UIMP-Aaman at SemEval-2025 Task11: Detecting Intensity and Emotion in Social Media and News

Aisha Aman Parveen
Universidad Menendez Pelayo
100012653@alumnos.uimp.es

Abstract

This paper presents our participation in SemEval task 11, which consists of emotion recognition in sentences written in multiple languages. We use in-context learning and fine-tuning methods to teach LLMs how to predict labels for Track A, Track B and Track C. The best results depends on track and language predicted.

1 Introduction

In the 21st century, emotions remain a fascinating and complex subject of study. From an evolutionary perspective, they have played a vital role in human survival, shaping decision-making, fostering social bonds, and influencing intelligence by bridging reason and instinct. Despite significant scientific progress, emotions remain an enigmatic phenomenon that is difficult to fully understand due to their subjective nature. This complexity makes their analysis particularly challenging. In this study, we focus on emotion recognition in text, specifically in sentences written in multiple languages, in the context of the SemEval task 11 (Muhammad et al., 2025b). The languages selected for the analysis are English, German, Spanish, and Brazilian Portuguese. Our approach begins by examining the range of emotions present in the text files for each language (Muhammad et al., 2025a). Then, we outline the specific tasks involved in each track, outline our analytical methods, and describe the experiments conducted. Finally, we present the results from our best-performing experiments. The study is structured around three tasks: Track A, Track B, and Track C. In the first track, we will experiment with different models in order to find the one with best performance, including the Llama 3 series of models (Dubey et al., 2024) and the Phi-4 model (Abdin et al., 2024), using the HuggingFace transformers library (Wolf et al., 2019). Then, for Track B and C, we will implement the best model for each of these tracks. We report results using macro F1 metric computed over the dev set.

2 Track A

In Track A, the emotions to identify are Anger, Fear, Joy, Sadness, Surprise, and Disgust. However, it is important to note that Disgust is not present in English or Spanish. Table 1 provides an overview of the training data distribution. Specifically, we observe that in English, the most common emotions are Fear, Sadness, and Surprise. However, in German, Anger and Disgust appear most frequently. A similar pattern emerges in Portuguese and Spanish, where Anger and Joy are the most predominant emotions in both languages. Overall, we find that Disgust is more prevalent in German than in Portuguese, while Fear is more strongly expressed by English speakers.

In our analysis, we will use two key techniques: in-context learning and fine-tuning. In the first technique (In-context learning), we will provide the model with a prompt made up of training examples and their corresponding labels. The weights of the model are not updated, instead the provided examples are used by the model to learn the general pattern of the problem. On the other hand, fine-tuning involves taking a pre-trained model, which has learned general language patterns from large datasets, and adapting it to perform well on more specialized tasks using specific, domain-related data. Together, these techniques allow the model to leverage prior knowledge while tailoring its capabilities to address the specific needs of our task, enhancing both accuracy and efficiency.

2.1 In-context learning

In-context learning prompting involves providing a model with specific input to guide its understanding and responses. This can include task-specific prompts, where the model is directly told what to

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 331–335

Table 1: Track A training data label distribution. Green color indicates the two most common emotions.

	Anger	Fear	Joy	Sadness	Surprise	Disgust
English	333	1157	674	878	839	
German	768	239	541	516	159	832
Portuguese	718	109	581	332	153	75
Spanish	492	317	642	309	421	

do (e.g., "Summarize this text"), few-shot learning prompts, where the model is given a few examples to learn from, or zero-shot prompts, where the model is given no examples and must rely on its pre-existing knowledge. Prompting helps the model perform tasks more effectively by tailoring its responses based on the context provided.

In our project, the prompt is designed to instruct the model to identify emotions from a given sentence. The prompt starts by clearly outlining the task as we can in Table 2. Then we will give few examples of sentences to be classified as input and the model has to give as a response the emotion perceived.

The key here is that the prompt asks the model to return only the selected emotions separated by commas and no other information. The data is provided as a CSV file with the text and emotion columns containing either 0s or 1s. We transform this into the prompt. Specifically, if a sentence has '1' in the "Joy" column, we add "Joy" to the corresponding prompt. This allowed us to directly label the emotions based on the data. For converting the LLM response into labels during inference, we carry out the reverse procedure. That is, if the LLM response is a sentence followed by the emotion 'Joy', we will set 1 to the Joy column.

We then experimented with different models to identify the one that provided the best performance. The models tested include various versions of the Llama model, ranging from 3.2-1B to 3.3-70B. We also varied the number of examples used in the prompt to determine the optimal configuration for performance (n=20, n=30, n=40, n=60). Table 4 reports the results obtained by the different systems. We start by evaluating n=20. The results show some interesting trends. For instance, Llama 70B models consistently provide better performance across languages, specially for English and Spanish, where the scores are higher across various configurations. For example, the Llama 70B model gives a score of 0.77 for Spanish and 0.71 for English. These results suggest that increasing the number of parameters in the model and adjusting the number of examples in the prompt can significantly improve performance. Overall, the combination of using a large model like Llama 70B and fine-tuning the number of examples in the prompt seems to yield the best results. Further testing with additional languages or more fine-tuning could potentially improve these outcomes even more.

