GT-NLP at SemEval-2025 Task 11: EmoRationale, Evidence-Based Emotion Detection via Retrieval-Augmented Generation

Daniel Saeedi[1], Alireza Kheirandish[1], Sirwe Saeedi[2], Hossein Sahour[3], Aliakbar Panahi[4], Iman Ahmadi Naeeni[5]

[1]Georgia Institute of Technology [2]Johnson & Johnson [3]Novastraum LLC
[4]C3 AI [5]Georgia State University

Correspondence:
dsaeedi3@gatech.edu

Abstract

Despite advancement in large language models (LLMs), emotion detection in multilingual settings remains challenging especially in low-resource languages with limited labeled datasets. In this research, we introduce *EmoRationale*, a novel framework leveraging Retrieval-Augmented Generation (RAG) to address cross-lingual generalization in emotion detection. We combined vector-based retrieval with in-context learning in LLMs, using semantically relevant examples to enhance classification accuracy and interpretability. This system provides evidence-based reasoning for its predictions, making emotion detection more transparent and adaptable across diverse linguistic contexts. Experimental results on the SemEval-2025 Task 11 dataset demonstrate that our RAG-based method achieves strong performance in multi-label emotion classification, emotion intensity assessment, and cross-lingual emotion transfer, surpassing conventional models in interpretability while remaining cost-effective. We release our code and model implementation to facilitate further research[1].

1 Introduction

Emotions are the unseen threads that weave together human communication, influencing decision-making, social interactions, and personal well-being (Barrett and Russell, 2014). With the rapid advancement of large language models (LLMs) and the increasing sophistication of AI-driven text analysis (Team et al., 2024; DeepSeek-AI, 2024; Fatemi and Hu, 2024), understanding emotions embedded within text has become critically important across multiple domains including mental health monitoring, customer engagement, and human-computer interaction (Hong et al., 2025). While LLMs have significantly improved sentiment-aware applications, their effectiveness across diverse linguistic and cultural contexts remains inconsistent (Tafreshi et al., 2024). Existing research has primarily focused on high-resource languages, leaving significant gaps in emotion detection for low-resource languages due to a lack of annotated datasets. This disparity hinders the development of inclusive and effective emotion-aware applications (Feng and Narayanan, 2023). To address these shortcomings, BRIGHTER (BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages) takes a major step forward, particularly for low-resource languages to collect multilabeled emotion-annotated datasets particularly for 28 different low-resource languages (Muhammad et al., 2025a; Belay et al., 2025). This paper presents *EmoRationale*, an AI-driven framework designed for the SemEval-2025 shared task "Bridging the Gap in Text-Based Emotion Detection" (Muhammad et al., 2025a), addressing challenges in multi-label emotion detection, emotion intensity assessment, and cross-lingual emotion detection. Leveraging Retrieval-Augmented Generation (RAG), our system integrates the retrieval capabilities of vector databases such as FAISS (Douze et al., 2024) with the generative power of LLMs like the multilingual MiniLM sentence transformer (Reimers and Gurevych, 2019) to incorporate relevant contextual information and bridge gaps in low-resource languages. By combining few-shot prompting with retrieval-augmented generation, *EmoRationale* produces accurate emotion predictions accompanied by explicit, evidence-based reasoning, with extensive ablation studies demonstrating that incorporating in-context examples significantly boosts performance—surpassing traditional fine-tuning approaches such as those based on RoBERTa (Liu et al., 2019). Moreover, our framework exhibits notable cross-lingual transfer capabilities and cost-effectiveness, paving the way for more inter-

[1]https://github.com/daniel-saeedi/SemEvalTask11_EmoRationale

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 640–650
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

pretable and efficient emotion recognition systems in multilingual settings.

