UCO-UC-CICESE-Plenitas Team at Exploring the Historical Text-Based Emotion Detection in Early Modern Spanish Correspondence Using Large Language Model and Embeddings

Yoan Martínez-López[1,2,*], Yanaima Jauriga[3], Mayte Guerra Saborit[3], Julio Madera[3], Luis Quintero Domínguez[4], Ansel Rodríguez-González[5], Carlos de Castro Lozano[1,2] and José Miguel Ramírez Uceda[1,2]

[1]*Universidad de Córdoba, Córdoba, Spain*
[2]*Plénitas, C/ Le Corbusier s/n, 14005 Córdoba, Spain*
[3]*Universidad de Camaguey, Circunvalación Norte, Camino Viejo Km 5 y 1/2, Camaguey, Cuba*
[4]*Universidad de Sancti Spíritus, Cuba*
[5]*Centro de Investigación Científica y de Educación Superior de Ensenada, Unidad Académica Tepic, México*

Abstract

This paper describes the participation of the UCO-UC-CICESE-Plenitas team in the HISEMOTIONS 2026 shared task at IberLEF, which addresses Text-Based Emotion Detection (TBED) in Early Modern Spanish epistolary correspondence from the 16th and 17th centuries. The task poses substantial challenges arising from diachronic semantic shift, orthographic variation, and the lexical distance between historical Spanish and the modern web text dominant in mainstream pre-training corpora. We propose a modular ensemble system that combines two complementary inductive biases: (i) an embedding-based branch using multilingual sentence embedding models paired with L2-regularised logistic regression classifiers trained under a binary relevance scheme with SMOTE oversampling, and (ii) a zero-shot branch built on instruction-tuned Large Language Models served locally through Ollama. A diverse pool of open-weight LLMs from the Qwen, Mistral, Gemma, Falcon, and Llama families (ranging from 4B to 14B parameters) was systematically evaluated to identify the most suitable reasoning backbone. The preprocessing pipeline deliberately preserves historical orthographic features in order to retain affective cues encoded in period-specific lexicon. Across two evaluation rounds, our system obtained macro-F1 scores of 0.42 and 0.40. The first result was achieved during a preliminary evaluation round involving three participating teams, whereas the second corresponds to the final official evaluation involving ten teams. In these rounds, our system ranked second and sixth, respectively. Our findings indicate that model family and training generation are more decisive than raw parameter count, and that retrieval augmentation based on modern Spanish embedding spaces can be counterproductive when applied to historically distant registers. Overall, Qwen-family models consistently provided the strongest performance across the evaluated configurations.

Keywords

Emotion detection, Large Language Models, historical NLP, sentence embeddings

1. Introduction

Text-Based Emotion Detection (TBED) has evolved into a cornerstone of Natural Language Processing (NLP), demonstrating significant algorithmic success in modern contexts such as mental health analysis, marketing research, and fake news detection. Within the Digital Humanities and Computational Literary Studies, emotion detection has been extensively applied to historical plays, novels, and political

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
†These authors contributed equally.
ymlopez2022@gmail.com (Y. Martínez-López); yanaima.jauriga@reduc.edu.cu (Y. Jauriga); mayte.guerra@reduc.edu.cu (M. G. Saborit); julio.madera@reduc.edu.cu (J. Madera); laqdominguez@gmail.com (L. Q. Domínguez); ansel@cicese.edu.mx (A. Rodríguez-González); carlosdecastrolozano@gmail.com (C. d. C. Lozano); p52raucj@uco.es (J. M. R. Uceda)
0000-0002-1950-567X (Y. Martínez-López); 0009-0000-6891-0068 (Y. Jauriga); 0000-0002-9556-5869 (M. G. Saborit); 0000-0001-5551-690X (J. Madera); 0000-0002-3527-0516 (L. Q. Domínguez); 0000-0001-9971-0237 (A. Rodríguez-González); 0000-0001-6603-843X (C. d. C. Lozano); 0000-0002-5027-7521 (J. M. R. Uceda)

texts. However, despite these advancements, the emotional dimension of other critical historical sources—specifically historical correspondence—remains largely under-examined. The HISEMOTIONS 2026 shared task addresses this gap by focusing on TBED in Early Modern Spanish epistolary texts from the 16th and 17th centuries [1, 2].

The primary challenge in analyzing historical language varieties lies in the limitations of current resources. Previous shared tasks within IberLEF have predominantly focused on contemporary Spanish and modern genres, such as social media and news. Consequently, the transferability of these modern models to historical varieties is hindered by distinct semantic, cultural, and linguistic conventions. A significant obstacle is semantic shift, where affective meanings in historical lexicons deviate substantially from modern usage. This diachronic evolution underscores the necessity for robust methods capable of capturing the complexity of affective states in Early Modern Spanish without relying solely on direct transfer from modern models.

