Language Model Prompting and Fine-tuning for Early Modern Spanish Corpus Text-based Emotion Detection

Liudmila Babakova[1], Ilia Stepin[2,*]

[1]Moscow State Pedagogical University, Malaya Pirogovaya St. 1-1, 119435 Moscow, Russian Federation
[2]Departamento de Ingeniería Informática, Escuela Politécnica Superior, Universidad Autónoma de Madríd, Calle Francisco Tomás y Valiente 11, 28049 Madrid, Spain

Abstract

Text-based emotion detection (TBED) is one of essential natural language processing problems. Such a well-known challenge as a low degree of data availability (as in the case of early modern Spanish) as well as the underexplored potential of the newly emerging large language models (LLMs) present unique opportunities for researchers to advance the state-of-the-art solutions applied to TBED. In this work, we propose several techniques for (large) language model prompting and fine-tuning, as we participate in the shared task "Historical Text-Based Emotion Detection in Early Modern Spanish Correspondence". The results show that the tested LLMs require further elaboration on prompt and fine-tuning design to be successfully applied to the TBED task.

Keywords

text-based emotion detection, natural language processing, digital humanities,

1. Introduction

Text-based emotion detection (TBED) is among the most widely addressed natural language processing-related problems. While language models (including newly emerging large language models (LLMs)) appear a promising tool for tackling this challenge, their potential remains underexplored for this task. This particularly applies to the cases where poor language resources are available, such as historical text emotion mining.

This work presents the results of an attempt to resolve the text-based emotion detection problem within the framework of the shared task "Historical Text-Based Emotion Detection in Early Modern Spanish Correspondence" (HISEMOTIONS) [1] organized as part of the Iberian Languages Evaluation Forum (IberLEF 2026) [2]. This shared task addresses semantic and diachronic challenges that the Spanish language migh pose for conveying emotion. Regarding this work, we prompt and fine-tune several LLMs using the original pieces of early modern Spanish text as well as their translations into English. In addition, we perform encoder tuning applied to several BERT-based language models. The corresponding multilabel dataset's sentences are mapped to a set of emotions (classes) that distinguishes sadness, joy, surprise, anger, fear, and hope. If the given sentence is not associated with any of the aforementioned classes, it is assigned the label "none".

The remainder of this manuscript is structured as follows. Section 2 describes the methodology applied throughout the experiments. Section 3 reports the corresponding experimental results. Section 4 concludes the paper stating the main achievements and limitations of the present work.

2. Methodology

Four groups of experiments have been carried out. First, zero-shot prompting experiments over the baseline model were carried out (see Section 2.1 for details). On a similar note, (Spanish-to-English) translation-based prompting was performed to estimate the potential of multilingual LLMs for TBED

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
babakovalyuda@gmail.com (L. Babakova); ilia.stepin@uam.es (I. Stepin)
0000-0002-4508-7555 (I. Stepin)

(see Section 2.2 for details). Then, one LLM was fine-tuned over a limited range of hyperparameters (see Section 2.3 for details). Subsequently, three encoders were fine-tuned over a wider range of hyperparameters to evaluate the impact on the corresponding models' performance (see Section 2.4 for details). Finally, the best performing model was selected to be tested using the test set data (see Section 2.5 for details).

2.1. Zero-shot prompting

First, three zero-shot prompting experiments were conducted. Those included the following three baseline LLMs: Qwen 2.5 [3][1], Qwen 3 [4][2] and DeepSeek 3.2 [5][3]. Importantly, these open-source models are directly available via API[4] and were exposed to Spanish when pretraining. Each of them was prompted with a query corresponding to the given emotion classification task. The following prompt was used to classify the emotion (possibly) encapsulated in text: *"You are a model for detecting emotions in text. For each text, determine the presence of the following emotions: joy, sadness, fear, anger, surprise, hope. A text may contain one emotion, multiple emotions, or no emotions. Return only the detected emotion names separated by commas. Examples: joy, sadness, fear, anger, surprise, hope. If no emotions are present, return: none. Do not add explanations. For single-text requests, return exactly one line"*.

2.2. Translation-based prompting

In the second round of experiments, the original texts were translated into English using GPT-5.2. This state-of-the-art model was selected due to its easiness to use via API. To do so, the following prompt was used: *"Translate each text into English. Preserve meaning, names, places, dates, and tone. Return only a valid JSON array of strings in the same order as the input array. Do not add markdown or explanations"*. The emotion classification task was then addressed using the models indicated in Section 2.1.