2 Background

Emotion detection in text is a multifaceted area of research within natural language processing (NLP) that seeks not only to identify basic sentiment but also to capture the rich spectrum of human emotions. Previous research mainly focused on sentiment analysis—differentiating between positive and negative affect (Dang et al., 2020)—but subsequent research has expanded the scope to include discrete emotions such as joy, sadness, anger, fear, surprise, and disgust (Mohammad et al., 2018) (Gupta et al., 2018). While LLMs trained on high-resource languages such as English benefit from extensive corpora, their performance in low-resource settings is hindered by data scarcity, inadequate tokenization, and the difficulty of capturing nuanced emotional cues unique to these languages. BERT (Devlin et al., 2019) has revolutionized NLP by providing deep contextualized representations that capture subtle semantic nuances in text (Abas et al., 2021). Fine-tuning BERT on emotion-labeled datasets enhances its effectiveness in tasks such as multi-label emotion classification and emotion intensity assessment by enabling the model to learn complex emotion-specific features(Qin et al., 2023). These capabilities have been further enhanced through the adoption of multilingual variants of BERT, which facilitate cross-lingual emotion detection by leveraging shared representations learned from extensive and diverse corpora (Hassan et al., 2022). Recent advancements in Retrieval-Augmented Generation (RAG) (Lewis et al., 2021) have introduced a promising avenue for enhancing emotion recognition systems. RAG frameworks leverage retrieval mechanisms to access semantically relevant examples from external knowledge bases, integrating this contextual information with the generative capabilities of large language models (LLMs) to improve accuracy and robustness. Additionally, Chain-of-Thought (CoT) prompting has significantly enhanced LLM reasoning and interpretability, enabling models to decompose complex NLP tasks into structured intermediate steps. This approach has demonstrated notable improvements in various applications, including emotion detection, by guiding models to explicitly articulate their decision-making process (Wei et al., 2023). This evolution underscores the potential of RAG and CoT-based methodologies in advancing emotion-aware AI systems and bridging gaps in interpretable affective computing (Bhaumik and Strzalkowski, 2024).

3 Tasks

In SemEval Task 11 (Muhammad et al., 2025a,b; Belay et al., 2025), participants are invited to tackle one or more challenges that address the complex and nuanced nature of emotional expression in text. Below, we detail the three tracks of the shared task. In Track A, the objective is to determine the set of emotions that the speaker conveys in a given text snippet. Each snippet is labeled with one or more of the following emotions: joy, sadness, fear, anger, surprise, and disgust. Track B builds on the multi-label classification framework by incorporating emotion intensity. In this task, for each text snippet paired with a target emotion, the objective is to predict an ordinal intensity value that indicates the strength of the expressed emotion. Intensity is scaled from 0 (none) to 3 (high). Track C addresses the challenge of cross-lingual emotion detection. In this track, participants are provided with a labeled training set in a source language and are required to predict the perceived emotion labels for texts in a different target language. The label set includes the same six emotions as in Track A.

4 System Overview

EmoRationale employs a retrieval-augmented generation (RAG) framework to classify emotions in text. Our system is composed of three main components: a vector store of labeled training data, a Large Language Model (LLM) prompt that incorporates relevant examples, and an LLM-based inference step.

4.1 Sentence Embedding and Vector Store Construction

As shown in Figure 1(a), we first embed each training text using the multilingual MiniLM sentence transformer (Reimers and Gurevych, 2019). These embeddings capture semantic information in a 384 dimensional vector space. We store the resulting vectors, along with the original text and emotion labels, in a FAISS (Facebook AI Similarity Search) vector store (Douze et al., 2024). This vector store serves as a searchable database that allows for efficient similarity-based lookups. When a new (unlabeled) text is presented as input, we embed it

