JNLP at SemEval-2025 Task 11: Cross-Lingual Multi-Label Emotion Detection Using Generative Models

Jieying Xue[1], Phuong Minh Nguyen[2,1], Minh Le Nguyen[1], Xin Liu[3]

[1]Japan Advanced Institute of Science and Technology
[2]ROIS-DS Center for Juris-Informatics, NII, Tokyo, Japan
[3]National Institute of Advanced Industrial Science and Technology
{xuejieying,phuongnm,nguyenml}@jaist.ac.jp, xin.liu@aist.go.jp

Correspondence: phuongnm@jaist.ac.jp

Abstract

With the rapid advancement of global digitalization, users from different countries increasingly rely on social media for information exchange. In this context, multilingual multi-label emotion detection has emerged as a critical research area. This study addresses SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection. Our paper focuses on two sub-tracks of this task: (1) Track A: Multi-label emotion detection, and (2) Track B: Emotion intensity. To tackle multilingual challenges, we leverage pre-trained multilingual models and focus on two architectures: (1) a fine-tuned BERT-based classification model and (2) an instruction-tuned generative LLM. Additionally, we propose two methods for handling multi-label classification: the *base* method, which maps an input directly to all its corresponding emotion labels, and the *pairwise* method, which models the relationship between the input text and each emotion category individually. Experimental results demonstrate the strong generalization ability of our approach in multilingual emotion recognition. In Track A, our method achieved Top 4 performance across 10 languages, ranking 1st in Hindi. In Track B, our approach also secured Top 5 performance in 7 languages, highlighting its simplicity and effectiveness[1].

1 Introduction

With the rapid proliferation of social media, particularly in the context of global digital communication, online platforms have emerged as the primary medium for information dissemination (Nandwani and Verma, 2021). Users from diverse linguistic backgrounds frequently express their opinions through comments, highlighting the growing need for cross-lingual sentiment detection (Nandwani and Verma, 2021). Consequently, multilingual sentence-level sentiment analysis has become a critical task for tracking public sentiment (Wankhade et al., 2022). Sentiment analysis is one of the most extensively studied applications in natural language processing (NLP). In text emotion recognition, it is common for a single sentence to express multiple emotions with varying intensities (Deng and Ren, 2020). However, developing reliable multi-label emotion analysis systems remains particularly challenging due to the scarcity of training data, especially for low-resource languages. Additionally, pre-trained language models often have limited knowledge of these languages, further complicating the task. To address these challenges, this paper presents our approach for SemEval-2025 Task 11, "Bridging the Gap in Multilingual Multi-Label Emotion Detection from Text Using Large Language Models" (Muhammad et al., 2025b). We participated in two tracks: Track 1 (Multi-Label Emotion Detection) using the BRIGHTER dataset, which includes 28 languages (Muhammad et al., 2025a; Belay et al., 2025), and Track 2 (Emotion Intensity Prediction), which covers 11 languages.

In this study, to address the challenges of multilingual sentiment analysis, we leveraged pre-trained models (such as RoBERTa) and large language models (LLMs) to perform multi-label sentiment analysis on both high-resource languages like English and Chinese, as well as low-resource languages such as African languages. We formulated multi-label emotion recognition as a text generation task. To overcome the challenge of limited training data, we utilize the capabilities of multilingual pre-trained language models to enhance both semantic understanding and the recognition of emotional tone, especially in low-resource languages. Furthermore, to tackle the multi-label classification challenge, we propose two methods: the *pairwise* method and the *base* method. Our findings also indicate that training the model on a combined multilingual dataset improves performance compared

[1]Our code is available at https://github.com/yingjie7/mlingual_multilabel_emo_detection

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 20–27

to training on individual language datasets. We present experiments comparing the applicability of these methods and conduct ablation studies to validate their effectiveness. Our approach demonstrates strong performance in both multi-label emotion recognition and emotion intensity detection. In Track A, it achieved top-four rankings in 10 languages, including first place in Hindi. In Track B, our method ranked within the top five for 7 languages, further highlighting its simplicity and effectiveness. Moreover, our approach exhibits strong generalization across both competition sub-tasks, making it particularly beneficial for low-resource languages.