HISEMOTIONS 2026 provides an opportunity to investigate how modern NLP techniques transfer to a historically distant language variety. In this work, we evaluate a diverse set of instruction-tuned Large Language Models and multilingual sentence embedding models for multi-label emotion detection in Early Modern Spanish correspondence. We further propose a modular ensemble architecture that combines embedding-based classifiers with zero-shot LLM reasoning in order to exploit both data-driven semantic representations and the contextual knowledge encoded in modern LLMs. Finally, we analyze the impact of model family, parameter scale, and retrieval-based augmentation strategies on performance in this historically grounded emotion classification task.

The remainder of this paper is organized as follows. Section 2 introduces the task definition and dataset. Section 3 describes the proposed methodology. Section 4 presents the evaluation protocol. Section 5 reports the experimental results. Finally, Section 6 concludes the paper and outlines directions for future work.

2. Task Definition and Dataset

2.1. Task Formalization

HISEMOTIONS 2026 is formalized as a multi-label text-based emotion detection (TBED) task over fragments of Early Modern Spanish epistolary correspondence (16th–17th centuries). Following the binary multi-label annotation framework proposed by Muhammad et al. [3] and adopted by the shared task organizers, each input fragment is associated with a binary indicator for each emotion category, capturing the possibility that a single clause may simultaneously express multiple affective states from the author's perspective.

Formally, let $\mathcal{F} = \{f_1, f_2, \ldots, f_N\}$ denote the set of emotion-bearing fragments, where each f_i is a contiguous span of text corresponding to a clause or sentence expressing a coherent affective state. Let $\mathcal{E} = \{\text{anger}, \text{fear}, \text{joy}, \text{sadness}, \text{surprise}, \text{hope}\}$ be the set of six emotion labels adopted from Ekman's basic emotion framework [4]—excluding *disgust* due to insufficient examples in the corpus—and extended with the *hope* category. The goal is to learn a function

$$\phi : \mathcal{F} \rightarrow \{0, 1\}^{|\mathcal{E}|} \tag{1}$$

that maps each fragment f_i to a binary vector $\mathbf{y}_i = (y_i^{\text{anger}}, y_i^{\text{fear}}, y_i^{\text{joy}}, y_i^{\text{sadness}}, y_i^{\text{surprise}}, y_i^{\text{hope}}) \in \{0, 1\}^6$, where $y_i^e = 1$ indicates that the author expresses emotion e at the time of writing.

The task focuses on the *perceived* emotions of the author, as interpreted by the annotator, rather than on third-party emotions referenced in the text or on rhetorical formulas typical of the epistolary genre (greetings, closings, courtesy expressions). This perception-based framing follows Muhammad et al. [5] and is explicitly enforced in our system through the prompt design described in Section 3.5.

2.2. Corpus Description

The shared task corpus is sourced from the *P. S. Post Scriptum* project [6, 7], a digital scholarly edition of unpublished private correspondence written in Portugal and Spain during the Early Modern period. The HISEMOTIONS 2026 release uses a selection of letters from the Spanish subcorpus (16th–17th centuries), authored by individuals of diverse social backgrounds: men and women, adults and children, masters and servants, soldiers, artisans, clergy, and political figures. Texts are segmented into emotion-bearing fragments, with all fragments derived from the same letter kept within a single split to prevent data leakage. The dataset is partitioned as follows:

- **Training set** (∼2,500 fragments): released with baseline LLM-annotated labels.
- **Development set** (∼400 fragments): released with gold human-annotated labels, used for model selection and threshold tuning.
- **Test set** (∼800 fragments): released without labels, used for final evaluation through the Codabench platform.

The dataset exhibits a pronounced class imbalance among the six emotion categories, a direct consequence of its authentic, naturally occurring historical content. In particular, emotions such as *hope* and *joy* are substantially more frequent than *surprise* and *fear*.

These characteristics motivate the methodological choices described in the next section, particularly regarding preprocessing, class imbalance handling, and model selection.

3. Methodology

3.1. Embedding Models

To evaluate the role of semantic representations in historical emotion detection, we selected a set of multilingual sentence embedding models that provide complementary encoding capabilities and have demonstrated strong performance on retrieval and semantic similarity benchmarks. The selected models differ in their training objectives, multilingual coverage, and retrieval strategies, allowing us to assess how modern embedding spaces transfer to Early Modern Spanish texts, a language variety characterized by substantial lexical, orthographic, and semantic differences from contemporary corpora.