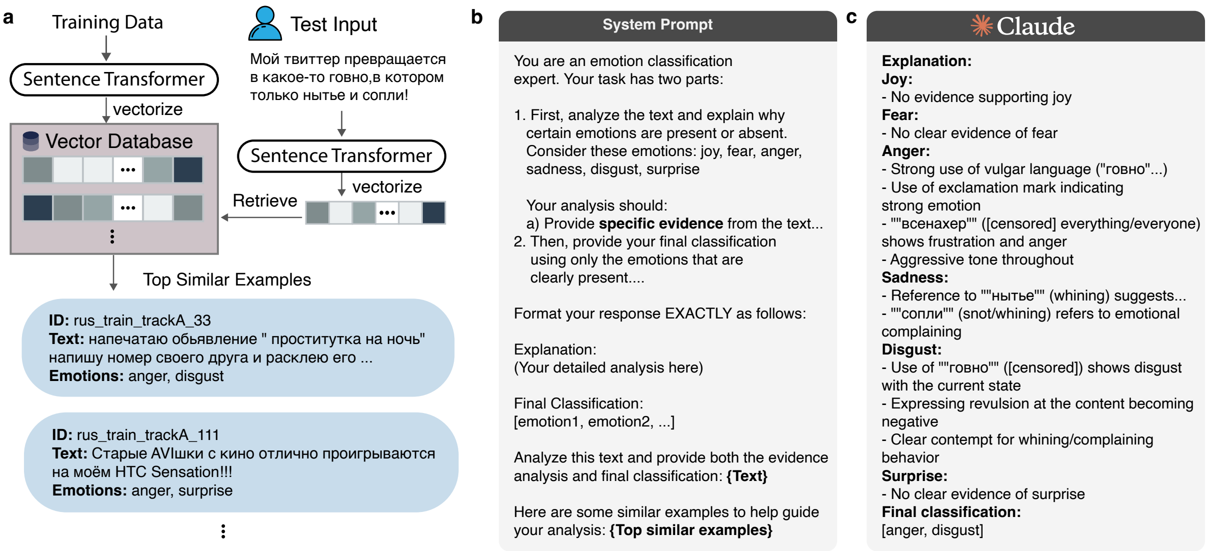

Figure 1: Overview of *EmoRationale*. The system is divided into three primary components. **a.** employs a sentence transformer to convert the training dataset into a vector database; during inference, the input text is vectorized and matched against the database using cosine similarity to retrieve the most similar examples. **b.** displays the system prompt for classification, where the model is first asked to provide evidence-based reasoning for the presence or absence of specific emotions, followed by a final classification. Note that the prompt includes top similar examples to guide the analysis. **c.** shows the output produced by Claude Sonnet 3.5 for the test input featured in panel (a).

using the same sentence transformer model. We then query the FAISS vector store to retrieve the k most similar examples from the training data. In our experiments, we use k=5. Each retrieved example includes its text and the associated ground truth emotion labels.

4.2 Few-shot prompting

Our method constructs a system prompt by first providing clear instructions on how to detect evidence supporting or contradicting each emotion (Figure 1**a**), followed by appending the most relevant examples from the training set as demonstrations. These instructions guide the LLM to methodically analyze linguistic cues and contextual hints, while the retrieved examples offer a concrete reference for how each emotion was identified in similar texts (see Figure 1**b** and appendix A for complete system propmpt). During inference, the LLM is prompted to produce both a concise “Explanation,” highlighting which features of the text suggest or rule out particular emotions, and a final bracketed “Classification” (e.g., [joy, fear]). If no emotion is detected, it outputs [none]. This two-part structure provides transparency and interpretability, enabling the system to provide both the reasoning behind its predictions and the specific emotional categories that apply (Figure 1**c**).

4.3 RoBERTa fine-tune

For baseline, we use a pre-trained *RoBERTa-base* model that is fine-tuned for the multi-label classification task. The model architecture is modified by adding an intermediate fully connected layer followed by a ReLU activation and layer normalization, with dropout applied both before and after these layers to mitigate overfitting. A mean-pooled embedding over all token representations is computed, and this pooled embedding is fed into the final classification layer, which maps the processed features to the required number of output classes (six for tracks A and C, and 24 for track B), producing raw logits for each label. The model is fine-tuned using a binary cross-entropy loss with logits, and the AdamW optimizer is employed with differentiated learning rates—a base learning rate of 2×10^{-5} for the pre-trained RoBERTa parameters and a higher learning rate of 1×10^{-4} for the newly added layers. Training is performed over 20 epochs with a batch size of 32 for training, validation, and testing.

