UMUTeam at HISEMOTIONS 2026@IberLEF 2026: Few-Shot Prompting with Large Language Models for Emotion Detection in Early Modern Spanish Letters

Tomás Bernal-Beltrán[1], Ronghao Pan[1], Jorge Gómez-Navalón[1], José Antonio García-Díaz[1], Ghassan Beydoun[2] and Rafael Valencia-García[1,*]

[1]Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain
[2]School of Information, Systems and Modelling, Faculty of Engineering and IT, University of Technology Sydney, Sydney, Australia

Abstract

Emotion detection in historical texts remains a challenging problem due to semantic shift, orthographic variation and the scarcity of annotated resources for historical language varieties. This paper describes the UMUTeam participation in the HISEMOTIONS 2026 shared task on Text-Based Emotion Detection in Early Modern Spanish epistolary texts. The task consists of identifying the presence of six emotion categories in historical letter fragments, formulated as a multi-label classification problem. Our approach relies on Few-Shot Learning (FSL) with the Qwen3.5-9B model, using instruction-based prompting and avoiding task-specific fine-tuning. The proposed pipeline combines task instructions, five randomly selected few-shot examples from the training set and structured JSON generation to obtain the final emotion predictions. On the official test set, our system achieved a macro-F1 score of 0.47, ranking 4th out of 10 participating teams. The results show that prompting-based approaches using general-purpose LLMs can provide competitive performance for historical emotion detection despite the strong semantic and linguistic differences between Early Modern and contemporary Spanish.

Keywords

Emotion Detection, Historical NLP, Early Modern Spanish, Few-Shot Learning, Prompt-Tuning, Large Language Models, Natural Language Processing

1. Introduction

Emotion detection in text has become a central task in Natural Language Processing (NLP), with applications ranging from sentiment analysis and opinion mining [1] to mental health assessment [2] and social media monitoring [3]. These applications rely on the ability of computational models to capture affective signals from textual data, which are often subtle, context-dependent and influenced by linguistic conventions. While most existing research has focused on contemporary language and modern textual genres, significantly less attention has been devoted to historical language varieties, where linguistic, semantic and cultural differences introduce additional challenges for automatic analysis [4]. In particular, Early Modern Spanish texts present a unique scenario in which affective expression is shaped by diachronic variation, non-standardized orthography and distinct rhetorical conventions, making the direct application of modern NLP methods non-trivial. As a result, approaches that perform well on contemporary Spanish may fail to generalize when applied to historical corpora, where both the surface form and the underlying semantic representations differ substantially.

Automatically identifying emotions in historical texts is inherently difficult. One of the main challenges is semantic shift, as the affective meaning of words may differ substantially from their modern equivalents [4]. This phenomenon can lead to systematic misinterpretations when models trained on contemporary data are directly applied to historical texts. Recent studies have highlighted the

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
tomas.bernalb@um.es (T. Bernal-Beltrán); ronghao.pan@um.es (R. Pan); jorge.gomezn@um.es (J. Gómez-Navalón); joseantonio.garcia8@um.es (J. A. García-Díaz); Ghassan.Beydoun@uts.edu.au (G. Beydoun); valencia@um.es (R. Valencia-García)
0009-0006-6971-1435 (T. Bernal-Beltrán); 0009-0008-7317-7145 (R. Pan); 0009-0005-7185-6054 (J. Gómez-Navalón); 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0001-8087-5445 (G. Beydoun); 0000-0003-2457-1791 (R. Valencia-García)

importance of explicitly modeling lexical semantic change to improve robustness in diachronic settings [5]. In the case of Spanish, this issue is further exacerbated by historical variation in vocabulary and usage across time [6]. In addition, emotional content in historical correspondence is often expressed implicitly or through culturally specific conventions, rather than through explicit affective markers. This makes the detection of emotions particularly challenging, as models must rely on contextual and pragmatic cues rather than surface-level lexical signals. These factors highlight the need for approaches that are robust to diachronic variation and capable of capturing subtle and context-dependent emotional information.

The HISEMOTIONS 2026 [7] shared task, which is part of IberLEF 2026 [8], addresses these challenges by proposing a Text-Based Emotion Detection (TBED) task on Early Modern Spanish epistolary texts from the 16th and 17th centuries. The goal is to identify the presence of multiple emotions in historical letter fragments, framing the problem as a multi-label classification task. Unlike previous IberLEF shared tasks, which focused on contemporary Spanish and modern domains, this task explicitly targets historical language varieties, aiming to evaluate the generalization capabilities of current models and to analyze the impact of semantic change on emotion detection.

Within Digital Humanities and Computational Literary Studies, emotion analysis has been widely explored [9], particularly for the study of affective patterns in literary and historical corpora. Prior work has examined the emotional structure of novels [10] as well as the use of sentiment analysis techniques in historical political texts [11]. However, relatively few studies have focused on historical correspondence, despite its relevance as a primary source for understanding affective communication in past societies. Recent work has started to address this gap by exploring emotion analysis in epistolary collections [12].