2.3. LLM fine-tuning

In the third round of experiments, Llama 3.1 [6][5] was fine-tuned to assess the impact of tuning against the previously prompted models (see Sections 2.1 and 2.2 for details), among others. This model was selected for fine-tuning due to hardware limitations. Two learning rates were used (1×10^{-4} and 1×10^{-5}) in the corresponding experiments. The quantized low-rank adaptation technique (QLoRA) [7] was used when fine-tuning the baseline model in both experiments.

2.4. Encoder tuning

Next, three encoders were fine-tuned additionally. Namely, those were (1) the case-sensitive multilingual BERT [8][6], (2) the BERT-based model trained on a Spanish corpus BETO [9][7] and (3) the multilingual BERT-based encoder mmBERT [10][8]. Encoders are cheap to both train and infer. While the downstream usage is limited to the data exposed during training, it remains a classic approach to classification problems.

Two types of BETO encoders were used in the corresponding experiments: the case-sensitive and the non-case-sensitive ones. All the three encoders were fine-tuned using three learning rates: 1×10^{-4}, 1×10^{-5} and 1×10^{-6}. In addition, BERT was fine-tuned using both weighted and unweighted binary cross-entropy loss.

[1] huggingface.co/Qwen/Qwen2.5-72B-Instruct
[2] Qwen/Qwen3-Max-Thinking
[3] huggingface.co/deepseek-ai/DeepSeek-V3.2
[4] deepinfra.com
[5] huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[6] huggingface.co/google-bert/bert-base-multilingual-cased
[7] huggingface.co/dccuchile/bert-base-spanish-wwm-cased
[8] huggingface.co/jhu-clsp/mmBERT-small

Table 1
Zero-shot prompting experimental results (all of them sorted by macro F1-score)

Model	Precision (macro)	Recall (macro)	F1-score (macro)
Qwen-2.5	0.516	0.639	0.568
Qwen-3	0.533	0.658	0.567
DeepSeek-3.2	0.678	0.551	0.542

Table 2
Zero-shot prompting experimental results (F1-scores by class)

Model	Anger	Fear	Joy	Sadness	Surprise	Hope	None
Qwen-2.5	0.625	0.467	0.523	0.622	0.429	0.451	0.860
Qwen-3	0.680	0.516	0.529	0.579	0.364	0.486	0.812
DeepSeek-3.2	0.667	0.500	0.492	0.623	0.250	0.397	0.867

Table 3
Translation-based experimental results (all of them sorted by macro F1-score)

Model	Precision (macro)	Recall (macro)	F1-score (macro)
Qwen-3	0.598	0.648	0.611
Qwen-2.5	0.674	0.568	0.598
DeepSeek-3.2	0.637	0.611	0.582

Table 4
Translation-based experimental results (F1-scores by class)

Model	Anger	Fear	Joy	Sadness	Surprise	Hope	None
Qwen-2.5	0.694	0.469	0.510	0.604	0.600	0.457	0.851
Qwen-3	0.763	0.517	0.610	0.641	0.333	0.544	0.866
DeepSeek-3.2	0.692	0.482	0.465	0.600	0.444	0.554	0.840

2.5. Final model selection

Finally, the weighted majority voting strategy was applied to select the best performing model among those used in Sections 2.1 - 2.4. This strategy was applied within the framework of another shared task on TBED [11]. Hence, inspired by a work from that shared task [12], we created several ensemble models based on the macro F1-scores of the n best performing models used to estimate which ensemble showed best performance. Ensembles of the top-3, top-5, top-7, and top-10 models with the best macro F1-scores were selected.

3. Experimental results

This section reports the results obtained during each experimental phase. Table 1 demonstrates the overall results (macro precision, recall, and F1-scores) achieved using the models mentioned in Section 2.1. It turns out that both Qwen models slightly outperform the DeepSeek-3.2 model. Table 2 reports the F1-scores computed for each of these models for each class.

Table 3 shows the results obtained for the previously reported models when the input text was first translated into English. Again, Qwen-3 [4] performs slightly better than the other two models tested. Table 4 reports the class-based F1-scores obtained for each model.

Table 5 reports the overall LLM fine-tuning results. All the computed evaluation metrics appear to be of lower values than those computed for the directly prompted LLMs or those translation-based.