5 Experimental Setup

In our experiments, we utilized a T4 GPU from Colab Pro for the multilingual MiniLM sentence transformer. We also conducted evaluations using DeepSeek R1(DeepSeek-AI et al., 2025) with the

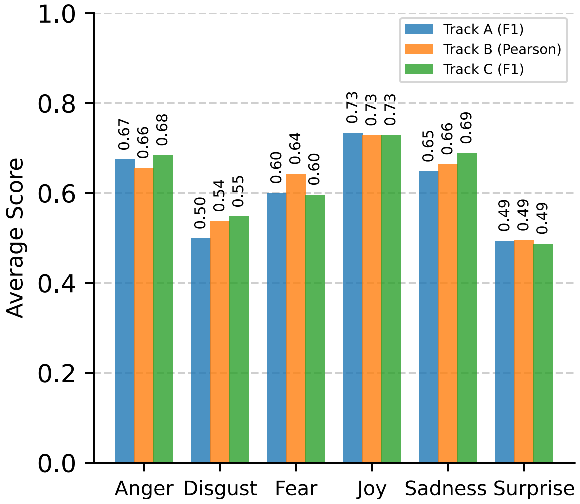

Figure 2: **Average performance scores across emotions for all three tracks in the SemEval 2025 Task 11.** Track A and Track C use F1-scores while Track B uses Pearson correlation coefficients. Joy consistently achieves the highest scores (≈ 0.73) across all tracks, while Surprise shows the lowest performance (≈ 0.49). Anger and Sadness also demonstrate strong performance, while Disgust exhibits the second-lowest scores among the emotions.

685B model for track A, DeepSeek V3 (DeepSeek-AI, 2024) for track B, and Claude Sonnet 3.5 (Anthropic) for track C. The total cost for development and evaluation on the test set amounted to $80.65 for Claude Sonnet 3.5, whereas DeepSeek R1 cost $22.22. However, one notable shortcoming of the DeepSeek API was its unreliability, frequently experiencing interruptions and errors due to high demand.

6 Results

We evaluated our system on all three tracks of SemEval-2025 Task 11: Multi-label Emotion Detection (Track A), Emotion Intensity (Track B), and Cross-lingual Emotion Detection (Track C). Below, we detail our overall performance, present findings from ablation studies, and discuss the impact of our retrieval-augmented generation (RAG) approach. Across all three tracks, models consistently exhibited lower performance for the emotions Surprise (approximately 0.49) and Disgust (approximately 0.53), as illustrated in Figure 2. Detecting surprise and disgust poses significant challenges, primarily because these emotions often depend on nuanced, context-specific, and culturally influenced expressions, which are difficult to interpret accurately through textual analysis alone. Unlike more explicit emotions such as anger or joy, surprise and disgust typically lack distinct verbal cues and thus require additional contextual understanding (Mohammad et al., 2018; Demszky et al., 2020). Furthermore, emotions like surprise can convey either positive or negative sentiments, while disgust may frequently be misinterpreted as anger or fear. Such inherent ambiguities and overlaps further complicate accurate emotion classification for language models.

6.1 Track A: Multi-label Emotion Detection

Table 1 presents the emotion recognition performance for Track A across eight languages. Our system achieves strong and consistent results, with average macro-F1 scores ranging from 0.519 (AFR) to 0.864 (RUS). Notably, the system exhibits high accuracy on languages such as Russian, while maintaining competitive performance on others despite linguistic diversity. Ablation studies further highlight the benefits of incorporating few-shot prompting with relevant examples (Table S1). When compared to the RoBERTa baseline (which was fine-tuned on the entire training dataset), configurations using retrieval-augmented prompts lead to substantial improvements. Although Claude Sonnet 3.5 with RAG provides competitive results, DeepSeek R1 with RAG was ultimately selected for Track A due to its comparable performance and approximately 20-fold cost reduction.