In this paper, we present the system developed by UMUTeam for the HISEMOTIONS 2026 shared task. Our approach is based on Few-Shot Learning (FSL) using a Large Language Model (LLM), specifically Qwen3.5-9B [13], prompted in English with a small set of examples extracted from the training set. We selected Qwen3.5-9B because of its strong instruction-following capabilities and competitive performance in multilingual and few-shot settings, while maintaining a moderate model size. FSL prompting has shown strong performance across a wide range of NLP tasks, enabling models to generalize to new domains with limited supervision [14]. More recent work has demonstrated that carefully designed prompts can further improve performance by guiding the reasoning process of LLMs [15]. In our approach, the model is instructed to perform multi-label emotion classification by generating a structured JSON output containing a binary decision for each emotion category. This output format allows for straightforward parsing and evaluation, ensuring consistency between model predictions and the required submission format. The prompt design explicitly constrains the model to produce only the desired structure, reducing variability in the generated responses. This strategy leverages the generalization capabilities of LLMs while avoiding the need for task-specific fine-tuning, which can be particularly challenging in low-resource and domain-shifted scenarios such as historical text analysis.

On the official test set, our system achieved a macro-F1 score of 0.47, ranking 4th out of 10 participating teams. The results show that FSL prompting with a general-purpose LLM can provide competitive performance even in a challenging historical setting characterized by semantic shift and limited annotated data. Our analyses suggest that the system performs more reliably when emotions are expressed explicitly, while implicit or context-dependent affective cues remain considerably more difficult to identify. These findings highlight both the potential and the limitations of instruction-based LLM prompting for emotion detection in historical correspondence.

The remainder of this paper is organized as follows. Section 2 reviews related work on emotion detection, historical text analysis and prompting strategies for LLMs. Section 3 describes the proposed methodology, including the prompting strategy and FSL configuration. Section 4 outlines the dataset and details the experimental setup and evaluation procedure. Section 5 presents the experimental results obtained on the official test set. Section 6 provides a discussion of the results, including error analysis and limitations of the proposed approach. Finally, Section 7 concludes the paper and outlines potential directions for future work.

2. Background Information

Text-based emotion detection has evolved from lexicon-oriented methods to supervised and transformer-based approaches. Early systems were largely built around affective lexical resources such as WordNet-Affect or NRC Emotion Lexicon, relying on word-level associations to infer emotional content [16]. These approaches provided simple and interpretable baselines, but tended to be limited when emotion depended strongly on context, composition or domain-specific usage. More recent work has instead consolidated annotated benchmarks and contextual modeling as the standard evaluation setting for emotion analysis in NLP [9]. The availability of labeled datasets has enabled the adoption of supervised and pre-trained transformer-based models, which capture contextual dependencies more effectively than lexicon-based methods [17]. Additionally, large-scale emotion datasets have further reinforced this paradigm by enabling fine-grained evaluation of emotion classification systems [18].

Recent benchmarks have also strengthened the view of emotion detection as a multi-label problem rather than a purely single-label one. Early shared tasks demonstrated the complexity of modeling affective signals in noisy and domain-specific data, particularly in social media contexts [19]. More recent datasets, such as GoEmotions [18] have expanded the granularity of emotion annotations, showing that fine-grained emotion categories introduce additional challenges in terms of label ambiguity and class imbalance. Furthermore, recent multilingual and low-resource benchmarks have highlighted that model performance varies substantially across languages, domains and annotation schemes, reinforcing the importance of domain adaptation strategies [20]. This is particularly relevant in the context of IberLEF tasks, where cross-domain and cross-lingual generalization is a recurring challenge.

Within Digital Humanities, emotion analysis has also been applied to literary and historical materials, but historical correspondence remains comparatively underexplored with respect to drama, fiction or political prose. Early work in this area has shown the potential of sentiment analysis techniques for historical texts, although often relying on adapted lexicons or simplified models [11]. More recent studies have explored emotion detection in specific historical corpora, highlighting the importance of contextual and cultural factors in affective expression [21]. In parallel, research on historical correspondence has begun to address the specific challenges of epistolary texts, where emotion is often expressed indirectly and shaped by social conventions [12]. This is particularly relevant for HISEMOTIONS, since epistolary discourse combines personal stance, conventional formulae and socially conditioned affective expression that are only weakly represented in modern benchmarks.

Historical NLP introduces difficulties that are only partially shared with contemporary NLP: non-standard orthography, unstable tokenization, sparse gold data, editorial or transcription noise, and diachronic semantic drift. These factors affect both the input representation and the reliability of downstream linguistic processing, often leading to a significant degradation in model performance when tools trained on modern language are directly applied to historical corpora. General work in the area repeatedly identifies orthographic instability as one of the main causes of performance degradation, since spelling variation directly impacts lexical coverage and hampers the effectiveness of standard tokenization and normalization pipelines [22].