In addition to the generative LLMs described in the following section, several pre-trained sentence embedding models were integrated into the experimental configurations to provide dense semantic representations of the input texts, either as features for downstream classifiers or as retrieval signals for context augmentation.

The LaBSE model (Language-agnostic BERT Sentence Embedding) [8] was employed as a strong multilingual baseline. Trained on parallel corpora covering 109 languages with a dual-encoder translation-ranking objective, LaBSE produces 768-dimensional embeddings that are well aligned across languages and have been shown to perform robustly on cross-lingual semantic similarity tasks, including those involving low-resource and historically distant language varieties.

The Nomic Embed Text model [9], released by Nomic AI, was also evaluated. It is a fully open-source long-context encoder (up to 8,192 tokens) trained with contrastive learning on a large-scale web corpus, designed to provide competitive performance on retrieval and clustering benchmarks while remaining lightweight enough for local deployment.

Finally, the BAAI/bge-m3 model [10] was considered for its multi-functional capabilities, combining dense, sparse, and multi-vector retrieval within a single encoder. Trained on a large multilingual corpus with explicit support for over 100 languages and document-level context windows of up to 8,192 tokens, bge-m3 has consistently ranked among the top open-weight multilingual encoders on retrieval benchmarks such as MIRACL and MTEB.

These embedding models were used either as the encoding component of retrieval-augmented configurations or as feature extractors for classical classifiers within the proposed pipeline. Their diversity—covering language-agnostic alignment (LaBSE), open long-context retrieval (Nomic), and

multi-functional multilingual representation (bge-m3)—enables a controlled assessment of how different encoder paradigms interact with Early Modern Spanish.

3.2. Large Language Models

To identify the most suitable reasoning backbone for emotion detection in Early Modern Spanish correspondence, we evaluated a diverse set of open-weight instruction-tuned Large Language Models (LLMs). The selected models represent different architectural families, parameter scales, and training generations, allowing us to analyze how these factors influence performance on a historically distant language variety. In particular, we were interested in determining whether recent advances in instruction tuning and multilingual training provide greater benefits than simply increasing model size.

A diverse set of open-weight instruction-tuned Large Language Models (LLMs) was therefore evaluated as the reasoning backbone of our pipeline. The selection covers four major model families released between 2024 and 2025, spanning a parameter range from 4B to 14B parameters and representative of the current state of the art in publicly available LLMs.

The Qwen family, developed by Alibaba Cloud, was represented by three checkpoints: Qwen2.5-7B and Qwen2.5-14B [11], both instruction-tuned variants of the second-generation Qwen series trained on a multilingual corpus of approximately 18 trillion tokens with extended context and strong multilingual coverage, and the more recent Qwen3-8B [11], which incorporates an improved post-training pipeline and enhanced reasoning capabilities through hybrid thinking modes.

The Mistral family was represented by Mistral-7B-Instruct [12], a dense decoder-only model based on grouped-query and sliding-window attention, and by Ministral-14B, a more recent extension within the same architectural lineage designed to provide stronger reasoning at moderate scale.

From the Gemma family, developed by Google DeepMind, we tested Gemma-2-9B [13] and the third-generation models Gemma-3-4B and Gemma-3-12B [14], which introduce improved tokenization, longer context windows, and refined alignment procedures relative to their predecessors.

Additionally, the Falcon family, developed by the Technology Innovation Institute, was represented by Falcon-3-10B-Instruct, a model designed with an emphasis on efficient inference and competitive performance across multilingual benchmarks. Finally, the Llama 3.1 8B Instruct model [15] was included as a widely adopted multilingual baseline frequently used in recent NLP shared tasks.

All models were evaluated in their instruction-tuned variants and served locally through Ollama, ensuring reproducibility and full control over inference parameters. This heterogeneous selection enables a controlled comparison of model family, parameter count, and training generation for historical Spanish emotion detection, a domain that remains largely underrepresented in mainstream pre-training corpora.

3.3. Preprocessing Pipeline

The preprocessing strategy is intentionally conservative, in order to preserve the linguistic features that carry diachronic and affective information. The pipeline implemented in our system (function `preprocess_historical_text`) performs the following operations:

1. **Character preservation**. Historical orthographic variants (e.g., *ç*, double consonants, etymological *x*/*j*, diacritical conventions) are deliberately retained rather than normalized to Modern Spanish. This decision is motivated by two considerations: (i) normalization risks erasing semantic and emotional cues encoded in period-specific lexicon (e.g., *esperar* as "to await" versus "to hope"), and (ii) modern multilingual LLMs and contextual encoders are sufficiently robust to recover meaning from minor surface variation, while fine-tuning further adapts them to the historical register.
2. **Whitespace normalization**. Multiple whitespace characters, line breaks, and trailing spaces are collapsed into a single space (`re.sub(r"\s+", " ", text)`) to ensure consistent tokenization across heterogeneous transcription conventions.