2 Background

Sentence-level sentiment analysis (SLSA) has advanced significantly with the rise of deep learning and multilingual sentiment detection. Early research primarily focused on extracting handcrafted sentiment features such as n-grams (Tripathy et al., 2016), lexicons, rule-based heuristics (Chikersal et al., 2015) to enhance SVM-based classifiers (Kumari et al., 2017) and deep neural networks, such as CNNs and RNNs (Chikersal et al., 2015; Minaee et al., 2019). Nonetheless, their reliance on static word embeddings limited their ability to handle complex linguistic phenomena such as long-range dependencies and cross-lingual variations. To address these limitations, researchers turned to Transformer-based pretrained language models (PLMs) such as BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020), which more effectively capture fine-grained emotional representations (Zhou et al., 2016; Li et al., 2018) by modeling richer linguistic semantics. In multilingual sentiment analysis, models like mT5 (Xue et al., 2021) and XLM-RoBERTa (XLM-R) (Conneau et al., 2020) further advanced the field by learning cross-lingual representations, making them the standard for multilingual applications (Hu et al., 2020). More recently, the widespread adoption of large language models (LLMs) such as LLaMA 2 (Touvron et al., 2023) has driven major breakthroughs in various NLP tasks (Upadhye, 2024; Sharma et al., 2023). These models exhibit remarkable zero-shot and few-shot learning capabilities, making them highly adaptable to new sentiment analysis tasks (Maceda et al., 2024). Furthermore, in the domain of Emotion Recognition in Conversations, LLMs have been leveraged with prompt-based techniques to extract latent supplementary knowledge from text, injecting this information to facilitate emotion recognition (Xue et al., 2024). In the broader NLP landscape, various methodologies including fine-tuning, prompting, transfer learning, and domain adaptation have been pivotal in adapting pre-trained LLMs for sentiment analysis across specific domains and languages.

However, as most PLMs are predominantly pretrained on English text, their effectiveness in multilingual sentiment analysis is often limited without additional fine-tuning is performed to optimize performance across diverse linguistic contexts (Zhang et al., 2023). Numerous studies have explored leveraging embeddings from LLMs for sentiment classification, using various low-resource datasets to assess their adaptability across languages (Dadure et al., 2025; Mujahid et al., 2023). In our work, we leverage BERT-based multilingual models to extend multi-label classification tasks, enabling knowledge transfer across languages. By integrating LLMs, we also present a *pairwise emotional recognition* method, which efficiently captures both emotional intensity and sentiment polarity within each sentence. This approach ensures that the model concentrates on one label at a time. Additionally, we reformulate the multi-label classification task as a text generation problem, enhancing the model's adaptability and generalization across NLP tasks.

3 System Description

In this work, the target task involves the perception of emotions in various languages, which aims to identify the emotion that most people would attribute to the speaker based on a given sentence or short text snippet. Given a text input (x), a machine learning system needs to retrieve all the multi-label emotions (y_e) expressed in the given text (Track A) and the intensity (y_i) of each class (Track B).

3.1 System Overview

In general, we leverage the capabilities of pretrained multilingual models to tackle cross-lingual challenges. Our system primarily focuses on two architectures: fine-tuning BERT-based classification models (Devlin et al., 2019) and instruction fine-tuning generative LLMs, building upon recent SOTA methods in the field of emotion recognition (Xue et al., 2024). To handle multi-label classification, we design two strategies: (1) the *base* method,

which maps a given input to all its corresponding labels, and (2) the *pairwise* method, which models the relationship between the input text and each label class individually.

$$\textit{base}: \quad \mathbb{P}_A(\{y_e\} \mid x) \qquad \mathbb{P}_B(\{\langle y_e, y_i \rangle\} \mid x) \quad (1)$$

$$\textit{pairwise}: \quad \mathbb{P}_A(\{0,1\} \mid x, y_e) \quad \mathbb{P}_B(y_i \mid x, y_e) \quad (2)$$

where x is the given input text; $\mathbb{P}_A, \mathbb{P}_B$ are probability models for track A and B; y_e, y_i are the emotional label and its corresponding emotional intensity from a pre-defined label set, respectively.

