OPI-DRO-HEL at SemEval-2025 Task 11: Few-shot prompting for Text-based Emotion Recognition

Daniel Karaś and Martyna Śpiewak
National Information Processing Institute
Warsaw, Poland
{dkaras, mspiewak}@opi.org.pl

Abstract

This paper presents our system, developed as our contribution to SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection task (Muhammad et al., 2025b), in particular track A, Multi-label Emotion Detection subtask. Our approach relies on two distinct components: semantic search for top N most similar inputs from training set and an interface to pretrained LLM being prompted using the found examples. We examine several prompting strategies and their impact on overall performance of the proposed solution.

1 Introduction

Emotions are inherently complex and multifaceted, influencing daily interactions while often remaining challenging to articulate and interpret. Individuals employ language in intricate and nuanced ways to convey emotions, with expression and perception varying across linguistic and cultural contexts, as well as within the same societal or social group. This paper presents our system proposed as a solution to Task 11, track A of Semeval-2025 competition, which focuses on detecting perceived emotions, i.e., what emotion most people think the speaker might have felt given a sentence or a short text snippet (Muhammad et al., 2025a).

The recognition of emotions by machine learning (ML) systems has been an active area of research for several decades, with approaches evolving from rule-based models to neural networks such as Long Short-Term Memory (LSTM) (Gupta et al., 2024).

The advent of large language models (LLMs) has introduced significantly more complex architectures, which demonstrated efficacy in classical natural language processing downstream tasks but also phenomenons such as emergent abilities (Wei et al., 2022). The effectiveness of LLMs in emotion detection is further supported by benchmarks such as SemEval-2024 Task 10, where LLMs were widely adopted by participants (Kumar et al., 2024). Given their demonstrated performance, LLMs are now regarded as state-of-the-art solutions in this domain. Consequently, this study leverages 70B variant of Deepsek R1 LLM to solve given task using several prompting strategies such as chain-of-thought and few-shot-prompting with examples for in-context-learning being provided by an information retrieval subsystem based on embeddings generated by a fine-tuned RoBERTa encoder (Liu et al., 2019).

2 Task and dataset

The task at hand is an instance of classical multi-labeled text classification task with the set of labels spanning six categories of perceived emotions: anger, sadness, fear, disgust, joy, surprise, which align with Ekman's six basic emotions. Text snippets were mostly extracted from social media webpages such as Reddit, Youtube and Twitter among others.

For instance, the sentence "That was the last time anyone saw her." was annotated with "fear" and "sadness".

The organizers provided a separate dataset for each of 28 different languages from 7 language families, each dataset was further divided into train, dev and test splits. In this study, we focused on developing a solution for the English subset of the Track A dataset, which consists of 2,768 training examples, 116 development examples, and 2,767 test examples.

An analysis of the label distribution, as presented in the task dataset description paper (Muhammad et al., 2025a), suggests that class imbalance may introduce additional challenges. Specifically, the most frequent class, fear, appears in 3,218 instances, whereas the least frequent class, anger, is present in only 671 examples. Additionally, only

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1233–1240
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

545 instances do not belong to any of the six predefined emotion categories, classifying them as neutral.

The official evaluation metric selected by the organizers was the macro-averaged F1-score, computed based on the predicted and gold-standard labels.

3 Experiments

In the following section, we present results exclusively on the test dataset to ensure a consistent and reliable point of reference, which most closely aligns with the final ranking. All reported results are based on the official macro-averaged F1-score; therefore, unless otherwise specified, this metric should be assumed by default.

3.1 Baseline

At the time of developing our system, the baseline results provided by the organizers had not yet been published in the task ranking. Consequently, we sought an appropriate off-the-shelf candidate to serve as a baseline upon which improvements could be made. In our preliminary study, we identified the pretrained RoBERTa-based model, *j-hartmann/emotion-english-roberta-large* (Hartmann, 2022), as a suitable candidate. This decision was based on the fact that the model was pretrained on the same set of emotion labels as those used in this task, with the addition of a "neutral" category, which could be interpreted as the absence of any assigned emotion label. Furthermore, the training data for this model predominantly consisted of social media posts (e.g., Reddit, YouTube, Twitter), which closely resemble the characteristics of the dataset used in this task. Given these factors, we hypothesized that the model would generalize effectively to previously unseen data of a similar nature.

We employed the pretrained model using the Hugging Face text-classification pipeline (Hug, accessed February 27, 2024) and applied a fixed threshold to convert the obtained softmax probability distribution into the expected binary classification format. The final threshold value of 0.26 was determined through a basic grid search.

