SINAI at HISEMOTIONS in IberLEF 2026: Three-Phase Fine-Tuning of a Domain-Adaptive BERT for Historical Emotion Detection

César Espin-Riofrio[1,*], Oscar León Granizo[1], Arturo Montejo-Ráez[2], Jhosue Burgos-Bohorquez[1] and Luis Lino-Cevallos[1]

[1]University of Guayaquil, Delta Av. s/n, Guayaquil, 090510, Ecuador
[2]University of Jaén, Las Lagunillas s/n, Jaén, 23071, Spain

Abstract

We present the SINAI system for the HISEMOTIONS 2026 shared task on multi-label emotion detection in Early Modern Spanish epistolary texts. A central challenge of this task is that the official training set carries LLM-generated labels with severe systematic biases, while evaluation is performed against human-annotated gold labels. Our approach addresses this mismatch through three main contributions: a domain-adaptive pretrained model, `bert_siglo_oro`, trained on the same *Post Scriptum* corpus as the task data; a data augmentation strategy based on re-annotating additional authentic historical fragments using prompts calibrated against the human-annotated development set; and a three-phase fine-tuning pipeline that progressively leverages both the noisy training data and the reliable human-annotated development set. Two model variants with complementary inductive biases are combined into a final ensemble. Our system achieved a Macro-F1 of 0.4529 on the official test set, ranking 5th out of 10 participating teams and nearly doubling the best workshop baseline of 0.26.

Keywords

emotion detection, historical Spanish, domain-adaptive pretraining, multi-label classification

1. Introduction

Detecting emotions in Early Modern Spanish correspondence is far from a straightforward classification problem. Texts from the 16th and 17th centuries exhibit archaic vocabulary, inconsistent orthography, and rhetorical conventions that differ substantially from modern usage, making it difficult to transfer the representations learned by contemporary language models to this domain without adaptation.

HISEMOTIONS 2026 [1], organized within the IberLEF 2026 framework [2], proposes a multi-label emotion detection task over epistolary texts drawn from the *Post Scriptum* corpus [3], with six target emotion categories: anger, fear, joy, sadness, surprise, and hope. A particularly difficult aspect of the task is that training labels were generated by a large language model while development and test labels were produced by human annotators, introducing systematic biases that any competitive system must account for.

We approached this problem by building on `bert_siglo_oro`, a model with domain-adaptive pretraining on historical Spanish epistolary text, and designing a three-phase fine-tuning pipeline that progressively leverages both the noisy training data and the reliable gold development annotations. Our best submission achieved a Macro-F1 of 0.4529 on the official test set, surpassing the workshop baseline of 0.26 by a substantial margin.

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
cesar.espinr@ug.edu.ec (C. Espin-Riofrio); oscar.leong@ug.edu.ec (O. L. Granizo); amontejo@ujaen.es (A. Montejo-Ráez); jhosue.burgosboh@ug.edu.ec (J. Burgos-Bohorquez); luis.linocev@ug.edu.ec (L. Lino-Cevallos)
0000-0001-8864-756X (C. Espin-Riofrio); 0000-0002-1118-2108 (O. L. Granizo); 0000-0002-8643-2714 (A. Montejo-Ráez); 0009-0007-2758-4869 (J. Burgos-Bohorquez); 0009-0003-4671-2687 (L. Lino-Cevallos)

2. Related Work

Emotion detection in historical texts has received considerably less attention than its modern counterpart, partly due to the scarcity of annotated corpora and partly because the semantic shift that words undergo over centuries makes it hard to rely on contemporary lexical resources. Early work in this area focused on literary genres such as plays and novels [4, 5], while epistolary corpora have only recently begun to attract systematic computational treatment [6, 7].

The broader field of multi-label emotion classification has been shaped largely by shared tasks over modern social media text, most notably the SemEval-2025 framework [8]. Transferring these methods to historical domains, however, requires confronting the substantial lexical and stylistic distance between centuries-old correspondence and the text on which contemporary models are pretrained.