6.2 Track B: Emotion Intensity

For Track B, our task is to predict ordinal emotion intensity values, which introduces additional complexity. As shown in Table 1, performance on intensity estimation varies considerably across languages. For instance, while Russian texts achieve an impressive average macro-F1 of 0.880, other languages such as Ukrainian yield lower scores (0.319).

6.3 Track C: Cross-lingual Emotion Detection

Table 1 also summarizes the results for Track C, which examines the transferability of emotion detection across languages. In this setting, a Portuguese/Brazilian training set is used to predict emotions in Spanish, Russian, and Mandarin. Our experiments show that even with only five in-context examples, the retrieval-augmented prompting enables effective generalization across languages. Ablation studies (see Table S3) reveal that while configurations with full explanation (i.e., reasoning

Track	Language	Anger	Disgust	Fear	Joy	Sadness	Surprise	Average
	Afrikaans (AFR)	0.468	0.458	0.520	0.625	0.525	-	0.519
	German (DEU)	0.805	0.722	0.580	0.758	0.684	0.428	0.663
	Portuguese (PTBR)	0.758	0.229	0.586	0.780	0.716	0.526	0.599
A	Russian (RUS)	0.891	0.804	0.954	0.890	0.772	0.871	0.864
	Algerian Arabic (ARQ)	0.626	0.379	0.530	0.542	0.618	0.468	0.527
	Moroccan Arabic (ARY)	0.641	0.311	0.507	0.692	0.635	0.402	0.531
	Swedish (SWE)	0.713	0.633	0.369	0.920	0.541	0.211	0.564
	Ukrainian (UKR)	0.491	0.454	0.753	0.663	0.692	0.547	0.600
	German (DEU)	0.549	0.454	0.370	0.583	0.546	0.446	0.491
	English (ENG)	0.787	-	0.754	0.801	0.794	0.619	0.751
	Spanish (ESP)	0.723	0.699	0.826	0.791	0.820	0.689	0.758
B	Portuguese (PTBR)	0.734	0.361	0.651	0.782	0.749	0.420	0.616
	Russian (RUS)	0.900	0.849	0.927	0.896	0.848	0.861	0.880
	Algerian Arabic (ARQ)	0.618	0.440	0.573	0.647	0.462	0.426	0.528
	Chinese (CHN)	0.746	0.407	0.502	0.844	0.607	0.359	0.577
	Hausa (HAU)	0.515	0.769	0.669	0.695	0.709	0.440	0.633
	Ukrainian (UKR)	0.333	0.245	0.334	0.314	0.377	0.310	0.319
	Romanian (RON)	0.650	0.612	0.819	0.930	0.727	0.374	0.685
	Afrikaans (AFR)	0.605	0.518	0.507	0.447	0.611	-	0.538
	German (DEU)	0.821	0.777	0.564	0.776	0.747	0.437	0.687
	Hindi (HIN)	0.825	0.665	0.866	0.834	0.804	0.705	0.783
C	Indonesian (IND)	0.586	0.533	0.500	0.808	0.743	0.328	0.583
	Russian (RUS)	0.857	0.690	0.890	0.906	0.763	0.730	0.806
	Algerian Arabic (ARQ)	0.575	0.385	0.623	0.543	0.618	0.516	0.543
	Moroccan Arabic (ARY)	0.646	0.235	0.453	0.731	0.688	0.459	0.535
	Swedish (SWE)	0.732	0.697	0.330	0.876	0.544	0.270	0.575
	Ukrainian (UKR)	0.508	0.426	0.633	0.638	0.678	0.451	0.556