For Spanish, these issues have been studied directly. Previous work has shown that adapting FreeLing to Old Spanish required extending the dictionary, customizing tokenization and retraining tagging components, precisely because graphemic variation affects lexical coverage and downstream linguistic analysis [23] . In parallel, the IMPACT-es project [24] released an open diachronic corpus of historical Spanish with lexical variants and lemma links, providing an important resource base for diachronic NLP in Spanish. These efforts highlight that even well-established NLP pipelines require substantial adaptation when applied to historical language varieties.

A closely related issue is lexical semantic change. Work on semantic change detection has helped formalize the problem and has provided evaluation frameworks for measuring temporal meaning shift rather than treating it as a purely anecdotal phenomenon [25]. Shared evaluation campaigns have further contributed to standardizing the task and enabling systematic comparison between approaches [26]. For emotion detection, this is especially important because affective meaning is often encoded in lexical items whose connotations do not remain stable over time. Historical normalization can facilitate

downstream processing by reducing surface variation [27], but it is not a complete solution when the historically appropriate interpretation diverges from the modern one.

LLMs have reopened the possibility of solving classification tasks through inference time adaptation rather than task specific fine-tuning. Previous work has shown that sufficiently large models can perform new tasks from instructions and a small number of demonstrations, without requiring parameter updates [14]. Subsequent work on In-Context Learning (ICL) further demonstrated that the effectiveness of this paradigm depends not only on the presence of examples, but also on how these examples expose the label space, the input distribution and the expected output structure [28]. This has led to a growing interest in prompt design as a central component of LLM-based systems.

For a workshop setting with limited historical supervision, this makes FSL prompting a pragmatic alternative. It reduces the engineering overhead associated with fine-tuning and avoids committing to a task-specific model trained on a relatively small and strongly shifted dataset. In addition, it allows rapid iteration over prompts without retraining. More broadly, prior work has shown that carefully designed prompts can significantly improve inference behaviour [14]. In particular, techniques such as structured prompting and explicit reasoning cues have been shown to enhance model reliability and consistency [15]. Recent work has further highlighted the practical value of enforcing structured output formats to ensure that model predictions can be reliably parsed and integrated into downstream pipelines [29].

This background directly motivates our final design choice for HISEMOTIONS 2026. Our system relies on FSL prompting with Qwen3.5-9B [13], frames the problem as multi-label emotion classification, and explicitly constrains the model to generate a structured JSON output containing binary decisions for each emotion category. This output is subsequently parsed to obtain the final predictions. In a setting defined by historical Spanish, semantic change and limited annotated data, this approach provides a simple, reproducible and low-overhead way of exploiting a strong general-purpose LLM, while enabling adaptation to the specific characteristics of the task through instruction-based prompting without requiring additional training.

3. System Overview

Figure 1 illustrates the overall architecture of the proposed system for the HISEMOTIONS 2026 shared task. The system is based on a FSL prompting strategy using Qwen3.5-9B. Given a historical Spanish letter fragment, the model is instructed to perform multi-label emotion classification by generating a structured JSON output containing binary decisions for each target emotion category.

The prompting pipeline is composed of three main elements: a set of task instructions, a collection of five few-shot demonstrations extracted from the training set, and the target historical letter fragment to be classified. Each few-shot demonstration consists of a historical text fragment paired with its corresponding emotion annotations encoded as a JSON object. The target letter fragment, which is distinct from the demonstrations, is appended after the few-shot examples using the same input format. These components are combined into a single prompt, which is subsequently processed by Qwen3.5-9B. The generated response follows a predefined JSON schema containing one binary value per emotion label, which is later parsed to obtain the final multi-label predictions.

The task is formulated as a multi-label text classification problem over six emotion categories: anger, fear, joy, sadness, surprise and hope. Unlike traditional supervised approaches based on task-specific fine-tuning, our method relies entirely on inference-time prompting, avoiding any parameter updates or additional training stages. This design choice is particularly suitable for historical NLP settings, where annotated data is limited and semantic shift may reduce the effectiveness of directly transferring models trained on contemporary language.

The prompt instructions explicitly constrain the model to generate only the desired JSON structure, reducing variability in the responses and facilitating reliable downstream parsing. Each few-shot example consists of a historical text fragment paired with its corresponding binary emotion annotations, exposing the expected input format, label space and output structure to the model prior to inference.