3.2 Methods

BERT-based method. For the baseline, we design a BERT-based multi-label classification model. In detail, fully connected layers with nonlinear activation functions (sigmoid (σ) and tanh) are added to the top layer of the BERT architecture to transform the `[CLS]` feature vector (Devlin et al., 2019) from the hidden representation to the output dimension (number of labels).

$$h^{CLS}, h^{words} = \text{BERT}(x) \quad (3)$$

$$h^{out} = \sigma(\tanh(h^{CLS} \cdot W^h) \cdot W^o) \quad (4)$$

Finally, during the fine-tuning process, the learnable weights (W^*) are optimized using cross-entropy loss on annotated data to maximize the log-likelihood of the model.

LLM-based method. Leveraging the robust natural language understanding capabilities of large language models (LLMs) (Touvron et al., 2023), we employ instruction prompting (highlighted in blue in Table 1) to guide the model in comprehending the task requirements. Our methodology follows instruction fine-tuning as outlined by Chung et al. (2022), using a *causal language modeling* objective to train the LLM to generate emotional label text, which is highlighted in red in Table 1.

$$s = \text{instruction-prompting}(x, y) \quad (5)$$

$$\mathbb{P}(s) = \Pi_{z=1}^{|s|} \mathbb{P}(s_z | s_0, s_1, ..., s_{z-1}) \quad (6)$$

where s, x represent a sequence of tokens, and z denotes the token index within the prompting input (Table 1). To optimize efficiency, we employ LoRA (Hu et al., 2022), a lightweight training methodology that reduces the number of trainable parameters. The fine-tuned LLM is designed to learn the distribution of emotional labels (or emotional intensity) based on the given prompt (s). During inference, the emotional label (y), which is omitted from the input prompt, is generated by the fine-tuned model.

system
You are an expert in analyzing the emotions expressed in a natural sentence. The emotional label set includes {anger, fear, joy, sadness, surprise}. Each sentence may have one or more emotional labels, or none at all.
user
Given the sentence: "`{input text: x}`", which emotions are expressed in it?
assistant
`{emotional label in text:` y_e `or` $\langle y_e, y_i \rangle$`}`

Table 1: Instruction prompting with the *base* template (track A).

Task	Strategy	Input	Output	Output example
Track A	base	x	$\{y_e\}$	*"disgust, sadness"*
Track B	base	x	$\{\langle y_e, y_i \rangle\}$	*"moderate degree of anger, low degree of sadness"*
Track A	pairwise	x, y_e	$\{0, 1\}$	*"yes"*
Track B	pairwise	x, y_e	y_i	*"moderate"*

Table 2: Examples of output format for text generation.

As outlined in the overview section, we have devised two approaches, *base* and *pairwise*, to tackle this task. Both approaches employ the same training techniques across tracks A and B and between the two approaches themselves. We provide example outputs designed for both tracks in Table 2. Detailed examples for each track are provided in Appendix A.

4 Experimental Setup

Dataset. To evaluate our methods, we use the original emotional data provided by the SemEval Task 11 organization. This dataset consists of three subsets: training, development, and test sets, spanning two competition phases: development and test. However, to ensure greater generalization, we consistently set aside 10% of the training data from each language as an internal development set. This held-out portion is used for hyper-parameter tuning, ensuring that the optimized checkpoints are selected based on this internal dev set. Additionally, to handle multilingual data, we design two settings: (1) *separated langs*, where a separate model is trained for each language, and (2) *mixed langs*, where a single model is trained to learn all languages simultaneously.

Evaluation Metric. According to the competition guidelines, the evaluation metric for Track A is the macro-averaged F1-score, while for Track B, it is the Pearson correlation coefficient between the predicted and gold-standard labels.