Table 1: **Test Performance.** This table presents our results on SemEval 2025 Task 11. Track A and Track C scores are reported as F1-scores, while Track B scores are measured using Pearson correlations. For Track A, we utilized DeepSeek R1 (685B parameters), for Track B, we employed DeepSeek V3 (685B), and for Track C, we used Claude 3.5 Sonnet.

for predictions) provide interpretability, prompt variants that omit the reasoning component can yield a modest performance boost—reaching an average macro-F1 of 0.755. In the cross-lingual setting, even with training data from a different language (Portuguese/Brazilian), the system successfully generalizes to Spanish, Russian, and Mandarin. This highlights the model's ability to effectively transfer emotional recognition capabilities across languages using only a few in-context examples, while the fine-tuned RoBERTa-base model struggles to generalize learned knowledge from one language to another.

7 Conclusion and Future Work

In this work, we introduced *EmoRationale*, a retrieval-augmented generation (RAG) framework for explainable emotion recognition that addresses the challenges posed by multilingual and nuanced emotional expressions in text. By integrating robust sentence embeddings with a FAISS vector store and leveraging few-shot in-context learning, *EmoRationale* offers significant advantages in interpretability and cross-lingual transfer—even if it may not achieve the highest raw predictive metrics. The framework provides explicit, evidence-based reasoning for its predictions while effectively generalizing emotion recognition across languages using only a few in-context examples—a feat that conventional fine-tuning approaches, such as RoBERTa-base, often struggle to achieve.

References

Ahmed R. Abas, Ibrahim Elhenawy, Mahinda Zidan, and Mahmoud Othman. 2021. Bert-cnn: A deep learning model for detecting emotions from text. *Computers, Materials and Continua*, 71(2):2943–2961.

Anthropic. The claude 3 model family: Opus, sonnet, haiku.

Lisa Feldman Barrett and James A Russell. 2014. *The psychological construction of emotion*. Guilford Publications.

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.

Ankita Bhaumik and Tomek Strzalkowski. 2024. Towards a generative approach for emotion detection and reasoning. *Preprint*, arXiv:2408.04906.

Nhan Cach Dang, María N Moreno-García, and Fernando De la Prieta. 2020. Sentiment analysis based on deep learning: A comparative study. *Electronics*, 9(3):483.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *Preprint*, arXiv:2005.00547.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Sorouralsadat Fatemi and Yuheng Hu. 2024. Enhancing financial question answering with a multi-agent reflection framework. In *Proceedings of the 5th ACM International Conference on AI in Finance*, ICAIF '24, page 530–537, New York, NY, USA. Association for Computing Machinery.

Tiantian Feng and Shrikanth Narayanan. 2023. Foundation model assisted automatic speech emotion recognition: Transcribing, annotating, and augmenting. *Preprint*, arXiv:2309.08108.

Umang Gupta, Ankush Chatterjee, Radhakrishnan Srikanth, and Puneet Agrawal. 2018. A sentiment-and-semantics-based approach for emotion detection in textual conversations. *Preprint*, arXiv:1707.06996.

Sabit Hassan, Shaden Shaar, and Kareem Darwish. 2022. Cross-lingual emotion detection. *Preprint*, arXiv:2106.06017.

Xin Hong, Yuan Gong, Vidhyasaharan Sethu, and Ting Dang. 2025. Aer-llm: Ambiguity-aware emotion recognition leveraging large language models. *Preprint*, arXiv:2409.18339.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

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, New Orleans, Louisiana. Association for Computational Linguistics.

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, and 29 others. 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, and 2 others. 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.