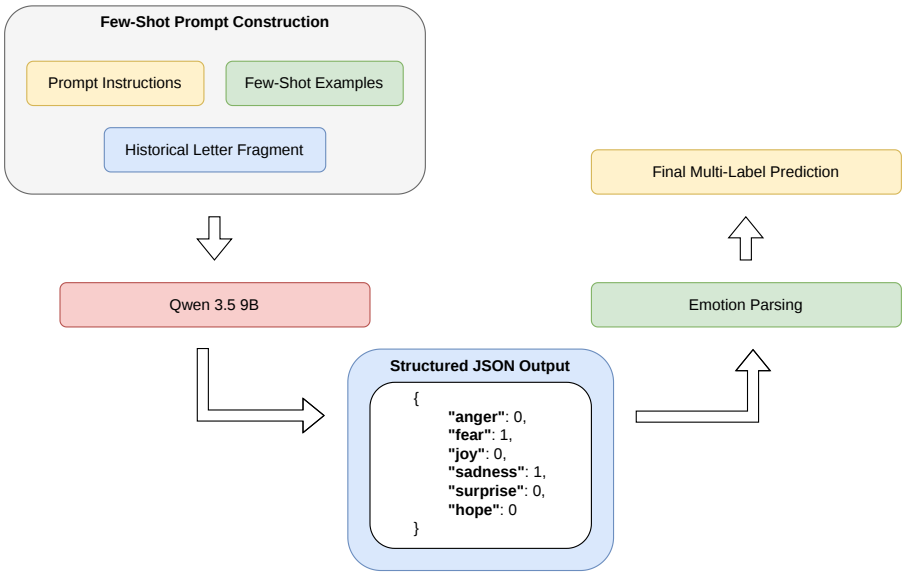

Figure 1: Overview of the proposed FSL pipeline.

The target text is appended after the few-shot demonstrations using the same structure employed in the examples. During inference, Qwen3.5-9B generates the predicted JSON response autoregressively. The generated content is subsequently processed through a lightweight parsing stage that extracts the binary predictions associated with each emotion category. If the response contains additional explanatory text or malformed structures, only the valid JSON fragment is retained for the final prediction extraction.

Overall, the proposed approach combines instruction-based prompting, few-shot demonstrations and structured generation into a simple and reproducible pipeline for emotion detection in Early Modern Spanish correspondence. By relying on a general-purpose LLM instead of task-specific fine-tuning, the system minimizes engineering overhead while remaining flexible enough to adapt to the linguistic variability and semantic ambiguity characteristic of historical texts.

We used the HuggingFace Transformers library to implement the system. Prior to inference, the historical letter fragments were minimally pre-processed by trimming redundant whitespace while preserving the original orthographic characteristics of the texts, since spelling variation constitutes an important feature of Early Modern Spanish. The resulting prompts were tokenized using the model tokenizer and processed autoregressively to generate the final structured predictions.

4. Experimental Setup

Our experiments were based exclusively on the training and development sets provided by the task organizers. Table 1 presents the number of examples and label distribution for each split.

As shown in Table 1, the dataset exhibits a strong class imbalance, particularly for emotions such as anger and surprise, whose positive instances represent less than 1% of the training data. Similarly, the development split also presents notable imbalance across emotion categories, although the label distribution differs considerably from the training data. For instance, anger and hope are substantially more represented in the development set, while sadness, despite remaining one of the most frequent categories, appears with a much lower proportion than in training. These differences between splits further increase the difficulty of the task, as models must generalize under both severe imbalance and distribution shift conditions. In contrast, sadness constitutes the dominant category in the training split, accounting for nearly half of all examples.

Given the strong class imbalance and the limited amount of annotated historical data available, we

Table 1
HISEMOTIONS 2026 dataset distribution. Each cell reports the number of positive examples and the percentage they represent within the corresponding split.

Emotion	Training	Development	Test
Total examples	2659	425	863
anger	12 (0.45%)	52 (12.24%)	– – – –
fear	303 (11.40%)	40 (9.41%)	– – – –
joy	129 (4.85%)	29 (6.82%)	– – – –
sadness	1180 (44.38%)	70 (16.47%)	– – – –
surprise	9 (0.34%)	7 (1.65%)	– – – –
hope	73 (2.75%)	85 (20.00%)	– – – –

opted for a prompting-based approach instead of task-specific fine-tuning. Since gold annotations for the test split were not publicly available, all prompt design decisions and methodological choices were based exclusively on the provided training and development data.

Our prompting strategy relied on five randomly selected few-shot examples extracted from the training split. Each example consisted of a historical letter fragment paired with its corresponding binary emotion annotations encoded as a JSON object. These demonstrations were incorporated into the prompt together with a set of task instructions and the target text to be classified, exposing the expected input structure, label space and output format to the model prior to inference.

The prompt used during inference is shown in Listing 1. In addition to the user prompt, the model was initialized with the following system instruction: *"You are an expert in emotion analysis of Early Modern Spanish letters."* This system-level instruction was designed to guide the model towards the specific linguistic and historical characteristics of the task domain.