Model	Strategy	Data	afr	amh	arq	ary	chn	deu	eng	esp	hau	hin	ibo	kin	mar	orm
(development)																
Qwen 32b	pairwise	separated langs	0.5143	0.5049	0.6574	0.5242	0.6909	0.7187	0.8189	0.8366	0.5724	0.8694	0.5049	0.4274	0.9507	-
Qwen 32b	base	separated langs	0.4610	-	-	-	-	-	0.8054	-	-	-	-	-	-	-
Qwen 32b	base	mixed langs	0.5140	0.557	0.64	0.537	0.732	0.677	0.751	0.839	0.57	0.899	0.509	0.477	0.959	0.478
Qwen 14b	base	mixed langs	0.4320	0.594	0.588	0.567	0.643	0.691	0.743	0.835	0.606	0.887	0.498	0.454	0.924	0.503
xml-roberta	base	mixed langs	0.5070	0.66	0.607	0.548	0.623	0.654	0.703	0.786	0.687	0.855	0.488	0.328	0.948	0.513
JNLP *(test)*			0.5925	0.6767	0.6407	0.609	0.6805	0.6990*	0.8036	0.8303	0.6504	*0.9257*	0.5404	0.4289	0.878	0.573

Model	Strategy	Data	pcm	ptbr	ptmz	ron	rus	som	sun	swa	swe	tat	tir	ukr	vmw	yor	Average
(development)																	
Qwen 32b	pairwise	separated langs	0.6202	0.6407	0.5161	0.7548	0.8809	0.3571	0.5307	0.2658	0.5915	0.6282	0.4581	0.6761	0.1265	0.4554	0.5933
Qwen 32b	base	separated langs	0.611	-	-	0.7230	-	-	-	-	-	-	-	-	-	-	-
Qwen 32b	base	mixed langs	0.638	0.546	0.571	-	0.902	0.416	0.557	0.332	0.509	0.72	0.429	0.639	0.114	0.355	0.5933
Qwen 14b	base	mixed langs	0.622	0.576	0.553	-	0.895	0.394	0.51	0.319	0.494	0.764	0.485	0.64	0.19	0.348	0.5863
xml-roberta	base	mixed langs	0.574	0.502	0.579	-	0.876	0.499	0.539	0.348	0.501	0.692	0.5	0.594	0.074	0.198	0.5718
JNLP *(test)*			0.6343	0.6184	0.4535	0.7787	0.8912	0.4965	0.4596	0.2949	0.6186	0.7223	0.4849	0.6873*	0.2261	0.3608	0.6163

Table 3: Results of Sub-task A. For a fair comparison, the *average* column is computed based on all languages except for *orm, ron, ptbr*, and *ptmz*, as these languages are missing in some settings. The red color indicates the best setting used for submission to obtain the test result. The notation (-) indicates that the experiment was not conducted. The asterisk (*) denotes results obtained during the post-evaluation phase.

Model	Strategy	Data	amh	arq	chn	deu	eng	esp	hau	ptbr	ron	rus	ukr	Average
(development)														
Llama2-13b	pairwise	separated langs	-	0.4411	0.73857	0.6197	0.8207	0.7221	0.5691	0.4938	0.691	0.8719	0.6229	0.6757
Qwen-32b	pairwise	separated langs	0.5433	0.6147	0.75	0.6793	0.8101	0.7715	0.6143	-	0.7245	0.9051	0.6428	0.7234
Qwen-32b	base	mixed langs	0.542	0.566	0.711	0.658	0.802	0.761	0.595	0.718	-	0.898	0.659	0.7063
Qwen-32b	pairwise	mixed langs	0.563	0.627	0.727	0.705	0.787	0.779	0.665	0.6	-	0.906	0.694	0.7363
JNLP *(test)*			0.6038	0.5873	0.6589	0.725	0.8129	0.7747	0.6496	0.6512	0.7055	0.9074	0.6719	0.7044

Table 4: Results of Sub-task B. The meanings of the denotations and colors are the same as in Table 3.