Upon obtaining the performance results of the official baseline solution, which was based on RemBERT and achieved a macro-averaged F1-score of 0.7083, it became evident that our selected baseline was significantly weaker, reaching only 0.4472.

3.2 RoBERTa fine-tuning

To determine whether the poor performance stemmed from the pretrained model's inability to generalize to unseen data from a different distribution or from inherent limitations of its architecture, we proceeded with fine-tuning the model on the training split of this task.

The model was trained for two epochs using the AdamW optimizer, a learning rate of 5.49e-05 and a batch size of 8, with hyperparameters optimized using the Optuna framework (Akiba et al., 2019). Adapting the competition dataset to the format expected by the pretrained model was relatively straightforward; the primary modification involved mapping instances without assigned labels to the "neutral" category. Additionaly, model's vocabulary was extended with unseen words present in task's dataset.

Fine-tuning the model led to a substantial improvement, yielding a macro-averaged F1-score of 0.6915.

3.3 Large language models

We hypothesized that the fine-tuned RoBERTa model had reached its performance limits and that further improvements would not be achievable without the introduction of additional data, likely through augmentation techniques. Given this constraint, we opted to explore the performance of large language models (LLMs) in a zero-shot or few-shot setting to assess their "out-of-the-box" effectiveness on the task.

We conducted an evaluation of several smaller large language models within the 8B–14B parameter range using the development dataset. Additionally, we assumed that the performance of these smaller models could serve as an indicator of their larger counterparts' capabilities. This approach enabled rapid iteration in a local environment. Additionally, the integration of tools such as Ollama (oll, accessed February 27, 2024) and LangChain (Lan, accessed February 27, 2024) streamlined the interaction with the models, allowing us to focus on experimental evaluations rather than addressing technical implementation challenges.

The evaluated models included Teuken-7B (Ali et al., 2024), Vicuna-13B (Chiang et al., 2023), LLaMA 3.1-8B (Grattafiori et al., 2024), and DeepSeek-R1-14B (DeepSeek-AI, 2025). However, these smaller models frequently exhibited issues such as hallucination of labels, failure to

Model	Prompt	N-shot	Chain-of-thought	F1-score
emotion-english-roberta-large (baseline)	-	-	-	0.4472
emotion-english-roberta-large (fine-tuned)	-	-	-	0.6915
RemBERT (official baseline)	-	-	-	0.7083
Deepseek-R1	unstructured	Zero	No	0.7307
Deepseek-R1	unstructured	Few	No	0.7356
Deepseek-R1	structured	Few	No	0.7159
Deepseek-R1	structured	Few	Yes	0.7039

Table 1: Results on test dataset for english language. N-shot denotes number of examples in the prompt and chain-of-thought marks the usage of said prompting technique, where "-" that it's not relevant for given model, see example prompts in Appendix

adhere to the task as specified in the prompt, or generation of outputs that did not conform to the expected format. Among the evaluated models, DeepSeek-R1 demonstrated the most promising results. Consequently, we proceeded with further evaluation of its larger variant, DeepSeek-R1-70B.

Initially, we aimed to assess the model's performance in a zero-shot setting using an unstructured prompt, which was primarily a paraphrased version of the task formulation. We hypothesized that this approach would serve as a strong baseline for further improvements. This evaluation yielded a promising macro-averaged F1-score of 0.7307. See Figure 1 for example prompt of this type.

3.4 Few-shot prompting

To further enhance this promising result, we implemented well-established prompting techniques, including few-shot prompting (FSP) and chain-of-thought (CoT) reasoning . Both techniques are widely recognized for their ability to potentially improve the performance of LLMs.

To select examples for few-shot learning, we repurposed our fine-tuned RoBERTa-based classifier. While this approach is not necessarily optimal—given that our model was not explicitly trained for metric learning tasks—it provided a practical means of example selection. We identified the top N examples by computing the cosine similarity between embeddings generated through mean pooling. Our underlying assumption was that this method would allow us to retrieve examples that are not only semantically similar but also aligned in terms of emotional labeling (i.e., associated with the same or similar sets of emotions) to the query embedding. The selected examples were drawn from the training set.

The few-shot prompting approach resulted in only a marginal improvement in performance compared to the zero-shot method, achieving an F1-score of 0.7356. While we did not conduct extensive benchmarking on the example selection process, a qualitative assessment suggests that the selected examples were generally relevant to the query. Therefore, we hypothesize that the limited performance gain is not primarily due to deficiencies in the example selection pipeline. Instead, we attribute this outcome to either a suboptimal choice of the number of shots or an insufficient model size.