Listing 1: Prompt used for emotion classification

```
For the following text, detect whether each emotion is present.

Emotions:
- anger
- fear
- joy
- sadness
- surprise
- hope

Return ONLY a JSON object with 0 or 1 for each emotion.
Do not explain your answer.

Few-shot examples:

{examples_text}

Now classify this text:

Text:
{text}

Output:
```

As shown in Listing 1, the instructions explicitly constrain the model to generate only a JSON object containing binary values for each emotion category. This structured format reduces variability in the

generated responses and facilitates reliable downstream parsing. The prompt itself was written in English, while the historical texts remained in their original Early Modern Spanish form.

Each few-shot example followed the same structure as the target inference instance, consisting of a historical text fragment paired with its corresponding JSON annotation. Listing 2 presents the template used to incorporate the demonstrations into the prompt.

Listing 2: Few-shot example formatting

```
Example {i}

Text:
{example_text}

Output:
{
  "anger": {value_for_anger_in_example},
  "fear": {value_for_fear_in_example},
  "joy": {value_for_joy_in_example},
  "sadness": {value_for_sadness_in_example},
  "surprise": {value_for_surprise_in_example},
  "hope": {value_for_hope_in_example}
}
```

Inference was performed using the HuggingFace Transformers library together with the Qwen3.5-9B model [13]. Text generation was carried out deterministically using greedy decoding (do_sample=False), ensuring that the model always selected the most probable token at each generation step. This strategy was adopted to reduce variability across generations and encourage stable, reproducible outputs, which is particularly important when enforcing a structured JSON format.

The maximum generation length was limited to 64 new tokens, which was sufficient to generate the complete JSON structure while preventing unnecessarily long or verbose responses. In addition, generation was performed with a beam size of 1, effectively disabling beam search and maintaining deterministic decoding behaviour throughout inference. Overall, this configuration was designed to prioritize output consistency and reliable parsing over generative diversity.

The generated responses were subsequently processed through a lightweight parsing stage designed to extract the first valid JSON fragment from the model output. Since LLM-generated responses may occasionally contain additional explanatory text or malformed structures despite the prompt constraints, the parsing procedure relied on regular-expression matching to identify the first JSON-like segment in the generated content. This fragment was then validated and decoded into the corresponding emotion predictions.

If the response did not contain a valid JSON structure, or if any emotion value could not be correctly interpreted as a binary label, a default prediction assigning value 0 to all emotion categories was returned. This conservative fallback strategy ensured that malformed generations did not interrupt the evaluation pipeline or produce inconsistent outputs. Finally, the extracted predictions were converted into binary labels corresponding to the six target emotion categories used in the shared task evaluation.

In practice, malformed generations were extremely rare. Thanks to the use of greedy decoding and explicit prompt constraints, the model consistently produced well-formed JSON outputs. During manual inspection of the generated predictions, we observed only a small number of cases containing additional explanatory text surrounding the JSON object. However, these cases were successfully handled by the parsing procedure, which extracts the first valid JSON fragment from the response. No invalid emotion labels were observed, and therefore the fallback strategy assigning zero values to all emotions was never triggered on the official test set.

5. Results

In this section, we present the results obtained by our final submission on the official test set. During the evaluation phase, we explored multiple approaches for the task, including supervised fine-tuning of encoder-based models, weighted and unweighted ensemble strategies and independent label-specific classifiers. However, the best overall performance was ultimately achieved using the proposed FSL prompting approach with Qwen3.5-9B.

To better understand the effectiveness of the proposed method, Table 2 summarizes the macro-F1 scores obtained by the different approaches evaluated. We first explored standard multi-label fine-tuning of several Spanish transformer-based encoder architectures, namely Rigoberta [30], BETO-cased and BETO-uncased [31], BERTIN [32], DistilBETO [33] and MR-BERT-es [34].

After identifying the best-performing individual models, we evaluated weighted and unweighted soft-voting ensembles. All possible model combinations were explored, and ensemble weights were selected through a grid search over the interval $[0, 2]$ with a step size of 0.25. The best unweighted ensemble corresponded to the combination of Rigoberta, MR-BERT and BETO-cased. The best weighted ensemble corresponded to the combination of Rigoberta, MR-BERT and BERTIN, with normalized weights of 0.50, 0.25 and 0.25, respectively.

Finally, we investigated a label-specific strategy in which six independent classifiers were trained, one per emotion category. Since this approach was explored after the previous experiments, only the best individual encoder was employed for all labels, namely Rigoberta. This configuration produced the lowest overall performance. In contrast, the proposed FSL prompting strategy with Qwen3.5-9B clearly outperformed all encoder-based alternatives.

Table 2
Comparison of the different approaches explored.