Experimental Environments. We implement all our experiments using widely adopted libraries such as PyTorch and HuggingFace. For pre-trained LLMs, we primarily experiment with XLM-RoBERTa-Large, Llama2 (7B-13B), and Qwen2.5 (14B–32B). For hyper-parameters, we train the model with a learning rate of $3e^{-4}$, using the AdamW optimization algorithm, 5–6 epochs.

5 Results

Overall, we evaluate our methods and their variants on the development set to select the best model and setting for each language (indicated in red in Tables 3, 4) for the final test submission.

5.1 Track A: Multi-label Emotion Detection.

Development Result. As shown in Table 3, we conducted experiments using both the *base* and *pairwise* methods on Qwen-32B, Qwen-14B, and RoBERTa models. The results indicate that the Qwen models with the *pairwise* method achieved the best overall performance. However, in datasets where the majority of samples contain zero or only one emotion label, the *base* method outperformed the *pairwise* approach. We attribute this to the fact that the *pairwise* method is inherently more suited for multi-label emotion recognition tasks. Additionally, in low-resource languages, LLMs performed poorly, whereas the RoBERTa-based approach yielded better results.

Test Result. Overall, our approach achieved 4th place in Track A for CHN, ESP, PCM, and PT-BR, secured 3rd place in ARQ, ARY, RON, and RUS, and ranked 2nd and 1st for SWE and HIN, respectively. These results demonstrate the strong generalization ability of our method, highlighting its simplicity and efficiency.

5.2 Track B: Emotion Intensity.

The results of Track A demonstrated the strength of LLMs compared to the XLM-RoBERTa model, leading us to primarily experiment with LLMs rather than XLM-RoBERTa in this track.

Development Result. As shown in Table 4, we conducted experiments using both the *base* and *pairwise* strategies on LLaMA 2 and Qwen-32B models. The results indicate that Qwen-32B outperforms LLaMA 2, and the *pairwise* strategy consistently achieves better overall performance compared to the *base* method.

Test Result. In Track B across 11 languages, our model achieved 3rd place in Ukrainian (ukr) and Algerian Arabic (arq), 4th place in Romanian (ron), and 5th place in Russian (rus), Brazilian Portuguese (ptbr), English (eng), and German (deu). With top-five rankings in seven languages, these results demonstrate the effectiveness and generalizability of our approach.

5.3 Result Analyses

Strategies Comparison. To gain a comprehensive understanding of the base and pairwise strate-

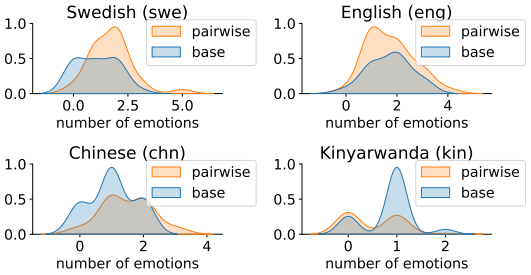

Figure 1: Distribution of improved samples between strategies *base* and *pairwise* with respect to the number of emotions (track A).

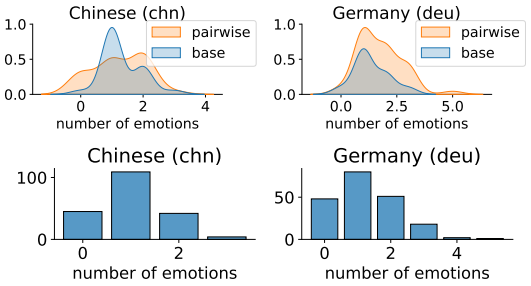

Figure 2: Distribution of improved samples between strategies *base* and *pairwise* with respect to the number of emotions (track B).

gies, we conducted experiments to analyze the distribution of improved examples, measured by the F1 score for each sample, across four languages: English, Swedish, Chinese, and Kinyarwanda (Figure 1 for Track A and Figure 2 for Track B). Our findings indicate that the pairwise strategy predominantly improves samples in languages that convey various emotions within a sentence, particularly in English and Swedish. Conversely, in languages or datasets with a limited variety of emotional labels, such as Chinese and Kinyarwanda (where each sample typically contains 0 to 2 emotions), the base strategy demonstrates a distinct advantage. We argue that this is because the pairwise strategy evaluates only one emotion at a time, making it more sensitive to label imbalance, which in turn leads to lower performance in languages with a limited variety of emotion labels compared to the base strategy. In contrast, the base strategy generates all emotions present in a sample simultaneously, highlighting its advantage in languages with fewer distinct emotion categories.