Domain-adaptive pretraining has emerged as an effective strategy to bridge this gap. Rather than fine-tuning a general-purpose model directly, continued pretraining on in-domain text allows the model to internalize domain-specific vocabulary and syntactic patterns before any supervised signal is introduced [9]. In our system, we rely on `bert_siglo_oro`, a model that underwent this process on historical Spanish text, which proved critical to obtaining competitive results on this task.

Finally, training on automatically generated labels and refining on smaller but cleaner human-annotated data is a well-established paradigm in low-resource settings [10]. Our three-phase pipeline draws on this intuition while addressing the specific distributional mismatch introduced by LLM-generated annotations.

3. System Description

Figure 1 provides an overview of the full system pipeline, which we describe in detail in the following subsections. The code for this system is publicly available at https://github.com/cespinr/sinai-hisemotions-2026.

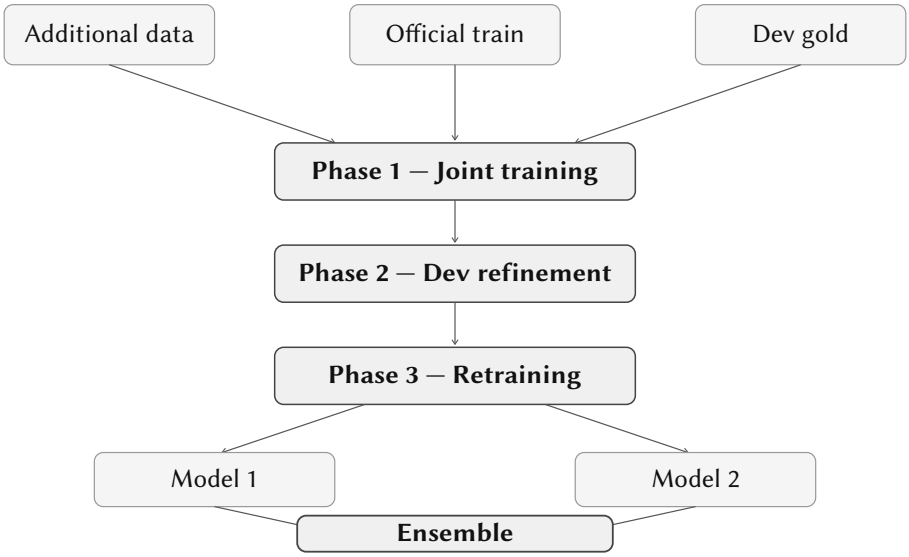

Figure 1: Overview of the three-phase training pipeline and ensemble.

3.1. Backbone Model

Our system is built on `bert_siglo_oro`, a model obtained by applying domain-adaptive pretraining on `bertin-project/bertin-roberta-base-spanish`, a RoBERTa-based model pretrained on modern Spanish. The adaptation corpus consists of 4,583 letters from the *Post Scriptum* collection spanning the 16th through 19th centuries, comprising 2,368 documents in Spanish and 2,215 in Portuguese,

totaling approximately 1.65 million tokens. Portuguese letters were deliberately included because they were produced in the same epistolary register and historical period as the Spanish ones, sharing clause structure, rhetorical conventions, and much of the courtly vocabulary that characterizes Early Modern correspondence. After sentence segmentation, 63,331 sentences were retained as training units for the masked language modeling objective. The adaptation reduced perplexity on historical text by 82.4% relative to the base model, confirming that `bert_siglo_oro` had internalized the vocabulary and orthographic conventions of Early Modern correspondence. Since the HISEMOTIONS dataset derives from this same corpus, `bert_siglo_oro` is already familiar with the linguistic register, orthographic conventions, and rhetorical patterns that characterize this genre, making it the component of our system that most closely captures the stylistic properties of Early Modern epistolary writing.