Approach	Macro-F1
Rigoberta	0.38
MR-BERT	0.36
BETO-cased	0.34
BERTIN	0.33
BETO-uncased	0.31
DistilBETO	0.30
Best unweighted ensemble (Rigoberta+MR-BERT+BETO-cased)	0.32
Best weighted ensemble (Rigoberta+MR-BERT+BERTIN)	0.35
Label-specific encoders (Rigoberta)	0.19
FSL prompting with Qwen3.5-9B	**0.47**

As shown in Table 2, the proposed prompting-based approach substantially outperformed all supervised encoder-based alternatives explored. In particular, the FSL configuration achieved an improvement of approximately +0.09 macro-F1 points over the best fine-tuned encoder configuration (Rigoberta) and +0.12 points over the best ensemble configuration. These results suggest that instruction-based LLM prompting is especially effective in scenarios characterized by strong semantic shift and limited annotated data, where traditional supervised approaches may struggle to generalize.

Interestingly, the ensemble configurations failed to surpass the best individual encoder. Although multiple combinations and weighting schemes were systematically explored, combining predictions from different encoder architectures did not lead to additional performance gains. This suggests that the errors made by the individual models were not sufficiently complementary to benefit from probability-level aggregation. Furthermore, the comparatively poor performance of the label-specific encoder

configuration indicates that independently modeling each emotion category is inadequate for this task, likely because the prediction of each emotion benefits from information shared across labels and from the broader contextual understanding provided by a unified model.

Table 3 presents the official ranking of the HISEMOTIONS 2026 shared task according to the official macro-F1 evaluation metric. Our system (*tbernal*) ranked 4th out of 10 participating teams, achieving a macro-F1 score of 0.47.

Table 3
Official HISEMOTIONS 2026 shared task ranking according to macro-F1 score.

Place	Team	Macro-F1
1	franrodrigo	0.58
2	carlosdf	0.57
3	explainators	0.56
4	**tbernal**	**0.47**
5	Sinai	0.45
6	UC-UCO-Plenitas	0.40
7	CE_LL_JB_UG	0.39
8	davidreyes	0.36
9	aldairverdugo	0.32
10	amrtodounam	0.13

As shown in Table 3, the top-performing systems achieved macro-F1 scores between 0.56 and 0.58, while our approach obtained a macro-F1 score of 0.47. This result positions our system within the upper half of the competition and comparatively close to the top-ranked submissions, despite relying on a comparatively simple prompting-based pipeline without task-specific fine-tuning or external resources.

To provide a more detailed analysis of the system behavior, Table 4 reports the official per-label precision, recall and F1-score obtained by our final submission on the test set.

Table 4
Per-label performance of the final system on the official test set.

Emotion	Precision	Recall	F1-score
anger	0.59	0.68	0.63
fear	0.45	0.67	0.54
joy	0.29	0.89	0.44
sadness	0.76	0.54	0.63
surprise	0.07	0.50	0.13
hope	0.29	0.81	0.43

6. Discussion

The results presented in Section 5 reveal several interesting patterns regarding the behavior of the proposed system.

As shown in Table 4, the per-label results reveal substantial variability across emotion categories. The best results were obtained for anger and sadness, both achieving F1-scores of 0.63. In particular, sadness achieved the highest precision (0.76), suggesting that the model is relatively reliable when predicting this category. This behavior is likely influenced by the predominance of sadness in the training set, which accounts for almost half of the examples.

In contrast, surprise proved to be the most difficult category, achieving an F1-score of only 0.13. This result is consistent with the extreme class imbalance observed in the dataset, where surprise represents less than 1% of the training examples. Similar difficulties were observed for hope and joy, where the model achieved high recall but comparatively low precision. This indicates that the system tends to

over-predict these categories, possibly because their affective signals are subtle, implicit or semantically ambiguous in historical texts.

6.1. Error Analysis

True positives are more frequent for emotions characterized by clear lexical or semantic cues, such as sadness. In contrast, false positives are more common for emotions whose manifestations depend strongly on context or pragmatic interpretation, particularly joy and hope, which exhibit high recall but comparatively low precision. Likewise, false negatives are expected when affective states are conveyed indirectly or through culturally specific expressions that differ from contemporary usage. These observations suggest that historical emotion detection remains challenging, even for modern LLMs.

6.2. Output Robustness

Interestingly, no malformed JSON outputs or invalid emotion labels were observed on the official test set. This confirms the effectiveness of the constrained prompting strategy and deterministic decoding for ensuring robust and reliable structured generation. Consequently, the prediction extraction process remained stable and no fallback mechanism was triggered during evaluation.

6.3. Limitations

Several limitations should be considered when interpreting these results. First, the dataset exhibits severe class imbalance, particularly for emotions such as anger and surprise. This restricts the ability of the model to learn rare affective patterns and likely contributes to the strong variability observed across emotion categories.

Second, the proposed approach relies exclusively on five randomly selected demonstrations and does not exploit more sophisticated example selection strategies or retrieval mechanisms. Alternative FSL strategies based on semantic similarity or dynamic example retrieval could potentially provide more informative demonstrations and improve generalization.