Table 4 reports the effects of changing the number of examples N shown in the prompt. From N=20 to N=40, we observe that English and German improve, while Spanish and Portuguese show a slight decline. Based on this, we decided to test with N=30 for Spanish and Portuguese to see if their performance improves. Since English and German perform better with N=40, we keep them at N=40, while Spanish and Portuguese remain at N=20.

2.2 Fine-tuning

In this section we will talk about fine-tuning the Llama 3.2-1B model. We only fine-tuned this model because we lack the computational and financial resources to fine-tune a bigger one. We trained a different model for each language, using the Adam optimizer (Kingma and Ba, 2015), learning rate $1e-5$, batch size 4 with 4 gradient accumulation steps (effective batch size 16). Training lasts for 3 epochs, and we apply a maximum gradient norm of 0.3 and a linear learning rate warm-up for the first 10% steps of training. Table 5 compares the performance of fine-tuning versus in-context learning for Track A across different languages. For English, in-context learning slightly outperforms fine-tuning, though the difference is small. In contrast, for the rest of languages, fine-tuning significantly outperforms in-context learning. Nevertheless, the Llama 3.3-70B is still better than our fine-tuned 1B model, so we selected the 3.3-70B model for the final submission.

Table 2: Prompt format used for LLM inference, Track A. This prompt includes 2 in-context examples for the Spanish task.

Instruction	Identifying emotions such as anger, fear, joy, sadness, surprise, and disgust, based on the sentence provided. Please only return the select emotions separated by a comma, and nothing else.
Input	2018 y lo sigo escuchando Like si tú también, te amo ...
Emotions	Joy
Input	No les aguas. Caso a los pen...sativos de los haters o como se escriba
Emotions	Anger, Disgust

Table 3: Track A results using different LLMs.

	Phi-4	Llama 3.1-8B	Llama 3.2-1B	Llama 3.2-3B	Llama 3.3-70B
English	0.63	0.59	0.43	0.45	0.71
German	0.53	0.47	0.26	0.36	0.59
Portuguese	0.68	0.68	0.36	0.55	0.77
Spanish	0.39	0.38	0.15	0.25	0.54

Table 4: Track A in-context learning results based on the number of examples shown in the prompt (n)

	n= 20	n= 40	n= 30	n= 60
English	0.71	0.73	-	0.73
German	0.59	0.63	-	0.60
Portuguese	0.77	0.76	0.77	-
Spanish	0.54	0.52	0.51	-

Table 5: Track A fine-tuning versus in-context learning results of Llama 3.2-1B.

	Fine-tuning	In-context learning
English	0.41	0.43
German	0.33	0.26
Portuguese	0.40	0.36
Spanish	0.24	0.15

3 Track B

Track B differs from Track A in that we also define the intensity of emotions across all languages. The intensity levels are as follows: 0 indicates no emotion, 1 represents a low degree of emotion, 2 indicates a moderate degree of emotion, and 3 corresponds to a high degree of emotion. Table 6 reports the full statistics for the training data. Upon examining the data for all languages, we find that high intensity emotions are the least frequent in the training set for the languages we analyzed. This suggests that the models have less data to learn from, making it more challenging to generate accurate predictions. Specifically, English has the most instances of high intensity emotions, while Portuguese has the fewest.

The prompt used for this second task is very similar to Track A, but with the addition of the emotion intensity. Table 7 shows the specific prompt. Based on results from track A, we will use the model Llama 3.3- 70B as it is the one with best performance. Results are shown in Table 8. The results obtained with the Llama 70B model show superior performance in English, with a score of 0.89, which suggests that the model is optimized or has more experience in this language. In German, the score is 0.67, indicating decent performance, although lower than in English. In Spanish and Portuguese, the model obtained similar scores of 0.65 and 0.66, respectively, suggesting that it faces more difficulties in these languages compared to English and German. It should be noted that the datasets used for Spanish and Portuguese are smaller (20 examples), which could have influenced performance. Overall, the model appears to perform better in languages with greater representation or training, such as English, and could benefit from further tuning to improve its performance in other languages.

4 Track C

Track C involves predicting the perceived emotion labels for a new text instance in a target language, using a labeled training set from one of the languages in Track A. The prompt used is identical to the one in Track A, and for all languages, 40 examples are provided in the prompt. The results are reported on Figure 1. We have decided to approach this task finding the best combination for each language in test. For instance, to predict English, we will train with German, Spanish and Portuguese dataset. The results obtained are 0.65 with German training, 0.69 with Spanish and 0.89 with Portuguese, which is the one that outperforms the other

Table 6: Track B: Emotion degree distribution across the different language pairs.