Xiangyu Qin, Zhiyu Wu, Jinshi Cui, Tingting Zhang, Yanran Li, Jian Luan, Bin Wang, and Li Wang. 2023. Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation. *Preprint*, arXiv:2301.06745.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Shabnam Tafreshi, Shubham Vatsal, and Mona Diab. 2024. Emotion classification in low and moderate resource languages. *arXiv preprint arXiv:2402.18424*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1331 others. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

A Appendix

Language	RoBERTa Baseline	Claude No Examples	Claude Random Ex.	Claude w/ RAG	Claude No Reason.	DeepSeek-R1 w/ RAG
ESP	0.797	0.871	0.873	**0.883**	0.882	0.876
PTBR	0.648	0.738	0.743	0.754	0.772	**0.779**
RUS	0.787	0.839	0.875	0.885	**0.914**	0.909
Avg	0.744	0.815	0.830	0.840	**0.856**	0.854

Table S1: **F1 scores across six emotions are presented for each ablation study in Track A.** A combined training dataset from three languages (Spanish, Portuguese/Brazilian, and Russian) was used, with evaluation on corresponding development sets of approximately 200 examples each. Six configurations were assessed: (a) fine-tuning RoBERTa-base on the combined dataset; (b) Claude Sonnet 3.5 without examples (examples refer to data from the combined training dataset); (c) Claude Sonnet 3.5 with random examples; (d) Claude Sonnet 3.5 with relevant examples; (e) Claude Sonnet 3.5 without analysis or explanation for predictions; and (f) DeepSeek R1 with relevant examples. A notable pattern is that, despite providing only five examples to both Claude Sonnet 3.5 and DeepSeek R1 while RoBERTa-base fine-tuning uses the entire training dataset, these models outperform the latter by a large margin. Moreover, they offer interpretability by providing explanations for the presence or absence of an emotion. For Track A, DeepSeek R1 was selected owing to its cost-effectiveness—approximately 20-fold cheaper than Claude Sonnet 3.5—while delivering comparable performance.

Language	RoBERTa Base	DeepSeek v3 No RAG	DeepSeek v3 Random Ex.	DeepSeek v3 w/ RAG	DeepSeek v3 No Reason. w/ RAG
ESP	0.694	0.701	0.724	**0.886**	0.763
CHN	0.562	0.505	0.549	**0.604**	0.575
ARQ	0.486	0.524	0.531	0.544	**0.552**
Avg	0.578	0.577	0.601	**0.678**	0.631

Table S2: **Average Pearson correlation coefficients across six emotional categories for each ablation study in Track B.** Evaluations were conducted on Spanish, Chinese, and Algerian Arabic datasets. Configurations include: (a) RoBERTa-base fine-tuned on the specified data; (b) DeepSeek v3 without examples (where examples refer to data from the training dataset); (c) with random examples; (d) with relevant examples; and (e) without prediction analysis/explanation. Despite using only five examples, DeepSeek v3 outperformed the fine-tuned RoBERTa-base by 20% in average Pearson correlation coefficient score, while also providing interpretable explanations for emotion presence or absence.

Language	RoBERTa Base	Claude No RAG	Claude Random Ex.	Claude w/ RAG	Claude No Reason. w/ RAG
ESP	0.519	0.865	**0.866**	0.864	0.864
RUS	0.527	0.840	0.864	0.871	**0.905**
CHN	0.462	0.474	0.489	0.481	**0.495**
Avg	0.503	0.726	0.739	0.739	**0.755**

Table S3: **F1 scores across six emotions are presented for each ablation study in Track C.** Evaluations were performed using a Portuguese/Brazilian training set, with development sets in Spanish (ESP), Russian (RUS) and Mandarin (CHN) (approximately 200 examples each). Configurations include: (a) RoBERTa-base fine-tuned on the Portuguese/Brazilian data; (b) Claude Sonnet 3.5 without examples (examples refer to data from the training dataset of Portuguese/Brazilian); (c) with random examples; (d) with relevant examples; and (e) without prediction analysis/explanation. Despite using only five examples, Claude Sonnet 3.5 exceeded the fine-tuned RoBERTa-base by 47% in average macro-F1, while also providing interpretable explanations for emotion presence or absence. This demonstrates Claude Sonnet 3.5 capacity to generalize and transfer emotion detection capabilities across languages using training data from a different language.