Third, although the prompting-based approach outperformed all encoder-based baselines, it still struggles with implicitly expressed emotions and culturally conditioned affective cues, which are common in historical correspondence. As a result, emotions whose expression depends strongly on context or pragmatic interpretation remain particularly difficult to detect.

Finally, the analysis presented in this work is primarily based on aggregate metrics and general error patterns. A more detailed qualitative study based on individual predictions and representative examples would provide further insight into the strengths and weaknesses of the system.

7. Conclusion and Future Work

In this work, we presented our system for the HISEMOTIONS 2026 shared task on emotion detection in Early Modern Spanish correspondence. Our approach relies on FSL prompting with Qwen3.5-9B, formulating the problem as a multi-label emotion classification task through structured JSON generation. The final system achieved a macro-F1 score of 0.47 on the official test set, ranking 4th out of 10 participating teams.

The obtained results demonstrate that prompting-based approaches using general-purpose LLMs constitute a competitive alternative for historical NLP scenarios characterized by semantic shift and limited annotated data. In particular, the proposed FSL strategy clearly outperformed all supervised encoder-based approaches explored during development, including fine-tuned transformer encoders, ensemble configurations and label-specific classifiers. These findings suggest that the inference-time adaptation capabilities of modern LLMs are especially beneficial when dealing with historical language varieties and complex affective phenomena.

Our discussion further revealed substantial variability across emotion categories and highlighted the difficulty of modeling implicit and context-dependent affective cues. These observations emphasize the intrinsic complexity of emotion detection in historical correspondence, where emotional expression is often indirect and culturally conditioned.

Despite the encouraging results, the analysis also revealed several limitations, including severe class imbalance, the reliance on a small set of demonstrations and the difficulty of recognizing implicitly expressed emotions.

For future work, several directions may further improve performance. First, exploring more advanced prompting strategies, such as retrieval-based example selection or dynamically generated demonstrations, could provide more informative contextual guidance during inference. Previous work has already shown the potential of optimizing few-shot example selection for hate speech detection using consistent retrieval mechanisms [28]. Second, incorporating reasoning-oriented prompting techniques or structured intermediate representations may help the model better capture subtle affective signals and multi-emotion interactions in historical texts.

Another promising direction involves combining prompting-based methods with lightweight parameter-efficient adaptation techniques, such as LoRA or QLoRA, to better specialize LLMs to the linguistic characteristics of Early Modern Spanish while maintaining relatively low computational requirements. Recent studies comparing prompting and fine-tuning strategies in Spanish NLP tasks suggest that combining both paradigms may provide complementary advantages depending on the domain and annotation setting [35, 36].

Finally, future work could also explore the integration of external historical lexical resources or diachronic semantic information in order to better address semantic drift and historically specific affective meanings. In addition, leveraging recent advances in Spanish emotion analysis and multilingual affective modeling may further improve the robustness of emotion detection systems for historical correspondence [37, 38].

Acknowledgments

This work is part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF/EU - FEDER/UE)-a way of making Europe. Mr. Tomás Bernal-Beltrán is supported by University of Murcia through the predoctoral programme.

Data availability

The implementation of the proposed FSL prompting approach, together with the scripts, prompts and experimental configuration required to reproduce the results presented in this work, is publicly available at:

https://github.com/tomasBernal/HISEMOTIONS2026.

The datasets used in this study were provided by the organizers of the HISEMOTIONS 2026 shared task and are described in Section 4. The repository includes the inference pipeline, prompt templates and evaluation scripts used to generate the submitted predictions.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL for grammatical and spelling correction. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] R. Kamal, M. A. Shah, C. Maple, M. Masood, A. Wahid, A. Mehmood, Emotion classification and crowd source sensing; a lexicon based approach, IEEE Access 7 (2019) 27124–27134.

[2] T. Zhang, A. M. Schoene, S. Ji, S. Ananiadou, Natural language processing applied to mental illness detection: a narrative review, NPJ digital medicine 5 (2022) 46.

[3] S. Moussa, An emoji-based metric for monitoring consumers' emotions toward brands on social media, Marketing Intelligence & Planning 37 (2019) 211–225.

[4] A. Kutuzov, L. Øvrelid, T. Szymanski, E. Velldal, Diachronic word embeddings and semantic shifts: a survey, in: Proceedings of the 27th international conference on computational linguistics, 2018, pp. 1384–1397.

[5] F. Periti, S. Montanelli, Lexical semantic change through large language models: a survey, ACM Computing Surveys 56 (2024) 1–38.

[6] T. Montes, L. Manrique-Gómez, R. Manrique, Historical ink: Semantic shift detection for 19th century spanish, in: Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change, 2024, pp. 29–41.

[7] 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).

[8] 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, pp. –.

[9] L. Canales, P. Martínez-Barco, Emotion detection from text: A survey, in: Proceedings of the workshop on natural language processing in the 5th information systems research working days (JISIC), 2014, pp. 37–43.