3. **Length capping**. Fragments are truncated at MAX_LENGTH $= 256$ whitespace tokens, which exceeds the natural length of the vast majority of fragments and preserves the full clause structure of the corpus.
4. **NaN handling**. Missing values in the label columns of the training and development splits are replaced with 0 rather than discarded (function `clean_dataframe_nan`), preserving the fragment text and treating absent annotations as negative evidence under the multi-label binary scheme.

3.4. Class Imbalance Handling

To mitigate the imbalance among emotion categories at training time, Synthetic Minority Over-sampling Technique (SMOTE) [16] is applied independently to each emotion in the embedding-based branch of the system. For each emotion $e \in \mathcal{E}$, SMOTE is invoked with k_neighbors $= 3$ on the embedding feature matrix whenever both the positive and negative classes contain at least five instances; otherwise the original distribution is preserved. This binary relevance decomposition allows each emotion-specific classifier to operate on a more balanced feature distribution without altering the joint label structure of the multi-label problem.

3.5. System Architecture

Our submission to HISEMOTIONS 2026 is a modular ensemble system that combines complementary inductive biases: (i) dense semantic representations from a multilingual sentence encoder coupled with calibrated linear classifiers, and (ii) zero-shot prompting of an instruction-tuned large language model served locally via Ollama. The architecture is intentionally designed so that each branch can also act as a standalone baseline, facilitating ablation analyses.

3.5.1. Branch A: Embedding-based Classifier

The first branch encodes each preprocessed fragment using a multilingual sentence encoder—in our configuration, `google/embeddinggemma-300M` [17], with `BAAI/bge-m3`, as an alternative tested during development—producing a dense vector $\mathbf{e}_i \in \mathbb{R}^d$ for each fragment. The choice of a multilingual model is motivated by its demonstrated robustness to non-canonical Spanish varieties and by its ability to capture lexical-semantic regularities that survive diachronic drift.

For each emotion $e \in \mathcal{E}$ we train an independent binary classifier under the binary relevance scheme. We adopt L_2-regularized logistic regression with the `lbfgs` solver, balanced class weights, and a maximum of 1,000 iterations. SMOTE is applied to the training embeddings as described in Section 3.4, and a common decision threshold $\tau = 0.5$ is applied to the predicted probabilities to produce the final binary output. The choice of logistic regression over deeper alternatives (e.g., MLPs, fine-tuned transformers) is grounded in the limited size of the training split ($\sim$2,500 fragments) and in the well-known sample efficiency of linear models over high-quality contextual embeddings.

3.5.2. Branch B: Zero-shot LLM Prompting

The second branch leverages an instruction-tuned LLM (`gemma2:9b`, served locally through Ollama) to perform zero-shot multi-label classification. The prompt is written in Spanish and includes:

1. A role specification framing the model as an expert in emotion analysis of Early Modern Spanish texts.
2. An explicit definition of each emotion in domain-appropriate Spanish vocabulary (e.g., *anger* $\rightarrow$ "ira, enojo, rabia, indignación"; *hope* $\rightarrow$ "esperanza, expectativa, deseo de futuro").
3. Four task-specific instructions: (i) detect only the author's emotions at the moment of writing, (ii) ignore epistolary courtesy formulas, (iii) account for the historical-semantic context of the period, and (iv) emit exactly six comma-separated binary digits in a fixed canonical order.

4. The fragment under analysis, delimited by quotation marks.

The LLM is queried with low-temperature decoding (`temperature` = 0.1, `num_predict` = 20) to maximize determinism. A robust regular-expression parser then extracts the six binary digits from the generated string, with a secondary pattern-matching fallback and a safe "all-zero" default for the rare cases in which the output cannot be parsed.

3.5.3. Branch B′: Few-shot Extension

A few-shot variant of Branch B was also developed and is retained as an optional component of the system. For each emotion, up to $n_{\text{ex}} = 2$ positive examples are sampled from the training set in a balanced manner, complemented by two fully negative examples (all-zero label vector). These examples are inlined into the prompt prior to the target fragment. While few-shot prompting consistently improves recall on low-frequency emotions during preliminary experiments, it also increases prompt length and inference time; we therefore expose it as a configurable strategy rather than a default.

