Procesamiento del Lenguaje Natural, Revista nº 77, septiembre de 2026, pp 551–563 recibido 03-07-2026 revisado 20-07-2026 aceptado 31-07-2026

Overview of HISEMOTIONS at IberLEF 2026: Historical Text-Based Emotion Detection in Early Modern Spanish Correspondence

Resumen de HISEMOTIONS en IberLEF 2026: Detección de Emociones en Textos Históricos de la Correspondencia Española de la Edad Moderna

Albina Sarymsakova,[1] Patricia Martín Rodilla,[1]
Eugenio Martínez Cámara,[2] Luis Alfonso Ureña López[2]
[1]Instituto de Estudios Gallegos Padre Sarmiento (IEGPS-CSIC)
[2]SINAI Research Group, Center for Advanced Studies in ICT,
Universidad de Jaén
{albina.s, p.m.rodilla}@iegps.csic.es, {emcamara, laurena}@ujaen.es

Abstract: This paper presents the results of HISEMOTIONS, a shared task at IberLEF 2026 focused on Text-Based Emotion Detection in Early Modern Spanish epistolary texts from the 16th and 17th centuries. The task aims to foster the development and evaluation of computational methods specifically adapted to historical Spanish language varieties, addressing the semantic and diachronic challenges of expressing emotion and affective states in Early Modern Spanish correspondence. The core challenge proposed within HISEMOTIONS is Multi-label Emotion Detection, oriented toward predicting the emotion perceived by a reader when interpreting an author's text. The final ranking encompasses the results of 7 teams, each of which proposed a distinct approach to the problem.
Keywords: emotions, historical texts, emotion detection in Spanish.

Resumen: Este artículo presenta los resultados de HISEMOTIONS, una tarea compartida en IberLEF 2026 centrada en la detección de emociones en textos epistolares del español de la Edad Moderna temprana de los siglos XVI y XVII. La tarea tiene como objetivo fomentar el desarrollo y la evaluación de métodos computacionales específicamente adaptados a las variedades históricas de la lengua española, abordando los desafíos semánticos y diacrónicos de la expresión de emociones y estados afectivos en la correspondencia de la Edad Moderna temprana en español. El desafío principal propuesto en HISEMOTIONS es la detección de emociones multi-etiqueta, orientada a predecir la emoción percibida por un lector al interpretar el texto de los autores. La clasificación final recoge los resultados de 7 equipos, cada uno de los cuales propuso un enfoque distinto para este desafío.
Palabras clave: emociones, textos históricos, detección de emociones en español.

1 Introduction

The Text-Based Emotion Detection (TBED) is a well-established task in modern Natural Language Processing (NLP). To date, TBED has demonstrated promising algorithmic performance, particularly for English-language data, mainly due to the availability of rich linguistic and lexical resources (Maks and Vossen, 2011; Valitutti, Strapparava, and Stock, 2004). Emotion Detection (ED) approaches have been successfully applied across a wide range of domains, including mental health analysis (Yang et al., 2023), customer behavior and marketing research (Wemmer, Labat, and Klinger, 2024), fake news detection (Zhang et al., 2023), and the humanities, where they support the analysis of emotion dimensions in literary texts (Santangelo, 2023).

Within the fields of Digital Humanities and Computational Literary Studies, ED has attracted substantial attention (Kim and Klinger, 2018). It has been used to analyze emotions in historical plays (Yavuz, 2020; Schmidt, Dennerlein, and Wolff, 2021), nov-

ISSN 1135-5948 DOI 10.26342-2026-77-38 © 2026 Sociedad Española para el Procesamiento del Lenguaje Natural

els (Reagan et al., 2016), fairy tales (Mohammad, 2011), and political texts (Sprugnoli et al., 2016).

Despite these advances, relatively few studies have examined the emotion dimension of other types of historical sources, such as historical correspondence (Gatti and Huesler, 2025; Turunen et al., 2022; Leemans et al., 2017; Mohammad, 2011). In particular, the existing literature lacks research assessing the broader applicability of emotion mining methods across specific linguistic and historical contexts, such as Spanish historical epistolary corpora.

The motivation for HISEMOTIONS (Sarymsakova et al., 2026a) at IberLEF 2026 (Bonet-Jover, González-Barba, and Chiruzzo, 2026) is to address the challenge that semantic shifting in Spanish poses for emotion detection in historical texts (Hu, Amaral, and Kübler, 2022; Periti and Tahmasebi, 2024; Montes, Manrique-Gómez, and Manrique, 2024), given that the lexicons in this domain often exhibit substantial deviations between historical and modern affective meanings. This challenge underscores the need for robust methods capable of capturing and detecting the complex nature of affective states expressed in historical correspondence. Another motivation for this shared task is to investigate how large language models (LLMs), language models (LMs), and Machine Learning (ML) approaches can contribute to the field of TBED in the domain of Spanish historical correspondence, given their promising performance on similar tasks in social media contexts (Plaza del Arco et al., 2020).

Based on these motivations, we formulate the following hypothesis: existing systems require adjustments to tackle TBED in Spanish historical correspondence effectively. This assumption is grounded in the observation that modern systems, including LLMs or LMs, are primarily trained on contemporary language, which limits their ability to interpret emotions in 16th-17th century Spanish texts due to semantic shifting in affective meanings. To test this hypothesis, we organized the HISEMOTIONS shared task at IberLEF 2026, with the goal of collecting participants' findings on which systems are applicable to this task and how they might be improved. Specifically, HISEMOTIONS is designed to address the following research questions:

RQ1: How well can existing systems tackle TBED in Early Modern Spanish correspondence?