System prompt for track A and C

You are an emotion classification expert. Your task has two parts:

1. First, analyze the text and provide evidence for and against the presence of each emotion:
   - **joy**
   - **fear**
   - **anger**
   - **sadness**
   - **disgust**
   - **surprise**

   Guidelines for analysis:

   - For each emotion, provide specific evidence from the text that supports or contradicts its presence.
   - Consider both explicit emotional words and contextual implications.
   - Base your analysis on linguistic patterns, word choice, and context.
   - Be objective in your analysis.
2. Then, based on your analysis, provide your final classification in the format: [emotion1, emotion2, ...]
   - Only include emotions that are clearly present.
   - Use only these emotions: joy, fear, anger, sadness, disgust, surprise, none.
   - Do not explain your choice, just provide the list.

Format your response EXACTLY as follows:
Explanation: (Your detailed analysis here)

Final Classification:
[emotion1, emotion2, ...]

Here are some similar examples to help guide your analysis:
[SELECTED EXAMPLES GOES HERE]

Analyze this text and provide both the evidence analysis and final classification:
[User Text]

Claude 3.5 Sonnet output for track C

Input text snippet:
Ist das nicht schön? Der wichtigste Beitrag der Ukraine für weniger Tote ist deren Korruption. Keine Pointe.

English Translation:
Isn't that beautiful? Ukraine's most important contribution to fewer deaths is its corruption. No punchline.

Claude 3.5 Sonnet Output:

1. **Evidence Analysis:**
   1. **Joy:**
      - **Against:** While "schön" (beautiful) appears, it's used sarcastically in a rhetorical question.
      - **Against:** Overall tone is cynical rather than joyful.
   2. **Fear:**
      - **Against:** No expressions of fear or threat present.
   3. **Anger:**
      - **For:** The sarcastic tone and cynical commentary about corruption suggests underlying anger.
      - **For:** The rhetorical question "Ist das nicht schön?" shows bitter sarcasm indicating anger.
      - **For:** The mention of "Tote" (dead) in connection with corruption implies angry criticism.
   4. **Sadness:**
      - **For:** Reference to deaths ("Tote") suggests a somber, sad context.
      - **For:** The cynical observation about corruption contributing to deaths carries a sad undertone.
   5. **Disgust:**
      - **For:** The sarcastic comment about corruption suggests moral disgust.
      - **For:** The connection between corruption and deaths implies disgust at the situation.
   6. **Surprise:**
      - **Against:** No elements of surprise present in the text.
2. **Final classification:**
   [anger, disgust, sadness]

System prompt for track B

You are an emotion classification expert. Your task has two parts:

1. First, analyze the text and explain why certain emotions are present or absent, and provide an intensity level (from none to very high) for each emotion. Consider these emotions: joy, fear, anger, sadness, disgust, surprise

   Your analysis should:

   - Provide specific evidence from the text
   - Consider both explicit words and contextual implications
   - Be objective and clear
   - Assign an intensity level for each emotion: none, low, moderate, high, very high

2. Then, provide your final classification by listing all detected emotions along with their intensity levels. Use only these emotions: joy, fear, anger, sadness, disgust, surprise, and the intensity levels: none, low, moderate, high, very high.

IMPORTANT

The examples below are provided to help guide your reasoning. They contain insights and annotations from experts who labeled the dataset. Pay close attention to how emotions and their intensities were derived in these examples, and use this understanding to inform your own analysis.

Format your response EXACTLY as follows:
Explanation: (Your detailed analysis here)

Final Classification:
[emotion1, emotion2, ...]

Here are some similar examples to help guide your analysis:
[SELECTED EXAMPLES GOES HERE]

Analyze this text and provide both the evidence analysis and final classification:
[User Text]