3.5.4. Ensemble Aggregation

The final predictions are produced by an ensemble that combines Branch A (embedding classifier) and Branch B (zero-shot prompting) through a weighted average of their per-emotion binary outputs. For a fragment i and emotion e, let $p_i^{e,A}, p_i^{e,B} \in \{0, 1\}$ be the predictions of each branch and w_A, w_B their respective weights (both set to 1.0 in our submission). The ensemble score is

$$s_i^e = \frac{w_A \, p_i^{e,A} + w_B \, p_i^{e,B}}{w_A + w_B}, \tag{2}$$

and the final prediction is $\hat{y}_i^e = \mathbb{1}[s_i^e \geq \tau]$ with $\tau = 0.5$, i.e., an emotion is predicted whenever at least half of the weighted vote supports it. This simple but effective aggregation provides redundancy: the embedding branch contributes stable, training-data-informed decisions, while the LLM branch contributes world-knowledge-grounded judgments about historical semantics that are not fully recoverable from the small training set.

3.6. Implementation Details

The system is implemented in Python 3 using `scikit-learn`, `sentence-transformers`, `transformers`, `imbalanced-learn`, and the `ollama` client. All experiments fix the random seed to 42 for reproducibility. Training and inference are executed on a single GPU when available (`torch.cuda.is_available()`) and fall back transparently to CPU otherwise. The expected runtime of the full ensemble on the development set is dominated by the LLM branch.

4. Evaluation Protocol

Following the official HISEMOTIONS 2026 evaluation protocol, the primary ranking metric is the **macro-averaged F1 score** computed across the six emotion categories. This choice is consistent with the multi-label nature of the task and with the marked class imbalance of the corpus: macro-averaging treats each emotion as equally important regardless of its support, thereby penalizing systems that achieve high overall accuracy by neglecting low-frequency categories such as *fear* or *surprise*.

Formally, for each emotion $e \in \mathcal{E}$ we compute precision, recall, and F1 score in the standard binary fashion:

$$P_e = \frac{TP_e}{TP_e + FP_e}, \quad R_e = \frac{TP_e}{TP_e + FN_e}, \quad F_{1,e} = \frac{2P_eR_e}{P_e + R_e}, \tag{3}$$

where TP_e, FP_e, and FN_e denote, respectively, the number of true positives, false positives, and false negatives for emotion e. The macro-averaged F1 score is then

$$\text{Macro-}F_1 = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} F_{1,e}. \tag{4}$$

In addition to Macro-F_1, we report:

- **Micro-averaged F1** (Micro-F_1), which aggregates true and false positives/negatives across all emotions before computing F1. This metric is dominated by frequent emotions and provides a useful contrast against Macro-F_1 when diagnosing imbalance-related effects.
- **Macro-averaged precision and recall**, which jointly help determine whether the system is over- or under-predicting and guide threshold tuning.
- **Per-emotion F1 scores** ($F_{1,e}$ for each $e \in \mathcal{E}$), which are essential for diagnosing the behavior of the system on low-frequency categories and for prioritizing threshold adjustments per emotion in future iterations.

All metrics are computed using the standard `scikit-learn` implementations with the parameter `zero_division=0`, so that emotions with no predicted positives contribute a score of zero rather than raising an exception. Evaluation is performed on the development split during model selection and threshold calibration, and on the test split through the official Codabench platform for the final leaderboard ranking. Reported development-set scores should therefore be interpreted as an upper-bound estimate of test-set performance, given the moderate domain shift between the two splits.

The following section presents the results obtained under this evaluation protocol on both the development and official test sets.

5. Results

The system was developed in Python using the PyTorch ecosystem (Transformers, sentence-transformers, scikit-learn, pandas, NumPy, PIL). Local LLMs were managed via Ollama, launched as a background service. Inference was executed on a CUDA-enabled GPU when available, with automatic device detection and precision selection (float16 on GPU, float32 on CPU). Authentication with the Hugging Face Hub provided access to gated models and to the Inference API.