RQ2: What improvements are needed to better address this task?

The rest of this paper is organized as follows. Section 2 provides the HISEMOTIONS task description. Next, the dataset assembled for this task, as well as baseline experiments, are delineated in Section 3. The strategies and methods used by the participants to address the HISEMOTIONS task, along with the official leaderboard, are presented in Section 4. Section 5 discusses the participants' results for the proposed systems. The error analysis of the proposed systems is detailed in Section 6. Finally, Section 7 presents task-driven conclusions and debates further challenges.

2 Task Description

The HISEMOTIONS shared task constitutes, to the best of our knowledge, the first within the IberLEF campaign addressing ED in Spanish historical correspondence. This edition aims to serve as a foundational task, establishing benchmarks for automatic emotion detection in Early Spanish correspondence from the 16th and 17th centuries and providing the first publicly available annotated resources for this domain.

The core challenge of HISEMOTIONS 2026 is to predict the perceived emotion or emotions as expressed by the author of a given letter fragment. For each textual fragment, participants must determine the presence or absence of the following six emotions: joy, sadness, fear, anger, surprise, and hope. The task is thus framed as multi-label binary classification, where each emotion constitutes an independent binary decision: present (1) or absent (0), as examples in Table 1 exhibit.

We followed a similar benchmark previously used in SemEval 2025 (Muhammad et al., 2025), which supports multi-label emotion annotation to capture the complexity of emotions expressed in text and has proven suitable in previous emotion detection shared tasks. For emotion perception, we adopt the approach of Mohammad (2022), focusing on how readers interpret the emotions conveyed by the author. This perceptual approach also introduces challenges due to readers' cultural and individual biases (Wakefield, 2021).

The shared task was organized through a

Text	Anger	Fear	Joy	Sadness	Surprise	Hope
Y de un mes a esta parte ha sido Dios servido que he sabido lo que digo.[a]	0	0	0	0	0	0
He sentido tan grande alteración de placer que quisiere como pájaro volar a veros.[b]	0	0	1	0	0	1

[a] *"And for a month now, God has been pleased that I have known what I say."*
[b] *"I have felt such great excitement of pleasure that I would like, like a bird, to fly to see you."*

Table 1: Samples from the Post Scriptum corpus included in the HISEMOTIONS training split.

GitHub repository,[1] which provided participants with all necessary instructions and the publicly available datasets. Prediction submissions and the evaluation campaign were hosted on CodaBench.[2]

The task followed a structured two-phase timeline. In the first phase, participants developed their systems and were provided with training and development sets. In the second phase, participants submitted their final systems' predictions for evaluation using a test set with hidden gold labels, against which their predictions were scored via CodaBench. Participants also had access to baseline experiment results through the GitHub repository. Both the dataset details and the baseline experiments are described in Section 3. During the development phase, each team was allowed up to 100 submissions per day, reduced to 5 per day during the evaluation phase. Each team could select its best submission for the final ranking. Systems were evaluated using Precision, Recall, and F1-score. The primary ranking metric was macro-averaged F1, which is well suited for multi-label classification as it accounts for class imbalance and provides a meaningful summary of performance across all emotion categories.

3 Dataset and Baseline Experiments

3.1 Spanish Historical Texts

The dataset for HISEMOTIONS was built upon the corpus of Spanish correspondence of 16th-17th centuries, digitized and transcribed, and sourced from work by Vaamonde (2015), P.S. Post Scriptum corpus.[3] It includes private letters written in Portugal and Spain during the Early Modern period. Table 1 provides some samples from the P.S. dataset. The corpus consists mainly of previously unpublished correspondence from individuals of diverse social backgrounds, including men and women, adults and children, masters and servants, soldiers, artisans, clergy, and political actors. The P.S. corpus also provides expert Part of Speech (PoS) and syntactic dependency annotations, enabling systematic research on Early Modern epistolary texts.

Category	Train	Dev	Test
Fragments	2.659	425	863
Anger	12	52	53
Fear	303	40	36
Joy	129	29	36
Sadness	1.180	70	163
Surprise	9	7	2
Hope	73	85	58
Total	**4.365**	**708**	**1.211**

Table 2: Class statistics of the HISEMOTIONS dataset.

The HISEMOTIONS corpus is compiled from authentic Spanish letters, originals from P.S. corpus, and emotion annotations. The letters were manually selected based on potential emotion content. The representation of emotion categories is highly imbalanced, as expected. Since the selected subset of the P.S. corpus, which forms part of the HISEMOTIONS corpus, was annotated using a multi-label methodology, achieving a balanced class distribution was inherently challenging, particularly given the authentic nature of the P.S. historical correspondence. Table 2 provides an overview of the class distribution statistics.

Overall, the HISEMOTIONS corpus includes a manually chosen subset from P.S. corpus, totaling 3.947 fragments. The emotion categories are imbalanced across the dataset. Nonetheless, this imbalance may

[1] https://github.com/albinasarymsakova/HISEMOTIONS_2026
[2] https://www.codabench.org/competitions/14140/
[3] http://teitok.clul.ul.pt/postscriptum/es/index.php?action=downloads

prove beneficial as (1) it allows us to evaluate whether existing models can capture contextual relationships and semantic shifting in emotion conveying, and (2) it reflects a realistic scenario for systems development, where balanced data in this specific domain is exceptionally difficult to obtain.