3.2. Classifier Architecture

On top of `bert_siglo_oro` we place a classification head that aggregates token representations from the last hidden state via mean pooling, projects them to 256 units with ReLU activation and dropout of 0.3, and produces 6 independent sigmoid outputs, one per emotion. We train with Focal Loss ($\gamma = 2.0$) and label smoothing of 0.05, using AdamW with a cosine warmup scheduler and a maximum sequence length of 128 tokens.

3.3. Training Data

The official dataset comprises 2,659 training fragments with LLM-generated labels, 425 development fragments with human-annotated gold labels, and 863 test fragments used for final evaluation. No additional data splits were performed; the official train, development, and test partitions were used as provided by the task organizers.

A central difficulty of this task is that the official training set carries LLM-generated labels with systematic biases, as shown in Table 1. Any model trained naively on this data learns the annotation patterns of the LLM rather than the emotional judgments of human experts.

Table 1
Label distribution bias in the official training set relative to the human-annotated development set.

Emotion	Train (%)	Dev gold (%)	Bias
anger	0.5	12.2	≈30× under
fear	2.1	9.4	≈4× under
joy	1.5	6.8	≈4× under
sadness	47.0	16.5	≈2.5× over
surprise	0.3	1.6	≈5× under
hope	1.0	20.0	≈7× under

To extend the training data without introducing artificially generated text, we selected additional fragments from the broader *Post Scriptum* corpus that had not been used in any of the official splits, and annotated them automatically using GPT-OSS-120b [11], accessed through the Helmholtz Blablador platform [12]. Unlike the original training set annotation, the prompt was calibrated against the human-annotated development set, using its gold labels as the reference criterion, 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. From the resulting set, we retained the 2,344 fragments that received at least one positive emotion label, achieving a label distribution substantially closer to the gold development set than the official training data, with anger, surprise, and hope represented at 15%, 2.2%, and 47% respectively compared to near-zero rates in the official training set. These additional fragments, being authentic historical text from the same corpus as the shared task, carry both the domain characteristics and a more balanced class distribution than the official training data.

3.4. Three-Phase Training Pipeline

Our training strategy proceeds in three phases designed to extract reliable signal from noisy data while progressively incorporating human-annotated supervision.

Phase 1. The model is trained jointly on the additional annotated fragments and the official training set. To partially compensate for the distributional mismatch, we subsample 70–80% of sadness-positive and fear-positive examples from the official training data, and repeat the human-annotated development set multiple times within the combined training data to amplify its signal.

Phase 2. The model is refined exclusively on the development set, where per-class decision thresholds are calibrated and fixed. This phase constitutes the primary source of reliable supervision, as the development set carries human-annotated gold labels.

Phase 3. The model is retrained on the combined train and development data using the thresholds from Phase 2 without readjustment, ensuring that the calibration derived from human annotations is preserved in the final model.

3.5. Ensemble

Our final submission combines two independently trained model variants, referred to as Model 1 and Model 2, that share the same architecture and follow the three-phase training pipeline described above, but differ in the hyperparameters shown in Table 2.

Model 1 uses fewer training epochs and a lower development set repetition factor, reducing exposure to the gold development signal and limiting overfitting to its distribution. This produces stronger generalization on frequent emotion classes such as fear, joy, sadness, and hope.

Model 2 uses more training epochs, a higher development set repetition factor, and a more aggressive sadness subsampling rate, amplifying the influence of the gold development signal and improving recall on underrepresented emotions, most notably surprise.

Ensemble. The two models are combined by averaging their predicted probabilities with equal weights, after which per-class thresholds are applied to produce the final binary predictions. By combining models with complementary strengths, the ensemble captures both the precision of Model 1 on frequent emotion classes and the recall of Model 2 on underrepresented ones, yielding a more balanced per-class performance than either model achieves individually.

Table 2
Hyperparameter differences between the two ensemble components.

Hyperparameter	Model 1	Model 2
Learning rate	5e-6	3e-6
Training epochs	10	12
Dev repetition factor	3	5
Sadness subsampling	70%	80%

4. Experiments and Results

Table 3 compares our submitted ensemble against the strongest organizer baseline (Gemma-3-4B-it with few-shot prompting) and the two individual model variants that compose it, Model 1 and Model 2, trained with different hyperparameter configurations as described in Section 3.