5.1. Development

Table 1 reports the performance of three open-source instruction-tuned LLMs evaluated under identical zero-shot conditions on the historical emotion detection task. Among the tested backbones, Qwen2.5-7B achieves the highest score (0.42), outperforming both Gemma-3-4B (0.32) and Mistral-7B (0.27) by margins of 10 and 15 absolute points, respectively. Notably, Gemma-3-4B surpasses the larger Mistral-7B despite having approximately half the parameter count, suggesting that raw model size is not the dominant factor governing performance on this task. These results point to two relevant observations. First, the considerable gap between Qwen2.5-7B and Mistral-7B (a relative improvement of roughly 56%) indicates that pre-training data composition and instruction-tuning recipes have a stronger impact than parameter count for emotion classification in Early Modern Spanish correspondence, a domain that is linguistically distant from the high-resource modern text typically dominant in pre-training corpora. Second, the comparatively strong performance of Gemma-3-4B suggests that recent training advances can partially compensate for reduced model capacity, making it a competitive lightweight alternative when computational resources are constrained. Overall, Qwen2.5-7B emerges as the most suitable backbone for this task and is therefore retained as the base model for the subsequent ensemble and fine-tuning experiments.

Table 2 presents the official leaderboard of the shared task, ranking all participating teams by macro-F1 on the test set. Our system, UCO-UC-CICESE-Plenitas, obtained a macro-F1 of 0.42, placing second

Table 1
Internal Comparisons

Method	Score
qwen2.5.7b	0.42
mistral7b	0.27
gemma3.4b	0.32

among the three participants. The top-ranked submission, dbqui1706, achieved 0.46, while the third entry, amrtodounam, reached 0.25. Two observations are worth highlighting. First, the gap between our system and the top-ranked submission is relatively narrow (4 absolute points, or roughly 9% in relative terms), indicating that our approach is competitive with the best-performing system in the task. Second, the substantial margin separating the top two systems from the third-ranked submission (17 absolute points) suggests a clear divide between approaches that successfully captured the linguistic and pragmatic particularities of Early Modern Spanish correspondence and those that did not. This bimodal distribution of results is consistent with the inherent difficulty of emotion detection in historical Spanish, where lexical archaism, orthographic variation, and shifts in emotional conceptualization challenge models trained predominantly on contemporary corpora. The performance of UCO-UC-CICESE-Plenitas validates the design choices underlying our pipeline, particularly the selection of Qwen2.5-7B as the backbone model (cf. Section 3.2) and the combination of zero-shot LLM voting with classical classifiers calibrated for the imbalanced emotion distribution.

Table 2
All teams/participants

Rank	Team/Participant	Macro F1
1	dbqui1706	0.46
2	**UCO-UC-CICESE-Plenitas**	0.42
3	amrtodounam	0.25

5.2. Evaluation

Table 3 extends the internal evaluation to a broader set of 12 configurations, including standalone open-source LLMs of varying sizes (4–14B parameters) and several ensemble strategies combining sentence embeddings with LLM-based reasoning. Scores are reported on the development set under identical zero-shot conditions. The results reveal three main findings. First, the best individual performance is shared by three configurations at the top of the table: Qwen3-8B (0.40), Gemma-3-12B (0.40), and a narrow group around 0.39 comprising Qwen2.5-14B and Qwen2.5-7B. The fact that Qwen3-8B matches or exceeds substantially larger models confirms the impact of recent generational improvements over raw parameter scaling, an observation already reported in the literature on instruction-tuned LLMs. Second, model family matters more than size: within the Mistral family, neither Mistral-7B (0.27) nor Ministral-14B (0.32) reach the level of the Qwen and Gemma backbones, while Falcon-3-10B (0.34) sits in an intermediate position. Third, and perhaps most informative for system design, the ensemble configurations relying on retrieval or auxiliary embeddings underperform their corresponding standalone backbones: `infobase_qwen2.5.7b` (0.31) and `emsb_nomatic_qwen2.5.7b` (0.29) both fall below the standalone Qwen2.5-7B (0.39), and the Nomic-based ensemble with Mistral drops to 0.22. This suggests that the additional retrieval signal — likely based on modern Spanish embedding spaces — introduces noise rather than informative context when applied to Early Modern Spanish, where the lexical and orthographic gap between query and indexed material undermines retrieval quality. Taken together, these observations support the decision to base the final submitted system on a Qwen-family backbone without retrieval augmentation, prioritising the LLM's intrinsic linguistic capabilities over

external embedding-based context that proved counterproductive in this historical-text setting.

Table 3
Internal Comparisons

Method	Score
qwen.2.5.14b	0.39
qwen.2.5.7	0.39
gemma2.9b	0.32
infobase_qwen2.5.7b	0.31
emsemb_labse_mistral3.14b	0.27
emsb_nomatic_qwen2.5.7b	0.29
ensemb_nomatic_mistral	0.22
qwen3.8b	0.40
falcon3.10b	0.34
gemma3.12b	0.40
ministral3.14b	0.32
mistral7b	0.27

Table 4 reports the official leaderboard, ranking the ten participating teams by macro-F1 on the test set. Our system, UCO-UC-CICESE-Plenitas, achieved a macro-F1 of 0.40, placing sixth out of ten participants. The top of the ranking is led by franrodrigo (0.58), closely followed by carlosdf (0.57) and explainators (0.56), while the bottom of the table is occupied by amrtodounam with a markedly lower score of 0.13.