3.2 Annotation Methodology

Annotating a historical epistolary corpus presents a fundamental challenge: establishing a reliable gold-standard tagset. Given that both the authors and their contemporaries are long deceased, it becomes impossible to address the true nature of the emotions expressed. To mitigate this limitation, we used the experimental annotation approach proposed by Sarymsakova et al. (2026b) and formed an interdisciplinary annotation team of two experts, drawing on their expertise in Spanish Philology and Computational Linguistics. We adopt the same annotation guidelines as the aforementioned authors (op. cit.), given their applicability to the same domain of TBED in historical Spanish letters, and because, to the best of our knowledge, this guideline proposal is the unique developed for TBED in this specific textual domain.

As discussed in Section 2, we follow the SemEval 2025 multi-label emotion detection framework (Muhammad et al., 2025), which allows the assignment of multiple emotion labels to a single fragment of a given letter. Here, *fragments* are defined as emotion-bearing units, contiguous spans of text at the clause or sentence level that express a coherent affective state. Fragments derived from the same letter are kept together within a single split (train, test or dev) to prevent data leakage. Available metadata includes letter-level information (when known), such as approximate date, place of origin, author identity, as provided by the P.S. corpus.

Applying the methodology described above, Inter-Annotator Agreement (IAA) was computed for both the dev and test sets of the HISEMOTIONS subset derived from the P.S. corpus, each of which was annotated by Gemma-3-4B-it (Kamath et al., 2025), the baseline model used in this shared task (see Section 3.3), and manually reviewed by two experts. The training set, by contrast, was annotated solely by a baseline model and did not undergo manual expert review, for two main reasons: the high labor cost associated with manually annotating a large number of fragments, and, given this constraint, LLM-based annotation becomes the only viable alternative when human annotation is unavailable, reflecting a realistic scenario in which TBED systems cannot rely on fully manual annotations in practice.

Emotions	Dev	Test
Anger	0.159	0.189
Fear	0.295	0.315
Joy	0.572	0.582
Sadness	0.621	0.641
Surprise	-0.012	-0.002
Hope	0.402	0.422
Average	**0.339**	**0.358**

Table 3: Cohen's κ scores on HISEMOTIONS development and test sets.

We employed the framework proposed by Artstein and Poesio (2008) to compute the weighted Cohen's κ (Cohen, 1960), as shown in Table 3. Sadness reaches substantial agreement (Landis and Koch, 1977; Artstein and Poesio, 2008), indicating that annotators reliably identify this emotion across both splits. Joy achieves moderate agreement, while hope falls between fair and moderate. Fear exhibits fair agreement, and anger only slight agreement. Surprise, by contrast, yields near-zero IAA. These differences are conditioned by the representations of each label in the dev and test splits. Surprise is the lowest-frequency category in the dataset, with only 7 instances in the dev set and 2 in the test set (see Table 2), making its κ score statistically unreliable. Overall, we observed a fair IAA, reflecting the consistency among the dev and test splits. Nonetheless, there remains meaningful disagreement among experts in multi-label emotion classification, which requires further improvement for future editions of HISEMOTIONS task.

3.3 Experiments

We used LLMs to semi-automatically annotate selected fragments for the HISEMOTIONS dataset and to provide F1-score baselines against which participants may assess the performance of their systems. The results are shown in Figure 1. Two common prompting strategies were applied: zero-shot, which provides only the task description and instructions to the model, and few-shot, which

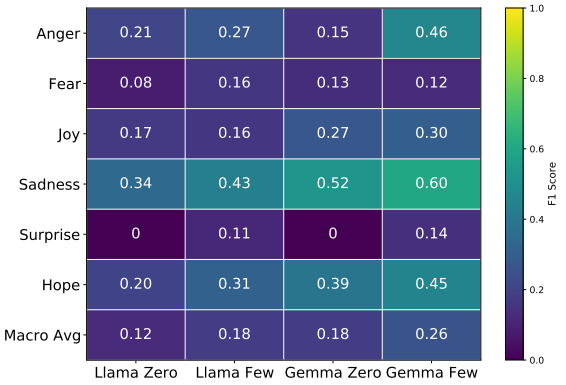

Figure 1: Llama-3.1-8B-Instruct and Gemma-3-4B-it F1 scores obtained with HISEMOTION test set.

supplements these instructions with annotated examples of the task. We selected two LLMs for our experiments: Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and Gemma-3-4B-it (Kamath et al., 2025), given their extensive use in text and emotion classification tasks (Muhammad et al., 2025). The code to obtain these baseline F1 score results on the HISEMOTIONS test set is released.[4]

Figure 1 shows that across all four baseline configurations, Gemma-3 few-shot setting achieves the highest macro-average F1 (0.26). Few-shot prompting consistently outperforms zero-shot across both models. Nonetheless, these data reflect the overall low performance of all baselines. Regarding per-emotion results, sadness emerges as the most reliably detected emotion, with F1 scores ranging from 0.34 (Llama-3.1 Zero-shot) to 0.60 (Gemma-3 Few-shot), while surprise shows low models' performance, with most configurations scoring at or near zero. These results suggest that class imbalance and semantic shifting in the emotion-conveying text of the fragments pose challenges for zero-shot and few-shot LLM-based classification in the historical domain.