[10] A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, P. S. Dodds, The emotional arcs of stories are dominated by six basic shapes, EPJ data science 5 (2016) 31.

[11] R. Sprugnoli, S. Tonelli, A. Marchetti, G. Moretti, Towards sentiment analysis for historical texts, Digital Scholarship in the Humanities 31 (2016) 762–772.

[12] G. Schneider, Affecting correspondences: Body, behavior, and the textualization of emotion in early modern english letters, Prose Studies 23 (2000) 31–62.

[13] Qwen Team, Qwen3.5: Towards native multimodal agents, 2026. URL: https://qwen.ai/blog?id=qwen3.5.

[14] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[16] S. M. Mohammad, P. D. Turney, Crowdsourcing a word–emotion association lexicon, Computational intelligence 29 (2013) 436–465.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[18] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, S. Ravi, Goemotions: A dataset of fine-grained emotions, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 4040–4054.

[19] S. Mohammad, F. Bravo-Marquez, M. Salameh, S. Kiritchenko, Semeval-2018 task 1: Affect in tweets, in: Proceedings of the 12th international workshop on semantic evaluation, 2018, pp. 1–17.

[20] S. H. Muhammad, N. Ousidhoum, I. Abdulmumin, J. P. Wahle, T. Ruas, M. Beloucif, C. De Kock,

N. Surange, D. Teodorescu, I. S. Ahmad, et al., Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 8895–8916.
[21] S. M. Mohammad, From once upon a time to happily ever after: Tracking emotions in mail and books, Decision Support Systems 53 (2012) 730–741.
[22] M. Piotrowski, Natural language processing for historical texts, Morgan & Claypool Publishers, 2012.
[23] C. Sánchez-Marco, G. Boleda, L. Padró, Extending the tool, or how to annotate historical language varieties, in: Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities, 2011, pp. 1–9.
[24] F. Sánchez-Martínez, I. Martínez-Sempere, X. Ivars-Ribes, R. C. Carrasco, An open diachronic corpus of historical spanish, Language resources and evaluation 47 (2013) 1327–1342.
[25] N. Tahmasebi, L. Borin, A. Jatowt, Survey of computational approaches to lexical semantic change, arXiv preprint arXiv:1811.06278 (2018).
[26] D. Schlechtweg, B. McGillivray, S. Hengchen, H. Dubossarsky, N. Tahmasebi, Semeval-2020 task 1: Unsupervised lexical semantic change detection, in: Proceedings of the fourteenth workshop on semantic evaluation, 2020, pp. 1–23.
[27] M. Bollmann, A large-scale comparison of historical text normalization systems, in: Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers), 2019, pp. 3885–3898.
[28] R. Pan, J. A. García-Díaz, R. Valencia-García, Optimizing few-shot learning through a consistent retrieval extraction system for hate speech detection, Procesamiento del Lenguaje Natural 74 (2025) 241–252.
[29] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
[30] I. de Ingeniería del Conocimiento, Rigoberta-2.0, 2025. URL: https://huggingface.co/IIC/RigoBERTa-2.0. doi:10.57967/hf/7048.
[31] 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, pp. –.
[32] J. de la Rosa, E. G. Ponferrada, P. Villegas, P. G. de Prado Salas, M. Romero, M. Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23.
[33] J. Cañete, S. Donoso, F. Bravo-Marquez, A. Carvallo, V. Araujo, Albeto and distilbeto: Lightweight spanish language models, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 4291–4298.
[34] D. Tamayo, I. Lacunza, P. Rivera-Hidalgo, S. Da Dalt, J. Aula-Blasco, A. Gonzalez-Agirre, M. Villegas, Mrbert: Modern multilingual encoders via vocabulary, domain, and dimensional adaptation, arXiv preprint arXiv:2602.21379 (2026).
[35] R. Pan, T. Bernal-Beltrán, A. Rodríguez-González, E. Menasalvas-Ruíz, R. Valencia-García, Evaluating spanish medical entity recognition: Large language models with prompting versus fine-tuning, Computers, Materials and Continua 87 (2026).
[36] T. Bernal-Beltrán, R. Pan, J. A. García-Díaz, M. del Pilar Salas-Zárate, M. A. Paredes-Valverde, R. Valencia-García, Detection of maliciously disseminated hate speech in spanish using fine-tuning and in-context learning techniques with large language models, Computers, Materials, & Continua 87 (2026).
[37] A. Salmerón-Ríos, J. A. García-Díaz, R. Pan, R. Valencia-García, Fine grain emotion analysis in spanish using linguistic features and transformers, PeerJ Computer Science 10 (2024) e1992.
[38] R. Pan, T. Bernal-Beltrán, J. A. García-Díaz, R. Valencia-García, speech-emotion: A multilingual and multimodal toolkit for emotion recognition from speech, SoftwareX 34 (2026) 102677.