Our submitted system achieved a Macro-F1 of 0.4529, placing 5th out of 10 participating teams and surpassing the best workshop baseline by 0.19 points. The ensemble improved over both individual models by combining their complementary strengths: Model 1 performed better on fear, joy, sadness, and hope, while Model 2 contributed the sensitivity to surprise that had been largely absent in Model 1, achieving an F1 of 0.182 for that class. The ensemble consolidated both contributions, improving anger substantially (+0.084 over Model 1) while maintaining competitive scores across the remaining classes.

Table 3
Per-class F1 scores on the official test set.

System	Ang	Fear	Joy	Sad	Sur	Hope	Mac-F1
Gemma-3 Few-shot	0.46	0.12	0.30	0.60	0.14	0.45	0.26
Model 1	0.369	0.356	0.645	0.672	0.050	0.510	0.4476
Model 2	0.393	0.308	0.607	0.602	0.182	0.398	0.4147
Ensemble	**0.453**	**0.351**	**0.611**	**0.631**	**0.182**	**0.490**	**0.4529**

The most persistent challenge was surprise, which has only 7 positive examples in the development set and is nearly absent in the official training data. Although the additional fragments annotated via Blablador provided enough signal for the model to detect some cases, precision remained very low at 0.111, reflecting the difficulty of distinguishing genuine surprise from superficially similar expressions in formal epistolary writing. Fear also remained below 0.40 throughout all configurations, likely because its expression in this corpus tends to be indirect and formulaic rather than lexically explicit.

In contrast, joy and sadness, which are more directly expressed in this genre, reached F1 scores above 0.60, and anger showed a substantial improvement over the baseline thanks to the bias-correcting annotation strategy applied to the additional training data.

A recurring observation throughout our experiments was that threshold calibration on the development set did not always transfer cleanly to unseen data, which we attribute to two compounding factors: the development set was used both for threshold calibration and as additional training signal in Phases 2 and 3, concentrating the available human-annotated signal in a single small split; and the official training data, despite our mitigation strategies, still dominated the learned representations given its volume relative to the development set. Addressing this structural imbalance remains the main open challenge for future work on this task.

The top-ranked system in the shared task achieved a Macro-F1 of 0.58 [1], leaving a gap of approximately 0.13 points that we attribute primarily to the remaining weakness in fear and surprise.

5. Discussion

Table 4 presents the per-class confusion matrix of the submitted ensemble on the official test set, derived from the Codabench evaluation results.

Table 4
Per-class confusion matrix on the official test set.

Emotion	TP	FP	FN	TN
anger	22	41	12	788
fear	11	33	9	810
joy	28	28	7	800
sadness	68	56	23	716
surprise	2	17	2	842
hope	82	62	108	611

The results reveal distinct error patterns across emotion classes. Anger and fear show a high number of false positives relative to true positives, suggesting that the system tends to over-predict these classes, likely as a consequence of the augmentation strategy that introduced more positive examples of minority emotions during training. Joy and sadness present the most balanced profiles, with false positives and false negatives at comparable levels, consistent with their stronger lexical signals in this genre.

Surprise stands out as the most problematic class: with only 2 true positives against 17 false positives, the system detects the class at very low precision, confirming that the model learned a noisy proxy for surprise rather than its genuine expression in Early Modern correspondence. A qualitative inspection of the development set reveals why: fragments correctly identified as surprise tend to contain explicit

markers of unexpected events, such as *"Que yo le declararé la verdad en viniendo"*, while false positives often involve emotionally charged language associated with hope or relief that the model conflates with surprise, as in *"sacarme del todo de este infierno pues me habéis dado ya luz de esta esperanza"*.

Hope presents the highest false negative count (108), indicating that the system misses a large proportion of hope-bearing fragments. This is consistent with the severe underrepresentation of hope in the LLM-annotated training data and suggests that even targeted augmentation was insufficient to fully recover this class.