4 Participants' Systems

A total of 24 participants registered for the HISEMOTIONS 2026 shared task, of whom 10 teams submitted their results to the leaderboard and 7 teams submitted working notes describing their systems. The official leaderboard is presented in Table 4. The following subsections present a brief summary of each team's contribution based on their submitted working notes, all of which include open-source code to ensure reproducibility.

[4] https://github.com/albinasarymsakova/HISEMOTIONS_2026/tree/main/baselines

Rank	Team	Macro F1
1	franrodrigo	0.58
2	carlosdf	0.57
3	explainators	0.56
4	tbernal	0.47
5	cespinr	0.45
6	ymlopez	0.40
7	luis_lino	0.39
8	davidra02	0.36
9	aldairverdugo	0.32
10	amrtodounam	0.13

Table 4: HISEMOTIONS 2026 Evaluation Phase Leaderboard.

4.1 Franrodrigo Team

Rodrigo-Ginés and Chamorro-Padial (2026) frame their experimental setup as an agentic research pipeline, in which a Claude Opus 4.7 coding agent runs the experimental loop autonomously while human researchers retain veto rights over strategic decisions. Their submitted system is a parameterless label-wise intersection of two prompted LLMs, GPT-5.4 and GPT-5.5: a label is predicted positive if and only if both models agree on it. The model uses greedy decoding. This approach scored 0.649 macro F1 on the development set and 0.578 on the test set. A key aspect of the system is prompt curation, which includes explicit guidance on archaic forms in XVI-century Spanish, abbreviations, and conventional courtesy formulae that should not be interpreted as emotion-bearing expressions. This permits the mitigation of the semantic shifting during the multi-label emotion detection.

4.2 Carlosdf Team

Díez-Fenoy (2026) fine-tunes a RoBERTa encoder pre-trained on historical Spanish (Gutiérrez-Fandiño et al., 2022). To address label imbalance and semantic shifting in the corpus, the system combines no-emotion subsampling, per-class threshold tuning, and synthetic data augmentation targeting the three most underrepresented classes: anger, fear, and hope. Training follows a multilabel stratified 5-fold cross-validation scheme, and the final submission is a validation

F1 weighted ensemble of the three best-performing fold checkpoints. The system achieves a macro F1 of 0.5656 on the test set.

4.3 Explainators Team

Babakova and Stepin (2026) explore four experimental strategies to tackle multilabel emotion detection in early modern Spanish and the semantic shifting presented in the texts: zero-shot prompting, translation-based prompting (Spanish-to-English with GPT-5.2), QLoRA fine-tuning of Llama 3.2, and encoder fine-tuning of BERT, BETO, and mmBERT. Translation-based prompting consistently outperforms zero-shot prompting, with Qwen-3 achieving the best result of 0.611 macro F1 on the development set, while fine-tuned encoders substantially underperform prompted LLMs. The final submission is a weighted majority-voting ensemble of the top-7 models ranked by development macro F1, achieving 0.56 macro F1 on the test set.

4.4 Tbernal Team

Bernal-Beltrán et al. (2026) address multilabel emotion detection and semantic shifting in early modern Spanish correspondence through few-shot learning (FSL) prompting with Qwen3.5-9B, avoiding task-specific fine-tuning. The prompt combines a system instruction that frames the model as an expert in early modern Spanish, five randomly selected training examples formatted as JSON demonstrations, and the target fragment, with the model constrained to generate a binary JSON object for the six emotion categories. FSL prompting substantially outperforms all supervised encoder-based alternatives explored during the development phase, including fine-tuned transformers, ensemble configurations, and label-specific classifiers, with the best fine-tuned encoder reaching only 0.38 macro F1 against 0.47 for the FSL prompting system. The final submission achieves 0.47 macro F1 on the test set.

4.5 Cespinr Team

Espin-Riofrio et al. (2026) build their system around a RoBERTa-based encoder (De la Rosa et al., 2022) that underwent domain-adaptive pretraining on 4.583 letters from the Post Scriptum corpus, comprising 2.368 documents in Spanish and 2.215 in Portuguese. To address the imbalance between human- and LLM-annotated instances in the HISEMOTIONS dataset, the authors annotate additional Post Scriptum fragments using GPT-OSS-120b, adding Portuguese instances to balance the emotion classes in the dataset. Unlike the original training set annotation, the prompt was calibrated against the gold labels of the human-annotated development set, with explicit instructions to focus on emotions expressed by the letter author at the moment of writing, to exclude rhetorical formulas, and to correct for the known biases of the original annotator. Training follows a three-phase pipeline: joint training on released and additional data, refinement exclusively on the gold development set with per-class threshold calibration, and retraining on the combined data with thresholds frozen. Following this training methodology, two models were obtained: one with fewer training epochs and one with more. The final sumbitted system in ensembe, which combines these two models by averaging their predicted probabilities with equal weights, after which per-class thresholds are applied to produce the final predictions. The system achieved a macro F1 of 0.4529 on the test set.