Several observations emerge from this distribution. First, the leaderboard exhibits a three-tier structure: a top cluster of three systems within a 0.02-point margin (0.56–0.58), an intermediate group of two systems around 0.45–0.47, and a third tier of five systems between 0.32 and 0.40, in which our submission is placed. The compactness of the top cluster suggests that the leading teams may have converged on similar architectural choices — likely involving fine-tuned transformer encoders rather than zero-shot LLM pipelines. Second, the gap separating our system from the top-ranked submission (0.18 absolute points) is substantial and indicates clear room for improvement, particularly in the modelling of the linguistic and pragmatic phenomena that characterise this task. Despite the mid-table position, our system outperforms four of the ten participants and remains within a competitive range of the third tier. The result is consistent with the internal evaluation reported in Section 3.2, in which the best zero-shot LLM configurations plateaued around 0.40 macro-F1. This convergence between development and test performance suggests that our pipeline generalises reliably but is constrained by the inherent limitations of zero-shot inference on a domain-specific task. We attribute the gap with respect to the top-ranked systems primarily to the absence of task-specific fine-tuning and to the lexical mismatch between modern instruction-tuned LLMs and the target text genre.

Table 4
Results for each team/participant

Rank	Team/Participant	Macro F1
1	franrodrigo	0.58
2	carlosdf	0.57
3	explainators	0.56
4	tbernal	0.47
5	Sinai	0.45
6	**UCO-UC-CICESE-Plenitas**	0.4
7	CE_LL_JB_UG	0.39
8	davidreyes	0.36
9	aldairverdugo	0.32
10	amrtodounam	0.13

Overall, the results indicate that modern instruction-tuned LLMs can provide competitive performance on historical emotion detection without task-specific fine-tuning. However, they also reveal clear limitations associated with domain mismatch and the use of retrieval components trained on contemporary corpora. These findings are further discussed in the concluding section.

6. Conclusions

This work has presented the UCO-UC-CICESE-Plenitas system for the HISEMOTIONS 2026 shared task on emotion detection in Early Modern Spanish correspondence. We approached the task as a multi-label binary classification problem over six emotion categories — anger, fear, joy, sadness, surprise, and hope — and addressed it through a modular ensemble combining an embedding-based classifier branch with a zero-shot LLM prompting branch, aggregated via weighted voting.

Several conclusions can be drawn from our experimental analysis. First, the systematic comparison of 12 configurations from five model families demonstrated that model family and training generation outweigh parameter count as predictors of performance on this task: Qwen3-8B and Gemma-3-12B matched or exceeded substantially larger models, while the Mistral family consistently underperformed regardless of scale. Second, contrary to common practice in modern NLP, retrieval-augmented configurations based on multilingual sentence encoders proved counterproductive in our setting: every ensemble configuration combining a backbone LLM with auxiliary retrieval signals underperformed its corresponding standalone backbone. We interpret this as evidence that the lexical and orthographic gap between Early Modern Spanish queries and the modern web text underlying the embedding spaces introduces more noise than informative context, undermining the value of retrieval augmentation for historically distant registers. Third, the conservative preprocessing strategy — retaining historical orthographic variants rather than normalising them to Modern Spanish — proved compatible with modern multilingual encoders and with instruction-tuned LLMs, supporting the hypothesis that period-specific surface forms carry affective and semantic information that should not be erased.

The official evaluation placed UCO-UC-CICESE-Plenitas 2nd out of three participants (macro-F1 = 0.42) in the first round and 6th out of ten (macro-F1 = 0.40) in the second, with results that are consistent between development and test splits. The convergence between internal and official scores indicates that our pipeline generalises reliably but is constrained by the inherent limitations of zero-shot inference on a domain-specific task. The 0.18-point gap separating our system from the top-ranked submission suggests that the leading teams have likely benefited from task-specific fine-tuning of transformer encoders on the training data, an avenue we did not explore in this edition.

Our study is subject to several limitations. First, all evaluated LLMs were used in zero-shot settings, without task-specific fine-tuning on the HISEMOTIONS training corpus. Second, the sentence embedding models and instruction-tuned LLMs employed in this work were primarily trained on contemporary text, which may limit their ability to capture lexical, orthographic, and semantic phenomena specific to Early Modern Spanish. Finally, the relatively small size of the annotated dataset restricts the exploration of more data-intensive architectures and may affect the robustness of the conclusions regarding low-frequency emotions.