6. Limitations

Threshold calibration was performed directly on the development set, which was also used as training signal in Phases 2 and 3. Cross-validated calibration or a dedicated held-out calibration split would have provided a less biased estimate of threshold generalization; however, with only 425 human-annotated fragments available, further subdividing the development set would have significantly reduced the reliable supervision signal on which the entire pipeline depends. This constraint is inherent to the task setup, where the scarcity of gold-annotated data limits the options for rigorous threshold validation.

7. Conclusions

We have presented the SINAI system for the HISEMOTIONS 2026 shared task, combining domain-adaptive pretraining, bias-correcting data augmentation, and a three-phase fine-tuning pipeline to address multi-label emotion detection in Early Modern Spanish correspondence. The central finding is that grounding both the backbone model and the re-annotation strategy in human-annotated gold labels yields substantial and consistent gains over prompting-based approaches: our system achieved a Macro-F1 of 0.4529, ranking 5th out of 10 participating teams and nearly doubling the best workshop baseline of 0.26.

Per-class results reveal a clear divide between emotions with direct lexical expression in this genre, such as joy and sadness, which both exceeded F1 = 0.60, and emotions whose expression is indirect or formulaic, namely fear and surprise, which remained the main bottlenecks despite targeted augmentation strategies. Of all system components, `bert_siglo_oro` contributes most directly to handling the stylistic register of Early Modern correspondence, as its domain-adaptive pretraining implicitly encodes the archaic vocabulary, orthographic conventions, and rhetorical patterns of the corpus.

As future work, we plan to investigate the quality and reliability of the automatically annotated additional fragments, for instance through human validation of a sample or inter-annotator agreement studies, as the annotation quality of this augmented data directly conditions the gains observed in minority emotion classes. We also plan to explore the explicit modeling of epistolary style features as a complementary signal for detecting indirectly expressed emotions, and the modernization of historical texts as a preprocessing step to improve transferability of contemporary emotion classifiers to this domain.

Acknowledgments

This work is 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. This research was also supported by Project FCI-036-2023 at the Universidad de Guayaquil, Ecuador.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-OSS-120b (accessed via the Helmholtz Blablador platform) in order to: automatically annotate additional training fragments from the *Post Scriptum* corpus. Further, the author(s) used Claude (Anthropic) in order to: review and edit this paper. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] A. Sarymsakova, P. Martín-Rodilla, E. Martínez-Cámara, L. A. Ureña-López, Overview of HISE-MOTIONS 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] G. Vaamonde, P. S. Post Scriptum. Dos corpus diacrónicos de escritura cotidiana, Procesamiento del Lenguaje Natural 55 (2015) 57–64.

[4] T. Schmidt, K. Dennerlein, C. Wolff, Towards a Corpus of Historical German Plays with Emotion Annotations, in: 3rd Conference on Language, Data and Knowledge (LDK 2021), Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, pp. 9:1–9:11.

[5] 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.

[6] F. Gatti, J. Huesler, Text Analysis Methods for Historical Letters: The Case of Michelangelo Buonarroti, in: EHES Working Paper No. 279, European Historical Economics Society, 2025.

[7] R. Turunen, I. Taskinen, L. Uusitalo, V. Kivimäki, Mining Emotions from the Finnish War Letter Collection, 1939–1944, in: Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), 2022, pp. 135–144.

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

[9] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't Stop Pretraining: Adapt Language Models to Domains and Tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 8342–8360.

[10] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, R. Klinger, A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021, pp. 2545–2568.

[11] Helmholtz Association, GPT-OSS-120b: Open-Source Large Language Model, https://helmholtz-blablador.fz-juelich.de, 2024. Accessed: 2026.

[12] Helmholtz Association, Helmholtz Blablador: An Inference Server for Scientific Large Language Models, https://helmholtz-blablador.fz-juelich.de, 2024. Accessed: 2026.