4.6 Ymlopez Team

Martínez-López et al. (2026) propose an ensemble system combining two features: an embedding-based classifier using multilingual sentence encoders (LaBSE, Nomic Embed, and BAAI/bge-m3) paired with regularised logistic regression under a binary relevance scheme with SMOTE oversampling, and a zero-shot LLM prompting branch used locally through Ollama with a Spanish prompt that explicitly instructs the model to detect only the author's emotions, ignore courtesy formulae, and account for historical-semantic context, mitigating semantic shifting. A systematic evaluation of 13 configurations spanning five model families (Qwen, Mistral, Gemma, Falcon, and Llama, 4B–14B parameters) reveals that model family matters more than parameter number, with Qwen2.5-7B consistently outperforming the rest. Notably, the ensemble configurations that combine LLMs with sentence embeddings consistently underperform their single backbones, which the authors attribute to the lexical and orthographic gap between Early Modern Spanish and the modern texts that underlie the embedding spaces. The final system, consisting of zero-shot setting using Qwen2.5-7B,

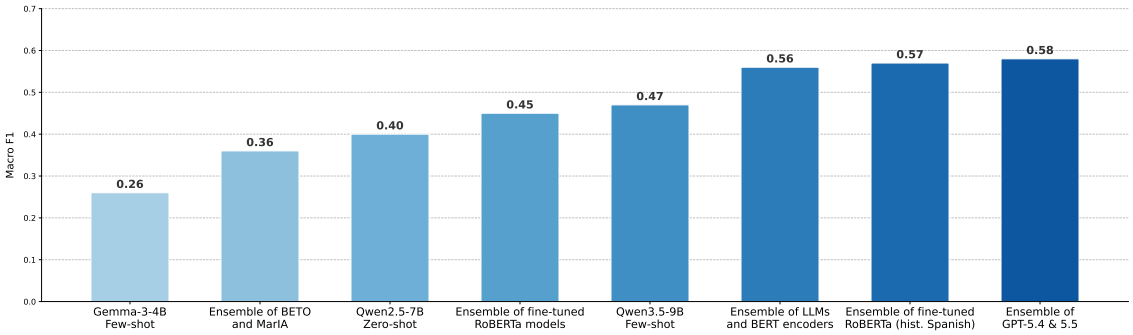

Figure 2: Summary of proposed systems' and the baseline model's macro F1 scores.

achieves a macro F1 of 0.40 on the test set.

4.7 Davidra02 Team

Reyes Alés et al. (2026) explore a range of approaches, from TF-IDF baselines through fine-tuned transformer encoders (BETO, mDeBERTa, and MarIA) to a weighted ensemble. To address class imbalance and semantic shifting in the dataset, they generate synthetic training instances for the five most underrepresented classes using Llama 3.1 8B with few-shot prompting. Class imbalance is further mitigated through per-label positive weights in the BCE loss, and per-class decision thresholds are optimized on the development set via both greedy search and Bayesian joint optimization with Optuna. The final ensemble combines BETO and MarIA through validation F1 weighted probability averaging. The best submission achieves a macro-F1 of 0.36 on the test set.

5 Results and Discussion

The results of the participants' systems reveal a consistent positive trend in the historical Spanish TBED performance. Figure 2 shows that all participant approaches surpassed the Gemma-3-4B Few-shot baseline of 0.26 macro F1. All submitted systems' results provide insights into answers to both research questions, mentioned in the Section 1.

RQ1: How well can existing systems tackle TBED in Early Modern Spanish correspondence?

The participating systems demonstrate that a combination of existing approaches, such as ensemble methods or prompting strategies, can achieve meaningful progress on text-based emotion detection in Early Modern Spanish correspondence. Ensemble strategies that combine fine-tuned models, particularly those incorporating historically adapted Spanish resources, prove especially effective, with the RoBERTa ensemble trained on historical Spanish achieving 0.57, nearly matching the best system. The ensemble setting combining GPT-5.4 and GPT-5.5 only outperforms the historically fine-tuned RoBERTa by 0.01.

RQ2: What improvements are needed to better address this task?

The authors of the submitted systems also offer multiple strategies for enhancing text-based emotion detection in Early Modern Spanish correspondence. For instance, multiple papers addressed the class imbalance or semantic shifting present in the original dataset by (1) generating synthetic instances; (2) translating the original instances to English; (3) annotating more instances from the original Post Scriptum corpus using the gold labels of the human-annotated development HISEMOTIONS set in the LLM prompt, etc. These findings are essential for further development of stronger historical Spanish TBED systems.

6 Error Analysis of the Proposed Systems

We performed a qualitative error analysis at the level of teams' individual predictions to identify the fragments that proved most challenging for the submitted systems. For each participating team, we ranked all test fragments by the Hamming distance between the predicted and gold binary emotion vectors and extracted the top-3 fragments per team exhibiting the largest divergence. The Hamming distance (Hamming, 1950; Li and Jain, 2009) between two strings of equal length is defined as the number of positions at which their corresponding symbols differ.