Emotional Type. To evaluate the model's effectiveness concerning emotional intensity across various emotions, we plotted the distribution of emotional labels alongside their corresponding intensities (Fig 3). We conducted experiments by aggregating all the languages and examined the correlation between performance and emotions, as well as their respective intensities. Our analysis revealed that classes with limited data, such as high-surprise or high-joy, typically exhibited poorer performance in our system. Conversely, the major emotional class, *"disgust"*, achieved the highest performance compared to other emotional classes, such as surprise, particularly high-surprise.

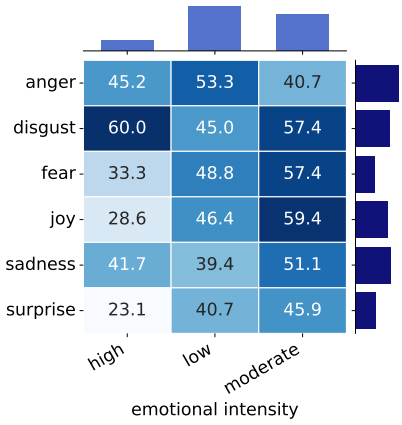

Figure 3: Overall performance of the pairwise strategy across all emotional labels in all languages (Track B).

Mixed languages. In both Sub-tasks A and B, mixed-language training, where a single model is fine-tuned for all languages, demonstrates superior performance compared to training separate models (Tables 3, 4). This improvement can be attributed to a more balanced distribution of emotion types across languages and the model's enhanced ability to generalize across linguistic variations.

6 Conclusion

In this work, we present a multilingual emotion recognition system for SemEval-2025 Task 11, which demonstrates strong performance and remains competitive with the top-performing teams. To address multilingual challenges, we design two architectures, BERT-based and LLM-based, and introduce two strategies, *pairwise* and *base*, for handling the multi-label classification task. We conduct extensive experiments to analyze the effectiveness and limitations of each approach, aiming to provide valuable insights for multilingual emotion recognition research. The results validate the simplicity and effectiveness of our methods, highlighting their strong generalization ability and applicability to other tasks.

Limitations and potential improvements. Despite the promising results achieved in our work, several limitations remain. First, the LLM-based approach is heavily dependent on the knowledge and capabilities within the LLMs themselves, which may limit its adaptability to evolving data. Furthermore, compared to BERT-based methods, LLM-based approach incurs higher computational costs, making it less efficient for large-scale or real-time applications.

There are several directions for future improvement. One potential enhancement lies in the utilization of finer-grained information contained in the logits output. Specifically, for the LLM Pairwise strategy, instead of relying solely on the final "yes" or "no" response, it is better to aggregate the logits corresponding to the tokens generating "yes" and compute a probability distribution via softmax. This would enable a more nuanced and probabilistic interpretation of the model's predictions, potentially improving robustness.

Another limitation is the incomplete handling of label imbalance. Our current framework does not fully address the issue, which may cause the model to overfit to dominant emotional categories. Future work could incorporate targeted data augmentation strategies, such as generating additional samples for underrepresented emotions, to mitigate this imbalance and enhance the overall performance and stability of the system.

References

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Prerna Chikersal, Soujanya Poria, and Erik Cambria. 2015. Sentu: sentiment analysis of tweets by combining a rule-based classifier with supervised learning. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 647–651.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Pankaj Dadure, Ananya Dixit, Kunal Tewatia, Nandini Paliwal, and Anshika Malla. 2025. Sentiment analysis of Arabic tweets using large language models. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 88–94, Abu Dhabi, UAE. Association for Computational Linguistics.