English	Anger	Fear	Joy	Sadness	Surprise
0: no emotion	2435	1157	2094	1890	1929
1: low degree emotion	207	857	449	505	588
2: moderate degree emotion	88	546	161	248	215
3: high degree emotion	38	208	64	125	36

Spanish	Anger	Fear	Joy	Sadness	Surprise
0: no emotion	1515	1679	1355	1689	1577
1: low degree emotion	518	242	538	228	333
2: moderate degree emotion	87	38	68	50	44
3: high degree emotion	46	37	35	29	42

German	Anger	Fear	Joy	Sadness	Surprise	Disgust
0: no emotion	1813	2349	2040	2068	2430	1745
1: low degree emotion	513	212	374	398	150	654
2: moderate degree emotion	262	36	172	118	23	185
3: high degree emotion	15	6	17	19	0	19

Portuguese	Anger	Fear	Joy	Sadness	Surprise	Disgust
0: no emotion	1508	2117	1645	1904	2073	2151
1: low degree emotion	459	81	286	209	116	66
2: moderate degree emotion	234	25	275	98	30	7
3: high degree emotion	25	3	20	15	7	2

Table 7: Prompt format used for LLM inference, Track B. This prompt includes 2 in-context examples for the Spanish task.

Instruction	I want you to identify which emotions and intensity would a person feel when reading a sentence. The list of possible emotions is: Anger, Fear, Joy, Sadness, Surprise. The intensity of emotions are: no emotion, low degree, moderate degree, high degree. Please only return the select emotions separated by a comma, and nothing else.
Input	2018 y lo sigo escuchando Like si tú también, te amo ...
Emotions	Joy high degree
Input	No les aguas. Caso a los pen...sativos de los haters o como se escriba
Emotions	Disgust moderate degree

Table 8: Track B results.

	Llama 3.3-70B
English	0.89
German	0.67
Portuguese	0.65
Spanish	0.66

languages. Moreover, Spanish best prediction is made with Portuguese training with a performance of 073. In contrast, for German, Spanish is the one with higher F1-score with 0.67. Finally, Portuguese is the best training predictor for Spanish. It is interesting to notice that English and Spanish are the languages with highest scores.

5 Final rankings

The final results of the competition are reported on Table 9. It can be observed how our approach our approach achieves good results in the ranking and is able to outperform the baseline for all tracks and languages, except for German Track A and Spanish Track B. The proposed LLM solution is powerful and convenient because it does not require training.

6 Conclusions

We have presented our participation at the SemEval task 11, which consists on emotion recognition in text, specifically in sentences written in multiple languages. In our approach we have implemented in- context learning and fine-tuning methods to

Table 9: Performance metrics for different tracks and languages. We report our submissions F1 score (**F1**), our ranking out of all participants in the Track (**Rank**) and the F1 of the proposed baseline by the organizers (**Baseline F1**).

Lang	Track A			Track B			Track C		
	F1	**Rank**	**Baseline F1**	**F1**	**Rank**	**Baseline F1**	**F1**	**Rank**	**Baseline F1**
deu	0.62	22/51	0.64	0.62	13/28	0.56	0.59	8/17	0.47
eng	0.72	48/98	0.71	0.73	22/44	0.64	0.66	7/18	0.38
ptbr	0.52	24/43	0.43	0.50	18/26	0.30	0.49	6/15	0.42
esp	0.77	24/48	0.77	0.70	18/30	0.73	0.69	9/16	0.57

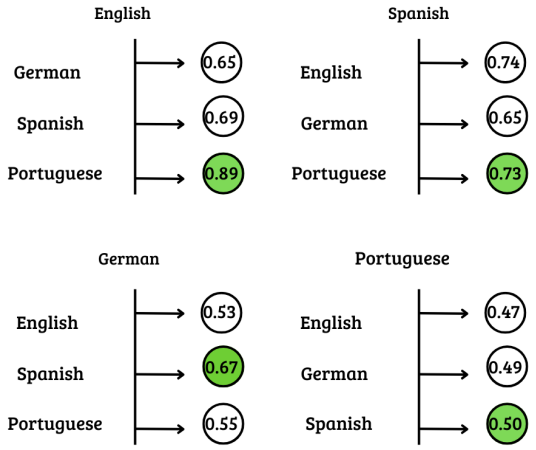

Figure 1: Track C results based on the training dataset. For each evaluation language, we report the performance depending on which of the other 3 languages is selected as training set.

teach LLM how to predict labels for Track A, Track B and Track C. We explore different Llama models in Track A until we found the one that outperforms and implement it on Track B and C. For future work, we would like to explore large LLM Fine Tuning with LoRa so we are able to improve results.

References

Marah Abdin et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905v1*.

Abhimanyu Dubey et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783v1*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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. *arXiv preprint arXiv:2502.11926v1*.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771v4*.