Table 5
LLM fine-tuning experimental results (all of them sorted by macro F1-score; lr stands for learning rate)

Model	Precision (macro)	Recall (macro)	F1-score (macro)
Llama-3.1, lr=1e-05	0.515	0.581	0.409
Llama-3.1, lr=1e-04	0.488	0.481	0.367

Table 6
LLM fine-tuning experimental results (F1-scores by class; lr stands for learning rate)

Model	Anger	Fear	Joy	Sadness	Surprise	Hope	None
Llama-3.1, lr=1e-05	0.448	0.279	0.333	0.370	0.222	0.551	0.658
Llama-3.1, lr=1e-04	0.038	0.443	0.507	0.384	0.000	0.380	0.819

Table 7
Encoder tuning experimental results (all of them sorted by macro F1-score; lr stands for learning rate)

Model	Precision (macro)	Recall (macro)	F1-score (macro)
BETO (lr=1e-05, cased, weighted)	0.424	0.553	0.449
BETO (lr=1e-04, cased, weighted)	0.413	0.576	0.440
BERT (lr=1e-05, cased, weighted)	0.352	0.537	0.411
BETO (lr=1e-05, cased)	0.443	0.450	0.386
BETO (lr=1e-05, uncased)	0.447	0.440	0.356
BERT (lr=1e-05, cased)	0.323	0.421	0.328
mmBERT (lr=1e-05)	0.486	0.339	0.291
mmBERT (lr=1e-04)	0.370	0.321	0.249
mmBERT (lr=1e-06)	0.304	0.381	0.284
BETO (lr=1e-04, cased)	0.320	0.335	0.278
BETO (lr=1e-06, cased)	0.286	0.385	0.275
BETO (lr=1e-04, uncased)	0.184	0.328	0.226
BERT (lr=1e-06, cased, weighted)	0.121	0.615	0.196
BERT (lr=1e-06, cased)	0.135	0.512	0.191
BETO (lr=1e-06, cased, weighted)	0.118	0.456	0.180
BETO (lr=1e-06, uncased)	0.085	0.714	0.149
BERT (lr=1e-04, weighted)	0.039	0.429	0.071
BERT (lr=1e-04)	0.031	0.286	0.055

Remarkably, the F1-scores of the fine-tune model (Llama-3.1) with both learning rates used do not achieve the 50% level. Table 6 presents them distinguishing between the dataset classes.

Table 7 presents the experimental results for LLMs where encoder tuning was applied. In general, the best performing LLM is the Spanish version of BERT [9]. Remarkably, its both cased and uncased versions appear four times among the top-5 models showing best performance. Table 8 reports the each encoder's F1-scores for each class.

Finally, Table 9 shows the evaluation metric scores computed for the ensemble models made up of the n-best performing models. Table 10 distinguishes the corresponding results among all the dataset classes. The top-3 models are all the LLMs used in the translation-based prompting experiments (see Table 3 for details). The top-5 models are all those making part of top-3 as well as Qwen-2.5 and Qwen-3 used in zero-shot prompting experiments. The top-7 models are the top-5 models as well as DeepSeek-3.2 used in zero-shot prompting experiments and BETO (learning rate = 1×10^{-5}, cased, weighted). Finally, the top-10 models are all those from top-7 as well as BETO (learning rate = 1×10^{-4}, cased, weighted), BERT (learning rate = 1×10^{-5}, cased, weighted), and the fine-tuned Llama-3.1 (learning rate = 1×10^{-5}).

Being the best performing model on the development set, the top-7 ensemble was submitted to be

Table 8
Encoder tuning experimental results (F1-scores by class; lr stands for learning rate)

Model	Anger	Fear	Joy	Sadness	Surprise	Hope	None
BETO (lr=1e-05, cased, weighted)	0.382	0.315	0.500	0.467	0.231	0.393	0.854
BETO (lr=1e-04, cased, weighted)	0.367	0.311	0.511	0.524	0.107	0.404	0.857
BERT (lr=1e-05, cased, weighted)	0.362	0.332	0.423	0.476	0.000	0.441	0.844
BETO (lr=1e-05, cased)	0.250	0.379	0.510	0.459	0.000	0.258	0.847
BETO (lr=1e-05, uncased)	0.172	0.347	0.491	0.460	0.000	0.167	0.856
BERT (lr=1e-05, cased)	0.263	0.297	0.209	0.435	0.000	0.260	0.834
mmBERT (lr=1e-05)	0.038	0.289	0.400	0.419	0.000	0.067	0.823
mmBERT (lr=1e-04)	0.000	0.328	0.229	0.549	0.000	0.065	0.822
mmBERT (lr=1e-06)	0.000	0.325	0.261	0.506	0.000	0.086	0.811
BETO (lr=1e-04, cased)	0.000	0.256	0.435	0.383	0.000	0.044	0.829
BETO (lr=1e-06, cased)	0.000	0.281	0.232	0.455	0.000	0.141	0.819
BETO (lr=1e-04, uncased)	0.000	0.271	0.000	0.485	0.000	0.000	0.823
BERT (lr=1e-06, cased, weighted)	0.211	0.187	0.211	0.384	0.095	0.285	0.000
BERT (lr=1e-06, cased)	0.252	0.252	0.107	0.441	0.035	0.253	0.000
BETO (lr=1e-06, cased, weighted)	0.169	0.282	0.124	0.400	0.000	0.286	0.000
BETO (lr=1e-06, uncased)	0.185	0.145	0.107	0.273	0.047	0.285	0.000
BERT (lr=1e-04, weighted)	0.000	0.145	0.107	0.241	0.000	0.000	0.000
BERT (lr=1e-04)	0.000	0.145	0.000	0.241	0.000	0.000	0.000