Jiawen Deng and Fuji Ren. 2020. Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning. *IEEE Transactions on Affective Computing*, 14(1):475–486.

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

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Upma Kumari, Arvind K Sharma, and Dinesh Soni. 2017. Sentiment analysis of smart phone product review using svm classification technique. In *2017 International conference on energy, communication, data analytics and soft computing (ICECDS)*, pages 1469–1474. IEEE.

Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. 2018. Hierarchical attention transfer network for cross-domain sentiment classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Lany Laguna Maceda, Jennifer Laraya Llovido, Miles Biago Artiaga, and Mideth Balawiswis Abisado. 2024. Classifying sentiments on social media texts: A gpt-4 preliminary study. In *Proceedings*

of the 2023 7th International Conference on Natural Language Processing and Information Retrieval, NLPIR '23, page 19–24, New York, NY, USA. Association for Computing Machinery.

Shervin Minaee, Elham Azimi, and AmirAli Abdolrashidi. 2019. Deep-sentiment: Sentiment analysis using ensemble of cnn and bi-lstm models. *arXiv preprint arXiv:1904.04206*.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Muhammad Mujahid, Khadija Kanwal, Furqan Rustam, Wajdi Aljedaani, and Imran Ashraf. 2023. Arabic chatgpt tweets classification using roberta and bert ensemble model. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(8).

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, 11(1):81.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Shivam Sharma, Rahul Aggarwal, and M. Kumar. 2023. Mining twitter for insights into chatgpt sentiment: A machine learning approach. *2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, pages 1–6.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melissa Hall Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Abinash Tripathy, Ankit Agrawal, and Santanu Kumar Rath. 2016. Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, 57:117–126.

A Upadhye. 2024. Sentiment analysis using large language models: Methodologies, applications, and challenges. *Int. J. Comput. Appl*, 186:30–34.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Jieying Xue, Minh-Phuong Nguyen, Blake Matheny, and Le-Minh Nguyen. 2024. Bioserc: Integrating biography speakers supported by llms for erc tasks. In *International Conference on Artificial Neural Networks*, pages 277–292. Springer.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of*

the 2023 Conference on Empirical Methods in Natural Language Processing, pages 7915–7927, Singapore. Association for Computational Linguistics.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Attention-based lstm network for cross-lingual sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 247–256.

A Examples of prompting

system
You are an expert in analyzing the emotions expressed in a natural sentence. The emotional label set includes {anger, fear, joy, sadness, surprise}. Each sentence may have one or more emotional labels, or none at all.
user
Given the sentence: "`bro dont do this to us`", which emotions and their corresponding intensities are expressed in it?
assistant
`fear`

system
You are an expert in analyzing the emotions expressed in a natural sentence. The emotional label set includes {anger, fear, joy, sadness, surprise}. Each sentence may have one or more emotional labels, or none at all.
user
Given the sentence: "`I could not unbend my knees.`", is the emotion anger expressed in it?
assistant
`No`

Table 5: Instruction prompting template in track A of *base* (top) and *pairwise* (bottom) strategies, respectively.

system
You are an expert in analyzing the emotions expressed in a natural sentence. The emotional label set includes {anger, fear, joy, sadness, surprise}, with three levels of intensity: low, moderate, and high. Each sentence may have one or more emotional labels, or none at all.
user
Given the sentence: "`A penny hit me square in the face.`", which emotions and their corresponding intensities are expressed in it?
assistant
`moderate degree of anger, low degree of sadness`

system
You are an expert in analyzing the emotions expressed in a natural sentence. The emotional label set includes {anger, fear, joy, sadness, surprise}, with three levels of intensity: low, moderate, and high. Each sentence may have one or more emotional labels, or none at all.
user
Given the sentence: "`Totally creeped me out.`", what is the intensity of the emotion fear expressed in it?
assistant
`high`

Table 6: Instruction prompting template in track B of *base* (top) and *pairwise* (bottom) strategies, respectively.