness	98
Disgust, Anticipation	87
Optimism, Disgust	50
Optimism	40
Anticipation, Trust	39
Joy	37
Joy, Disgust	28
Disgust, Anticipation, Trust	22
Love, Trust	20

4.4 Model Performance Evaluation

Our experiments are divided into three approaches: transformer-based models, instruction prompting, and instruction fine-tuning. The performance of the different models tested on our emotion detection task is summarized in Table 2, with evaluation metrics that include Micro F1 score, precision, precision, and recall.

For Transformer-Based Models, AraBERTv2 outperformed the other models with the highest Micro F1 score (0.74), accuracy (0.65), and precision (0.82). Similar results were obtained using CamelBERT and XLM-RoBERTa. They showed slightly lower F1-Scores of 0.72, and accuracy scores (0.63 and 0.64). Both models also demonstrated high precision and recall values, with scores of 0.79 and 0.66, respectively.

In the Instruction Prompting experiments, OpenAI's GPT-4o achieved a Micro F1-Score of 0.42, which was the highest among the instruction-based prompting models, although it was still much lower compared to the transformer-based models. The other models, Deepspeck, Fanar, and Gemini had significantly lower scores.

For the Instruction Fine-Tuning experiments, Mistral and Microsoft phi 4 showed notably lower performance metrics. Mistral achieved a Micro F1-Score of 0.24 and an accuracy of 0.25, while Microsoft phi 4 had the lowest performance with a Micro F1-Score of 0.11 and an accuracy of 0.32. Notably, all tested models exhibited a loss value of around 0.16, indicating similar levels of training error across the models. In general, transformer-based models, particularly AraBERTv2, demonstrate superior performance across all metrics compared to instruction prompting and fine-tuning approaches, as shown in 2 that the training loss decreases over epochs, starting at approximately 0.184 and decreasing steadily to around 0.099. This smooth decline indicates that the model is successfully learning from the training data without significant optimization difficulties. The validation loss does not increase significantly after epoch 1, indicating that the model is not severely overfitting to the training data. Figure 3, which presents the thresholds of the F1 score over epochs for AraBERTv2. The optimal threshold range of 0.4-0.5 represents the sweet spot where the model achieves the best balance between precision and recall, maximizing the F1 score for multi-label emotion classification in Arabic text. While instruction-based models show some promise, they fall short of achieving the level of performance seen with pre-trained transformer models. Instruction fine-tuning models, on the other hand, require further optimization to match the efficacy of the other two experimental approaches.

4.5 Discussion

In this study, we evaluated the performance of various models on the emotion detection task using three distinct experimental approaches: Transformer-Based Models, Instruction Prompting, and Instruction Fine-Tuning. The results reveal significant differences between these approaches, with transformer-based models consistently outperforming both instruction prompting and fine-tuning methods across most evaluation metrics.

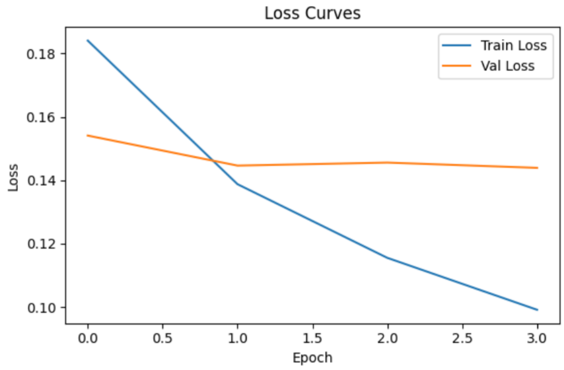

Figure 2: AraBERTv2 Loss Curve

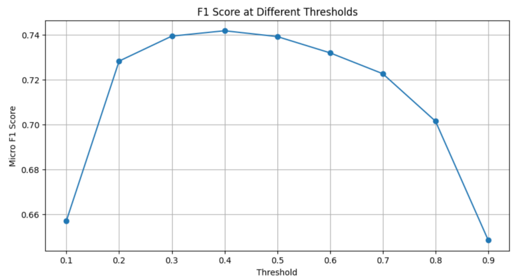

Figure 3: F1-Score threshold for AraBERTv2

While CamelBERT and XLM-RoBERTa showed slightly lower performance compared to AraBERTv2, their results still highlighted the strength of transformer-based models for emotion detection. The high performance of these models suggests they excel in correctly identifying emotion categories. For the Instruction Prompting approach, models such as OpenAI's GPT-4o and Deepseek achieved modest results compared to the transformer-based models. This indicates that while they can respond to a wide variety of instructions, they may not be specifically fine-tuned for emotion detection tasks. Regarding instruction fine-tuning approaches, although this methodology offers the potential for task-specific performance enhancement, our findings indicate that it is more challenging than initially anticipated, especially for complex NLP tasks. However, the loss value across all models remained around 0.16, suggesting that while the models performed differently in terms of metrics such as F1-Score, accuracy, and precision, their training stability was comparable. This may imply that, despite differences in model architectures and approaches,