Table 9
Final model selection experimental results

Model	Precision (macro)	Recall (macro)	F1-score (macro)
Top-3	0.681	0.616	0.615
Top-5	0.683	0.648	0.630
Top-7	0.702	0.638	**0.631**
Top-10	0.702	0.638	0.628

Table 10
Final model selection experimental results (F1-scores by class).

Model	Anger	Fear	Joy	Sadness	Surprise	Hope	None
Top-3	0.740	0.500	0.581	0.641	0.444	0.542	0.859
Top-5	0.763	0.571	0.563	0.662	0.444	0.536	0.867
Top-7	0.761	0.574	0.567	0.684	0.444	0.517	0.868
Top-10	0.727	0.539	0.576	0.688	0.444	0.546	0.872

tested using the test data provided by the shared task organizers. The macro F1-score registered on the test set is of 0.57.

4. Conclusion

This work underlines the complexity of the TBED task for early modern Spanish corpora. Indeed, the macro F1 scores the directly prompted LLMs barely exceed 50%. In case of the translation-based prompts, the corresponding macro F1 scores increase slightly. Remarkably, the macro F1 scores of both the fine-tuned LLMs and BERT-based language models do not reach 50%.

It is therefore important to acknowledge a number of limitations. First, further data normalization techniques appear needed, as F1-scores for specific classes (irrespective of the models) are oftentimes equal to 0. Second, only a limited number of LLM models were used in the experiments. Third, a wider range of hyperparameters should be used, as part of future work, to fine-tune the corresponding models.

Noteworthy, our team ranked third in the given shared task. It is also worth noting that the source code is made publicly available [9].

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] A. Sarymsakova, P. Martín-Rodilla, E. Martínez-Cámara, L. A. Ureña-López, Overview of hisemotions at iberlef 2026: Historical text-based emotion detection in early modern spanish correspondence, Procesamiento del lenguaje natural 77 (2026).

[2] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS. org, 2026.

[3] Q. Team, Qwen2.5: A party of foundation models, 2024. URL: https://qwenlm.github.io/blog/qwen2.5/.

[4] Q. Team, Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025). URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

[5] DeepSeek-AI, A. Liu, A. Mei, B. Lin, B. Xue, B. Wang, B. Xu, B. Wu, B. Zhang, C. Lin, C. Dong, et al., DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models, 2025. URL: https://arxiv.org/abs/2512.02556v1. arXiv:2512.02556.

[6] A. Grattafiori, et al., The Llama 3 Herd of Models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[7] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLORA: efficient finetuning of quantized LLMs, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[8] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[9] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[10] M. Marone, O. Weller, W. Fleshman, E. Yang, D. Lawrie, B. V. Durme, mmbert: A modern multilingual encoder with annealed language learning, 2025. URL: https://arxiv.org/abs/2509.06888. arXiv:2509.06888.

[11] S. H. Muhammad, N. Ousidhoum, I. Abdulmumin, S. M. Yimam, J. P. Wahle, T. L. Ruas, M. Beloucif, C. De Kock, T. D. Belay, I. S. Ahmad, et al., Semeval-2025 task 11: Bridging the gap in text-based emotion detection, in: Proceedings of the 19th international workshop on semantic evaluation (SemEval-2025), 2025, pp. 2558–2569.

[12] Z. Ruan, R. You, K. Yang, J. Lin, W. Dai, M. Zhou, M. Jin, X. Mei, PAI at SemEval-2025 Task 11: A Large Language Model Ensemble Strategy for Text-Based Emotion Detection, in: Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), 2025, pp. 1136–1142.

[9] https://github.com/lyuda-web/HISEMOTIONS_2026_explainators