Team	Fragment	Distance	Gold	Predicted
cespinr	437	4/6	0,0,0,1,0,1	1,1,0,0,0,0
	71[a]	3/6	0,1,0,1,0,0	1,0,0,0,0,0
	113	3/6	1,0,0,0,0,0	0,1,0,1,0,0
franrodrigo	285[b]	4/6	0,1,0,1,0,0	1,0,0,0,0,1
	71[a]	3/6	0,1,0,1,0,0	1,0,0,0,0,0
	304	3/6	0,1,0,1,0,0	0,0,0,0,0,1
explainators	285[b]	4/6	0,1,0,1,0,0	1,0,0,0,0,1
	526	4/6	1,0,1,0,0,0	0,1,0,1,0,0
	241	3/6	0,0,0,1,0,0	1,0,1,0,0,0
ymlopez	439	4/6	1,0,0,1,0,0	0,1,0,0,1,0
	72	3/6	0,0,0,0,1,0	0,0,0,1,0,0
	255	3/6	0,0,0,1,1,0	0,1,0,0,0,0
carlosdf	103	3/6	1,0,0,0,0,0	0,1,0,1,0,0
	282[c]	3/6	0,0,0,1,0,0	0,0,1,0,0,1
	415	3/6	1,0,0,0,0,1	0,0,1,1,0,1
tbernal	71[a]	3/6	0,1,0,1,0,0	1,0,0,0,0,0
	282[c]	3/6	0,0,0,1,0,0	0,0,1,0,0,1
	285[b]	3/6	0,1,0,1,0,0	1,0,0,0,0,0
davidra02	19	4/6	0,0,1,0,0,0	1,1,0,1,0,0
	110	4/6	1,0,0,1,0,0	0,1,1,1,0,1
	282[c]	4/6	0,0,0,1,0,0	1,0,1,0,0,1

[a] Line 71: *"Y veo que el interés le ciega y piensa que no hay verdad." / "And I see that self-interest blinds him and he thinks there is no truth."*

[b] Line 285: *"Y tengan entendido que los que me dan guerra de dentro y fuera y sepan que tengo de salir vencedor con vuestra firmeza o con acabar con todo muy largo los he escrito advirtiéndoos de el orden que habéis de guardar para que nuestro estado no se dilate que es lo que pretenden nuestros enemigos." / "And let it be understood by those who wage war on me from within and without and let them know that I am to come out victorious, whether through your steadfastness or by bringing an end to everything at great length I have written to you advising you of the order you must keep so that our state is not delayed which is what our enemies intend."*

[c] Line 282: *"Bien entiendo que no es por culpa vuestra sino por falta de ventura mía que el ser quien sois me da causa de tener esta confianza y me asegura de vuestra firmeza y valor." / "I well understand that it is not through your fault but through my own lack of fortune for your being who you are gives me cause to hold this confidence and assures me of your steadfastness and valor."*

Table 5: Results of the prediction analysis for the systems submitted to HISEMOTIONS. The "Distance" column reports the Hamming distance between the gold and predicted binary labels. Each binary label indicates the presence (1) or absence (0) of an emotion, given in the following order: anger, fear, joy, sadness, surprise, and hope. The superscripts [a], [b], and [c] identify the Spanish text (and its English translation) corresponding to each line.

In our multi-label setting, where each fragment is encoded as a six-dimensional binary vector over the emotions (anger, fear, joy, sadness, surprise, and hope), this amounts to the number of emotion labels that are either wrongly predicted or missed, and thus provides an intuitive per-instance error measure that penalizes both false positives and false negatives. Table 5 reports, for each of the seven systems, the top-3 fragments with the highest Hamming distance between predictions and gold, together with the corresponding gold and predicted label vectors.

The colour coding in Table 5 reveals that three fragments appear consistently across the majority of the systems, namely the instances located at lines 71, 285, and 282 of the test set. The fragment 71 (green) appears among the most difficult cases for `cespinr`, `franrodrigo`, and `tbernal`; line 285 (orange) for `franrodrigo`, `explainators`, and `tbernal`; and line 282 (yellow) for `carlosdf`, `tbernal`, and `davidra02`. This pattern, common to systems built on different architectures and strategies, suggests that the difficulty is intrinsic to the texts and emotion co-occurrences rather than to the modeling choice.

Table 6 resumes the misclassified fragments, each of which appears across three distinct systems, and reports their gold labels. Two of them, fragments 71 and 285, present two emotion co-occurrence (fear+sadness).

Fragment	Systems	Emotions
71	3	fear+sadness
285	3	fear+sadness
282	3	sadness

Table 6: Lines showing the greatest disagreement between the gold and predicted labels across all submitted systems.

The fragments 282 and 285, as the text transcription in Table 5 indicates, are extensive. These aspects may influence the misclassification.

Regarding the predictions of each fragment, the 71 is classified as anger by all three systems; the 285 is likewise predicted as anger by all three, and additionally as hope by two of them; and the 282 is predicted by most systems as joy and hope. Comparing these predictions with the actual textual cues, it becomes evident that the systems still struggle to account for the semantic shifting that Early Modern Spanish lexical cues exhibit in context. This aspect is clearly observed in fragment 282, which is interpreted by the systems as positive due to lexicons such as *confianza* (confidence), *firmeza y valor* (steadfastness and valor), while the emotional core of the fragment is anchored by *Bien entiendo que no es por culpa vuestra sino por falta de ventura mía* (I well understand that it is not through your fault but through my own lack of fortune). Moreover, unlike human annotators, the systems do not appear to rely on contextual relations between fragments: they infer the closest emotion within a fragment in isolation, without considering the preceding or following fragments of the same letter.

Taken together, these observations point to two directions for improvement. First, systems still leave considerable room for improvement in correctly interpreting Early Modern Spanish lexical cues in emotion classification. Second, the systems' tendency to classify each fragment in isolation should be addressed, so that systems can draw on the broader context to predict the emotion correctly.

7 Conclusion

The findings presented in this first edition of HISEMOTIONS at IberLEF 2026 indicate that our hypothesis, which arose in Section 1 and holds that existing systems require adjustments to effectively tackle multi-label TBED in Spanish historical correspondence, receives strong empirical support, providing a solid foundation for further research in this direction. The participants' submissions have shed light on how the TBED in Spanish historical correspondence must be addressed and what insights must be taken into account to improve it.