Table 2: Emotion Detection Task Evaluation Results

Experiment Type	Model	Micro F1-Score	Accuracy	Precision	Recall
Transformer Based Models	AraBERTv2	0.74	0.65	0.82	0.68
	CamelBERT	0.72	0.63	0.79	0.66
	XLM-RoBERTa	0.72	0.64	0.79	0.66
Instruction Prompting	OpenAI's GPT-4o	0.42	0.58	0.32	0.61
	Deepseek	0.11	0.08	0.11	0.30
	Fanar	0.34	0.58	0.26	0.50
	Gemini	0.35	0.44	0.28	0.47
Instruction Fine-Tuning	Mistral	0.24	0.25	0.22	0.24
	Llama3.1	0.66	0.58	0.72	0.61
	Microsoft Phi 4	0.11	0.32	0.11	0.16

the models were trained similarly, with comparable training errors.

When we used LLMs, they did not produce the higher results that we had originally hoped for. Despite their success in various NLP tasks, the LLMs used in our experiments did not perform well on the emotion detection task, even with instruction prompting and when fine-tuned. They showed significantly lower performance compared to transformer-based models like AraBERTv2 on all key evaluation metrics, such as Micro F1-Score, Accuracy, Precision, and Recall. This may be because LLMs are typically trained for general language tasks and are not specifically optimized for emotion detection, which requires a deeper understanding of emotional nuances. As evidenced in the results, this study confirms that LLMs struggle to detect emotion, and further improvements are needed.

4.6 Limitations

This study has several limitations that should be acknowledged. First, the dataset used in our analysis exhibited significant class imbalance, with a high frequency of the 'Disgust' emotion. This imbalance may have influenced the generalizability of model performance, particularly impacting the detection of less frequent emotion categories. Second, although instruction prompts were explicitly formatted to support reproducibility, differences in model-specific responsiveness and capabilities may have affected consistency across instruction-based models. Third, our findings show that general-purpose LLMs, while broadly applicable, are not specifically optimized for complex emotion detection tasks. Lastly, computational resource constraints limited the scope of experimentation with larger datasets or extensive hyperparameter tuning, which may have further improved model performance.

5 Conclusion and Future work

In this work, we conducted experiments on emotion detection in Arabic social media text, focusing on three approaches: LLM instruction prompting, LLM instruction fine-tuning, and transformer-based pre-trained models. Our goal was to investigate how these three approaches impact performance and identify which performs better. Our findings revealed that transformer-based models perform the best for the task at hand, whereas fine-tuning and prompting LLMs struggle to achieve similar success.

As future work, we intend to fine-tune LLMs using a larger dataset. Additionally, while existing LLMs are effective for tasks such as chat completion, text generation, and image generation, there is a need for LLMs specifically designed for classification tasks. Furthermore, we intend to extend our investigation to other low-resource languages, where data and resources are more limited.

Acknowledgments

This study was supported by the grant NPRP14C-0916-210015, awarded by the Qatar Research, Development and Innovation Council (QRDI).

References

Mohammed M Abdelgwad, Taysir Hassan A Soliman, Ahmed I Taloba, and Mohamed Fawzy Farghaly.

2022. Arabic aspect based sentiment analysis using bidirectional gru based models. *Journal of King Saud University-Computer and Information Sciences*, 34(9):6652–6662.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. 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)*, pages 7088–7105, Online. Association for Computational Linguistics.

Manal Abdullah, Muna AlMasawa, Ibtihal Makki, Maha Alsolmi, and Samar Mahrous. 2020. Emotions extraction from arabic tweets. *International Journal of Computers and Applications*, 42(7):661–675.

Ahmad Al-Qerem, Mohammed Raja, Sameh Taqatqa, and Mutaz Rsmi Abu Sara. 2024. Utilizing deep learning models (rnn, lstm, cnn-lstm, and bi-lstm) for arabic text classification. In *Artificial Intelligence-Augmented Digital Twins: Transforming Industrial Operations for Innovation and Sustainability*, pages 287–301. Springer.

Fatimah Aljwari. 2022. Emotion detection in arabic text using machine learning methods. *IJISCS (International Journal of Information System and Computer Science)*, 6(3):175–185.