The authors of (Brown et al., 2020) demonstrate that model performance tends to improve with increasing model scale, with the FSP approach exhibiting a more rapid performance gain compared to the zero-shot method. This suggests that increasing the parameter count of the prompted LLM could still have a significant impact on performance. Furthermore, a similar trend is observed with the number of examples: except for very small models (fewer than 2 billion parameters), performance generally improves as the number of shots increases. An example prompt is provided in Figure 2.

However, as highlighted in (Brown et al., 2020), the effectiveness of FSP is also dependent on the specific characteristics of the task. Therefore, it is possible that the task under investigation does not benefit substantially from few-shot prompting.

3.5 Prompt structuring

Additionally, we investigated the impact of structuring and formatting the prompt. Previous studies, such as (Wei et al., 2023) and (He et al., 2024), indicate that large language models (LLMs) can be highly sensitive to prompt formulation, with factors such as the order of few-shot examples and even capitalization influencing performance. An example is provided in Figure 3.

Enhancing the previously described unstructured prompt with additional structure and formatting led to a decrease in performance, yielding an F1-score of 0.7159. This finding is consistent with the observations reported in (He et al., 2024), where markdown formatting resulted in lower performance compared to plain text. However, given the inherent variability in how LLMs respond to prompt formulation, it remains possible that the opposite effect could occur under different downstream task.

3.6 Chain-of-thought

Unfortunately, the use of chain-of-thought prompting proved detrimental to overall performance, yielding an F1-score of 0.7039, which was lower than that of the zero-shot approach. This outcome is not entirely unexpected, as prior research (Wei et al., 2023) has demonstrated that the effectiveness of CoT prompting is highly dependent on model size. Notably, performance can improve significantly when scaling from a 62B model (which is relatively close in scale to our 70B model) to a 540B model.

Moreover, the robustness of CoT prompting is largely task-dependent. While CoT can outperform standard prompting in models as small as 8B for certain tasks, in other cases, a significant performance shift occurs around the 62B model threshold, or the performance of CoT prompting remains comparable to that of non-CoT prompting, regardless of model size. An example CoT prompt is provided in Figure 4.

4 Conclusions and limitations

Consequently, we selected the unstructured few-shot approach with RoBERTa-based semantic search as our final submission. This approach achieved a macro-averaged F1-score of 0.7356, ranking 32nd out of 75 teams and outperforming the official baseline solution, which scored 0.7083.

As discussed in the previous section, we believe that this ranking could be improved with minimal modifications to the overall system while maintaining the existing framework. Specifically, replacing DeepSeek-R1-70B with a larger model, optimizing the number and potentially the order of few-shot examples, and further fine-tuning RoBERTa for the metric learning task could yield performance gains. Furthermore, given the sensitivity of large language models (LLMs) to prompt formulation, refining prompt design presents an additional avenue for optimization and future research.

References

accessed February 27, 2024. Huggingface transformers text classification pipeline. https://huggingface.co/docs/transformers/main_classes/pipelines#transformers.TextClassificationPipeline.

accessed February 27, 2024. Langchain. https://python.langchain.com/docs/integrations/llms/ollama/.

accessed February 27, 2024. Ollama. https://github.com/ollama/ollama.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Jan Ebert, Alexander Arno Weber, Richard Rutmann, Charvi Jain, Max Lübbering, Daniel Steinigen, Johannes Leveling, Katrin Klug, Jasper Schulze Buschhoff, Lena Jurkschat, Hammam Abdelwahab, Benny Jörg Stein, Karl-Heinz Sylla, Pavel Denisov, Nicolo' Brandizzi, Qasid Saleem, Anirban Bhowmick, Lennard Helmer, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Alex Jude, Lalith Manjunath, Samuel Weinbach, Carolin Penke, Oleg Filatov, Shima Asaadi, Fabio Barth, Rafet Sifa, Fabian Küch, Andreas Herten, René Jäkel, Georg Rehm, Stefan Kesselheim, Joachim Köhler, and Nicolas Flores-Herr. 2024. Teuken-7b-base teuken-7b-instruct: Towards european llms. *Preprint*, arXiv:2410.03730.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. Comprehensive study on sentiment analysis: From rule-based to modern llm based system. *Preprint*, arXiv:2409.09989.

Jochen Hartmann. 2022. Emotion english distilroberta-base. `https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/`.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on llm performance? *Preprint*, arXiv:2411.10541.