The most popular choice among submitted systems at HISEMOTIONS 2026 was the ensemble method, the best-performing of which combined two LLMs (GPT-5.4 and 5.5), and the second-best-performing setting was built upon RoBERTa pre-trained on historical Spanish. The participants' results outperformed the established HISEMOTIONS baseline, indicating substantial progress in this first edition of the shared task.

The error analysis of the proposed systems suggests that, on one hand, there remains substantial scope for systems to improve in accurately interpreting the lexical traits of Early Modern Spanish when predicting emotions. On the other hand, the isolation of each fragment for emotion detection currently used in the proposed settings needs to be addressed to enable systems to leverage surrounding context to obtain accurate emotional predictions.

Further editions of HISEMOTIONS aim to be organized to tackle other issues related to TBED in historical Spanish, such as the extraction of emotion-bearing units from texts and their application to enhance emotion classification, the knowledge-editing and in-context learning challenges that this domain presents, and human label variation in historical Spanish TBED. In this context, we aim to collect authentic, expert-annotated textual data, if possible, to provide a high-quality dataset for future editions.

Acknowledgments

This study is part of the research project "History and emotions: Characterization and evaluation of computational linguistics and language models in emotion mining from historical sources" of Momentum CSIC Programme (ref. MMT24–IEGPS–01) within the framework of the Generación D initiative, promoted by Red.es, an entity attached to the Ministry for Digital Transformation and the Civil Service, to attract

and retain talent through scholarships and training contracts funded by the Recovery, Transformation and Resilience Plan supported by the Next Generation EU funds. This work is also partially supported by funding from the Spanish National Research Council (CSIC), Special Intramural Project No. 2024ICT028. This work is also funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project ROMANET (CERV-2024-CHAR-LITI-101215052), funded by the European Union under the Citizens, Equality, Rights and Values programme, Project HEART-NLP-UJA (PID2024-156263OB-C21) and project VERITAS-H (AIA2025-163322-C64) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, Project GALENO-IA (DGP_PIDI_2024_00852) funded by the European Union under the Smart Specialization Strategy Program for Sustainability in Andalusia (S4Andalucía) 2021–2027.

References

Artstein, R. and M. Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Babakova, L. and I. Stepin. 2026. Large language model prompting and fine-tuning for early modern Spanish corpus text-based emotion detection. 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 Workshop Proceedings, León, Spain. CEUR-WS.org.

Bernal-Beltrán, T., R. Pan, J. Gómez-Navalón, et al., J. A. García-Díaz, G. Beydoun, R. Valencia-García. 2026. UmuTeam at HISEMOTIONS 2026: Few-shot prompting with large language models for emotion detection in early modern Spanish letters. 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 Workshop Proceedings, León, Spain. CEUR-WS.org.

Bonet-Jover, A., J. Á. González-Barba, and L. Chiruzzo. 2026. 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.*

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

De la Rosa, J., E. G. Ponferrada, M. Romero, et al., P. Villegas, P. González de Prado Salas, M. Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Díez-Fenoy, C. 2026. Emotion detection in early modern Spanish correspondence: A multilabel transformer approach. 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 Workshop Proceedings, León, Spain. CEUR-WS.org.

Espin-Riofrio, C., O. León Granizo, A. Montejo-Ráez, et al., J. Burgos-Bohorquez, L. Lino-Cevallos. 2026. SINAI at HISEMOTIONS in IberLEF 2026: Three-phase fine-tuning of a domain-adaptive BERT for historical emotion detection. 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 Workshop Proceedings, León, Spain. CEUR-WS.org.

Gatti, F. and J. Huesler. 2025. Text analysis methods for historical letters, the case of michelangelo buonarrotti. EHES Working Paper 279, s.l.

Grattafiori, A., A. Dubey, A. Jauhri, et al., A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, others. 2024. The llama 3 herd of models.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, et al., J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68(0):39–60.

Hamming, R. W. 1950. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160.

Hu, H., P. Amaral, and S. Kübler. 2022. Word embeddings and semantic shifts in historical spanish: Methodological considerations. *Digital Scholarship in the Humanities*, 37(2):441–461, 06.

Kamath, G. T. A., J. Ferret, S. Pathak, et al. 2025. Gemma 3 technical report. *ArXiv*, abs/2503.19786.

Kim, E. and R. Klinger. 2018. A survey on sentiment and emotion analysis for computational literary studies. *arXiv preprint arXiv:1808.03137*.

Landis, J. R. and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

Leemans, I., E. Maks, J. {van der Zwaan}, et al., H. Kuijpers, K. Steenbergh. 2017. Mining embodied emotions: A comparative analysis of bodily emotion expressions in dutch theatre texts 1600-1800'. *Digital Humanities Quarterly*, 11(4).

Li, S. Z. and A. Jain, editors, 2009. *Hamming Distance*, pages 668–668. Springer US.

Maks, I. and P. Vossen. 2011. A verb lexicon model for deep sentiment analysis and opinion mining applications. In A. Balahur, E. Boldrini, A. Montoyo, and P. Martinez-Barco, editors, *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 10–18, Portland, Oregon, June. Association for Computational Linguistics.