Amer J Almahdawi and William J Teahan. 2019. A new arabic dataset for emotion recognition. In *Intelligent Computing: Proceedings of the 2019 Computing Conference, Volume 2*, pages 200–216. Springer.

Ghadah Alqahtani and Abdulrahman Alothaim. 2022a. Emotion analysis of arabic tweets: Language models and available resources. *Frontiers in Artificial Intelligence*, 5:843038.

Ghadah Alqahtani and Abdulrahman Alothaim. 2022b. Emotion analysis of arabic tweets: Language models and available resources. *Frontiers in Artificial Intelligence*, 5.

Mohamed Taher Alrefaie. 2019. Arabic stop words.

Maha Jarallah Althobaiti. 2023a. An open-source dataset for arabic fine-grained emotion recognition of online content amid covid-19 pandemic. *Data in Brief*, 51:109745.

Maha Jarallah Althobaiti. 2023b. An open-source dataset for arabic fine-grained emotion recognition of online content amid covid-19 pandemic. *Data in Brief*, 51:109745.

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, Yousef Ali, and Maged Al-shaibani. 2024. CIDAR: Culturally relevant instruction dataset for Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12878–12901, Bangkok, Thailand. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Muhammad Azeem Aslam, Wang Jun, Nisar Ahmed, Muhammad Imran Zaman, Li Yanan, Hu Hongfei, Wang Shiyu, and Xin Liu. 2024. Improving arabic multi-label emotion classification using stacked embeddings and hybrid loss function. *arXiv preprint arXiv:2410.03979*.

Santosh Kumar Bharti, S Varadhaganapathy, Rajeev Kumar Gupta, Prashant Kumar Shukla, Mohamed Bouye, Simon Karanja Hingaa, and Amena Mahmoud. 2022. Text-based emotion recognition using deep learning approach. *Computational Intelligence and Neuroscience*, 2022(1):2645381.

Md Rafiul Biswas. 2024. Createinstructionsetforllm: A pipeline to generate instruction-response datasets from multi-label corpora. `https://github.com/rafiulbiswas/CreateInstructionSetForLLM`. Accessed: July 2025.

Hasna Chouikhi, Manel Aloui, Cyrine Ben Hammou, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 2024. Gemmar: Enhancing llms through arabic instruction-tuning.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Eman Yaser Daraghmi, Sajida Qadan, Yousef Daraghmi, Rami Yussuf, Omar Cheikhrouhou, and Mohammed Baz. 2024. From text to insight: An integrated cnn-bilstm-gru model for arabic cyberbullying detection. *IEEE Access*.

Xin Hong, Yuan Gong, Vidhyasaharan Sethu, and Ting Dang. 2025. Aer-llm: Ambiguity-aware emotion recognition leveraging large language models.

Chenyang Huang, Amine Trabelsi, and Osmar R Zaïane. 2019. Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert. *arXiv preprint arXiv:1904.00132*.

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

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. *arXiv preprint arXiv:2103.06678*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

Mounika Marreddy and Radhika Mamidi. 2023. Chapter 6 - learning sentiment analysis with word embeddings. In Dipankar Das, Anup Kumar Kolya, Abhishek Basu, and Soham Sarkar, editors, *Computational Intelligence Applications for Text and Sentiment Data Analysis*, Hybrid Computational Intelligence for Pattern Analysis and Understanding, pages 141–161. Academic Press.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

El Habib Nfaoui and Hanane Elfaik. 2024. Evaluating arabic emotion recognition task using chatgpt models: A comparative analysis between emotional stimuli prompt, fine-tuning, and in-context learning. *Journal of Theoretical and Applied Electronic Commerce Research*, 19(2):1118–1141.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Abrar K Samara and Gheith A Abandah. 2021. Investigating fast bilstm neural networks for arabic language applications. In *2021 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, pages 250–255. IEEE.

Jennifer Santoso, Kenkichi Ishizuka, and Taiichi Hashimoto. 2024. Large language model-based emotional speech annotation using context and acoustic feature for speech emotion recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11026–11030.

Zhengyan Shi, Adam X. Yang, Bin Wu, Laurence Aitchison, Emine Yilmaz, and Aldo Lipani. 2024. Instruction tuning with loss over instructions. In *Advances in Neural Information Processing Systems*, volume 37, pages 69176–69205. Curran Associates, Inc.

V. K. Singh, R. Piryani, A. Uddin, and P. Waila. 2013. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, pages 712–717.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Wajdi Zaghouani, Hamdy Mubarak, and Md Rafiul Biswas. 2024. So hateful! building a multi-label hate speech annotated arabic dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15044–15055.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.