Several directions for future work emerge from this analysis. First, task-specific fine-tuning of a Qwen-family backbone — for instance via LoRA adapters or full fine-tuning of a smaller variant — on the HISEMOTIONS training set is the most immediate path to closing the gap with the top-ranked systems. Second, domain-adaptive pre-training on broader Early Modern Spanish corpora, such as the full P. S. Post Scriptum collection or comparable diachronic resources, could provide a more linguistically aligned starting point than current multilingual models. Third, historical-aware retrieval augmentation, for example through embeddings trained or adapted on period-specific corpora, would address the failure mode of modern-only embedding spaces identified in our internal evaluation. Finally, per-emotion threshold calibration and few-shot prompt optimisation — particularly for low-frequency categories such as fear and surprise — represent low-cost refinements that could yield measurable gains within the existing architecture.

Beyond the specific task, the HISEMOTIONS 2026 shared task highlights the broader value of shared tasks targeting historical language varieties: they expose the limitations of mainstream pre-trained models on linguistically distant registers and provide a controlled setting in which architectural and methodological choices can be evaluated against a common benchmark. We hope that the analysis presented here contributes to ongoing efforts to extend the reach of modern NLP to the rich, under-explored landscape of historical Spanish correspondence.

Declaration on Generative AI

During the preparation of this work, the authors used Claude(Opus 4.7) and Grammarly in order to: Grammar and spelling check. After using these services, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

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] S. H. Muhammad, N. Ousidhoum, I. Abdulmumin, S. M. Yimam, J. P. Wahle, T. L. Ruas, 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.

[4] P. Ekman, An argument for basic emotions, Cognition & Emotion 6 (1992) 169–200. doi:10.1080/02699939208411068.

[5] S. H. Muhammad, N. Ousidhoum, I. Abdulmumin, J. P. Wahle, T. Ruas, M. Beloucif, 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.

[6] G. Vaamonde, P. S. Post Scriptum. dos corpus diacrónicos de escritura cotidiana, Procesamiento del Lenguaje Natural 55 (2015) 57–64.

[7] H. Ehrlicher, R. Klinger, J. Lehmann, S. Padó, Measuring historical emotions and their evolution: An interdisciplinary endeavour to investigate the "emotions of encounter", Liinc em Revista 15 (2019).

[8] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic BERT sentence embedding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2022, pp. 878–891.

[9] Z. Nussbaum, J. X. Morris, B. Duderstadt, A. Mulyar, Nomic embed: Training a reproducible long context text embedder, arXiv preprint arXiv:2402.01613 (2024). URL: https://arxiv.org/abs/2402.01613. doi:10.48550/arXiv.2402.01613.

[10] J. P. Gujjar, H. P. Kumar, Text similarity technique using BGE-M3, in: 2025 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE), IEEE, 2025, pp. 1–7.

[11] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, others, Z. Qiu, Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025). URL: https://arxiv.org/abs/2505.09388. doi:10.48550/arXiv.2505.09388.

[12] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril,

T. Wang, T. Lacroix, W. El Sayed, Mistral 7B, arXiv preprint arXiv:2310.06825 (2023). URL: https://arxiv.org/abs/2310.06825. doi:10.48550/arXiv.2310.06825.
[13] Gemma Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, others, S. Iqbal, Gemma 3 technical report, arXiv preprint arXiv:2503.19786 (2025). doi:10.48550/arXiv.2503.19786.
[14] R. D. Labati, A. Ferrara, S. Picascia, V. Piuri, E. Rocchetti, F. Scotti, Investigating vision-language models biometric capabilities via sequence-based predictions (2025). URL: https://ssrn.com/abstract=5381090, available at SSRN 5381090.
[15] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, others, P. Vasic, The Llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024). URL: https://arxiv.org/abs/2407.21783. doi:10.48550/arXiv.2407.21783.
[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic minority oversampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357. doi:10.1613/jair.953.
[17] T. A. Laine, Quantum LLMs using quantum computing to analyze and process semantic information, arXiv preprint arXiv:2512.02619 (2025). URL: https://arxiv.org/abs/2512.02619. doi:10.48550/arXiv.2512.02619.

A. Online Resources

The results of the HISEMOTIONS 2026 are available via:

- HISEMOTIONS 2026,
- HISEMOTIONS 2026 Github,
- HISEMOTIONS 2026 Code