Shivani Kumar, Md. Shad Akhtar, Erik Cambria, and Tanmoy Chakraborty. 2024. SemEval 2024 - task 10: Emotion discovery and reasoning its flip in conversation (EDiReF). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1933–1946, Mexico City, Mexico. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

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.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

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 Example prompts

Unstructured zero-shot prompt

Given a target text snippet: "/ o So today I went in for a new exam with Dr. Polvi today, I had to file new paperwork for the automobile accident case which is being done differently then the scoliosis stuff. So he comes in and starts talking about insurance stuff and how this look bad since I was getting treatment on my neck and stuff already blah blah.", predict the perceived emotion(s) of the speaker, knowing that target text comes from twitter. Specifically, select whether each of the following emotions apply: joy, sadness, fear, anger or surprise.
In other words, label the text snippet with: joy (1) or no joy (0), sadness (1) or no sadness (0), anger (1) or no anger (0), surprise (1) or no surprise (0).
Output only labels and their corresponding scores (0 or 1) in following format: "Label":Score.

Figure 1: Example unstructured (written in plain, natural language, no formatting) zero-shot (no examples) prompt

Unstructured few-shot prompt

Given a target text snippet: "/ o So today I went in for a new exam with Dr. Polvi today, I had to file new paperwork for the automobile accident case which is being done differently then the scoliosis stuff.So he comes in and starts talking about insurance stuff and how this look bad since I was getting treatment on my neck and stuff already blah blah.", predict the perceived emotion(s) of the speaker, knowing that target text comes from twitter.
Specifically, select whether each of the following emotions apply: joy, sadness, fear, anger or surprise.
In other words, label the text snippet with: joy (1) or no joy (0), sadness (1) or no sadness (0), anger (1) or no anger (0), surprise (1) or no surprise (0).
Output only labels and their corresponding scores (0 or 1) in following format: {"Label":Score}.
Following are examples of similar sentences with assigned labels to help you with labeling:
"i have major headache, just want to sleep all day, and the worst part when i look in the mirror my lips is swollen to like two times the size." has following scores
{ "anger":0,"fear":1,"joy":0,"sadness":1,"surprise":0 }
(...)

Figure 2: Example unstructured (written in plain, natural language, no formatting) few-shot (examples present) prompt. Only one example is included for clarity and brevity.

Structured few-shot prompt

Role:
You are a multilabel classifier predicting the perceived emotion(s) of the author, knowing that text comes from twitter.
Label the text snippet with: joy (1) or no joy (0), sadness (1) or no sadness (0), anger (1) or no anger (0), surprise (1) or no surprise (0), fear (1) or no fear (0).

Pointers:
- Remember that you are predicting emotions of author, not the reader.
- Carefully consider each possible emotion(label).

Constraints:
- Classify text snippet provided in Input section
- Output only labels and their corresponding scores in following format: {"Label":Score}.
- Scores can only be either 0 or 1
- Use same format as provided by examples

Examples:

Example 1

Input:
"i have major headache, just want to sleep all day, and the worst part when i look in the mirror my lips is swollen to like two times the size."

Labels:
{ "anger":0,"fear":1,"joy":0,"sadness":1,"surprise":0 }

Input
"/ o So today I went in for a new exam with Dr. Polvi today, I had to file new paperwork for the automobile accident case which is being done differently then the scoliosis stuff.So he comes in and starts talking about insurance stuff and how this look bad since I was getting treatment on my neck and stuff already blah blah."

Figure 3: Example structured (formatting, clear constraints and instructions) few-shot (examples present) prompt. Only one example is included for clarity and brevity.

Structured few-shot prompt with chain-of-thought

##Role:
You are a multilabel classifier predicting the perceived emotion(s) of the author, knowing that text comes from twitter.
Label the text snippet with: joy (1) or no joy (0), sadness (1) or no sadness (0), anger (1) or no anger (0), surprise (1) or no surprise (0), fear (1) or no fear (0).

##Pointers:
- Remember that you are predicting emotions of author, not the reader.
- Carefully consider each possible emotion(label).

##Constraints:
- Classify text snippet provided in ##Input section
- Output only labels and their corresponding scores in following format: {"Label":Score}.
- Scores can only be either 0 or 1
- Use same format as provided by examples

##Examples:

Example #1

Input:
"i have major headache, just want to sleep all day, and the worst part when i look in the mirror my lips is swollen to like two times the size."

Reasoning:
Was author feeling anger? No.
Was author feeling fear? Yes.
Was author feeling joy? No.
Was author feeling sadness? Yes.
Was author feeling surprise? No.

Labels:
{ "anger":0,"fear":1,"joy":0,"sadness":1,"surprise":0 }

##Input
"/ o So today I went in for a new exam with Dr. Polvi today, I had to file new paperwork for the automobile accident case which is being done differently then the scoliosis stuff.So he comes in and starts talking about insurance stuff and how this look bad since I was getting treatment on my neck and stuff already blah blah."

Figure 4: Example structured (using formatting and additional isntructions), few-shot, chain-of-thought prompt. Only one example is included for clarity and brevity.