Martínez-López, Y., Y. Jauriga, M. Guerra Saborit, et al., J. Madera, L. Quintero Domínguez, A. Rodríguez-González, C. de Castro Lozano, J. M. Ramirez Uceda. 2026. UC-UCO-Plenitas team at exploring the historical text-based emotion detection in early modern Spanish correspondence using large language model and embeddings. 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 Workshop Proceedings, León, Spain. CEUR-WS.org.

Mohammad, S. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In K. Zervanou and P. Lendvai, editors, *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA, June. Association for Computational Linguistics.

Mohammad, S. M. 2022. Best practices in the creation and use of emotion lexicons. *arXiv preprint arXiv:2210.07206*.

Montes, T., L. Manrique-Gómez, and R. Manrique. 2024. Historical ink: Semantic shift detection for 19th century Spanish. In N. Tahmasebi, S. Montariol, A. Kutuzov, D. Alfter, F. Periti, P. Cassotti, and N. Huebscher, editors, *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 29–41, Bangkok, Thailand, August. Association for Computational Linguistics.

Muhammad, S. H., N. Ousidhoum, I. Abdulmumin, et al. 2025. SemEval-2025 task 11: Bridging the gap in text-based emotion detection. In S. Rosenthal, A. Rosá, D. Ghosh, and M. Zampieri, editors, *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2558–2569, Vienna, Austria, July. Association for Computational Linguistics.

Periti, F. and N. Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4262–4282, Mexico City, Mexico, June. Association for Computational Linguistics.

Plaza del Arco, F. M., C. Strapparava, L. A. Urena Lopez, et al., M. Martin. 2020. EmoEvent: A multilingual emotion corpus based on different events. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1492–1498, Marseille, France, May. European Language Resources Association.

Reagan, A. J., L. Mitchell, D. Kiley, et al., C. M. Danforth, P. S. Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):31.

Reyes Alés, D., A. I. Arteaga, S. Peñafiel Calderón, et al., A. Verdugo Valerio. 2026. Multi-label emotion detection in early modern Spanish correspondence: System description for HISEMOTIONS @ IberLEF 2026. 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 Workshop Proceedings, León, Spain. CEUR-WS.org.

Rodrigo-Ginés, F.-J. and J. Chamorro-Padial. 2026. An agentic multi-LLM intersection approach to historical Spanish emotion detection. 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 Workshop Proceedings, León, Spain. CEUR-WS.org.

Santangelo, P., 2023. *74 The role and symbolic meanings of emotion in literary language*, pages 1558–1590. De Gruyter Mouton, Berlin, Boston.

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

Sarymsakova, A., E. Martínez-Cámara, P. Martín-Rodilla, et al., L. A. Ureña-López. 2026b. Emotion detection in textual records of affective correspondence in spanish. *Procesamiento del lenguaje natural*, 77.

Schmidt, T., K. Dennerlein, and C. Wolff. 2021. Towards a Corpus of Historical German Plays with Emotion Annotations. In D. Gromann, G. Sérasset, T. Declerck, J. P. McCrae, J. Gracia, J. Bosque-Gil, F. Bobillo, and B. Heinisch, editors, *3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *Open Access Series in Informatics (OASIcs)*, pages 9:1–9:11, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Sprugnoli, R., S. Tonelli, A. Marchetti, et al., G. Moretti. 2016. Towards sentiment analysis for historical texts. *Digit. Scholarsh. Humanit.*, 31:762–772.

Turunen, R., I. Taskinen, L. Uusitalo, et al., V. Kivimäki. 2022. Mining emotions from the finnish war letter collection, 1939–1944. In K. Berglund, M. {La Mela}, and I. Zwart, editors, *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)*, CEUR Workshop Proceedings, pages 135–144. CEUR-WS. Jufoid=53269; Digital Humanities in the Nordic and Baltic Countries Conference ; Conference date: 15-03-2022 Through 18-03-2022.

Vaamonde, G. 2015. P. s. post scriptum: Dos corpus diacrónicos de escritura cotidiana. *Proces. del Leng. Natural*, 55:57–64.

Valitutti, A., C. Strapparava, and O. Stock. 2004. Developing affective lexical resources. *PsychNology J.*, 2:61–83.

Wakefield, J. 2021. Ai emotion-detection software tested on uyghurs. *BBC News*, 26.

Wemmer, E., S. Labat, and R. Klinger. 2024. EmoProgress: Cumulated emotion progression analysis in dreams and customer

service dialogues. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5660–5677, Torino, Italia, May. ELRA and ICCL.

Yang, K., S. Ji, T. Zhang, et al., Q. Xie, Z. Kuang, S. Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore, December. Association for Computational Linguistics.

Yavuz, M. C. 2020. Analyses of character emotions in dramatic works by using EmoLex unigrams. In J. Monti, F. Dell'Orletta, and F. Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 472–477, Bologna, Italy, March. CEUR Workshop Proceedings.

Zhang, Y., X. Su, J. Wu, et al., J. Yang, H. Fan, X. Zheng. 2023. Emoknow: emotion- and knowledge-oriented model for covid-19 fake news detection. In X. Yang, H. Suhartanto, G. Wang, B. Wang, J. Jiang, B. Li, H. Zhu, and N. Cui, editors, *Advanced Data Mining and Applications*, Lecture Notes in Computer Science, pages 352–367, United States. Springer, Springer Nature. 19th International Conference on Advanced Data Mining and Applications, ADMA 2023 ; Conference date: 21-08-2023 Through 23-08-2023.