Multi-Label Emotion Detection in Early Modern Spanish Correspondence: System Description for HISEMOTIONS @ IberLEF 2026

David Reyes Alés[1,*], Alejandro Iabin Arteaga[1], Santiago Peñafiel Calderón[1] and Aldair Verdugo Valerio[1]

[1]*Universidad Carlos III de Madrid, Leganés, Spain*

Abstract

This paper describes our system submitted to the HISEMOTIONS shared task at IberLEF 2026, which focuses on multi-label emotion detection in Early Modern Spanish epistolary texts (16^{th}–17^{th} centuries). We explore a range of approaches: TF-IDF with classical classifiers as baselines, fine-tuning of pre-trained Spanish and multilingual Transformer models (BETO, mDeBERTa, MarIA), LLM-based data augmentation to address severe class imbalance, and weighted model ensembling. Our best single-model result is achieved by BETO (*bert-base-spanish-wwm-cased*) fine-tuned with weighted binary cross-entropy and per-label threshold optimisation, reaching a Macro F1 of **0.3792** on the development set and **0.36** on the official Codabench hidden test set. We report experiments with three Transformer encoders, Focal Loss, and Bayesian joint threshold search via Optuna, and analyse the challenges posed by extreme label imbalance, near-zero-shot classes, and domain specificity of historical Spanish.

Keywords

multi-label classification, emotion detection, historical Spanish, transformer models, data augmentation, IberLEF 2026

1. Introduction

Emotion detection in natural language is a well-studied problem in modern text [1], yet applying these techniques to *historical* documents introduces challenges that the field has only begun to address: archaic vocabulary, non-standard orthography, unconventional punctuation, domain-specific abbreviations (*VM* for *vuestra merced*), and a near-total absence of annotated resources.

HISEMOTIONS @ IberLEF 2026 [2, 3] addresses precisely this gap. The task requires automatic identification of six emotions—*anger, fear, joy, sadness, surprise*, and *hope*—in a corpus of personal letters from 16^{th}–17^{th} century Spain. The evaluation metric is Macro F1 over the six binary labels, framing the problem as multi-label classification [4].

Three aspects make this task particularly challenging. First, the training set suffers from **severe class imbalance**: *sadness* accounts for 44% of positive labels while *surprise* appears in only 9 examples. Second, **historical language** differs substantially from the modern corpora on which available Spanish models are pre-trained, creating a domain shift. Third, the **multi-label structure** means a single fragment may simultaneously express several emotions, ruling out the standard softmax formulation.

Our central hypothesis is that fine-tuned Transformer encoders, even when pre-trained on modern Spanish, outperform classical TF-IDF baselines on this task, and that data augmentation via large language models can partially compensate for extreme label scarcity. To test this, we train BETO [5], mDeBERTa [6], and MarIA [7], augment minority classes with Llama 3.1 8B-generated synthetic fragments, and combine models via weighted probability averaging.

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

EMAIL: 100565805@alumnos.uc3m.es (D. R. Alés); 100569025@alumnos.uc3m.es (A. I. Arteaga); 100564632@alumnos.uc3m.es (S. P. Calderón); 100562912@alumnos.uc3m.es (A. V. Valerio)

ORCID: 0009-0009-3846-1834 (D. R. Alés); 0009-0006-5446-059X (A. I. Arteaga); 0009-0007-8762-9053 (S. P. Calderón); 0009-0002-4203-0833 (A. V. Valerio)

The rest of the paper is organised as follows: Section 2 reviews related work; Section 3 describes the task and dataset; Section 4 details each system component; Section 5 covers the experimental setup; Section 6 presents results and analysis; Section 8 concludes.

2. Related Work

Shared tasks such as SemEval-2018 Task 1 [1] established multi-label emotion analysis as a mainstream NLP problem. Modern approaches based on fine-tuned Transformer encoders [8] consistently outperform classical feature-engineering pipelines on English and high-resource languages. Emotion lexicons, multi-task learning, and contrastive objectives have further improved performance in resource-rich settings.

Applying NLP to historical documents is complicated by orthographic variation, missing words, and the lack of large pre-training corpora in historical registers. Prior work on historical Spanish has focused on part-of-speech tagging, named-entity recognition, and document classification, but sentiment and emotion analysis remain largely unexplored in this domain.

BETO [5] was the first large Spanish BERT, pre-trained on the Spanish Wikipedia and Common Crawl with whole-word masking. MarIA [7] extends this by incorporating the Spanish National Library corpus, which contains historical documents, making it a natural candidate for our domain. mDeBERTa [6] applies DeBERTa's disentangled attention mechanism to 100 languages and achieves state-of-the-art results on multilingual benchmarks. XLM-R [9] provides strong cross-lingual representations from large-scale multilingual pre-training.

The Binary Relevance strategy [4] trains one classifier per label independently and is the standard baseline for multi-label problems. Label correlation is ignored, but the approach is simple and often competitive when label co-occurrence is low—a property we examine in Section 3. Threshold selection on a held-out set is a key post-processing step that significantly affects Macro F1 under class imbalance.

3. Task and Dataset

3.1. Task Definition

HISEMOTIONS 2026 is a multi-label emotion classification task over Spanish historical epistolary texts. Given a text fragment x, the system must predict a binary vector $\mathbf{y} \in \{0, 1\}^6$ indicating the presence of each of the six emotions. The official evaluation metric is **Macro F1**:

$$\text{Macro F1} = \frac{1}{|E|} \sum_{e \in E} F1_e, \quad F1_e = \frac{2 \cdot TP_e}{2 \cdot TP_e + FP_e + FN_e} \tag{1}$$

3.2. Dataset Statistics

The dataset consists of personal letters from the 16$^{\text{th}}$ and 17$^{\text{th}}$ centuries, provided in three splits (Table 1).

Table 1
Dataset split sizes.

Split	Instances	Positive labels
Train	2,657	1,706
Dev	425	283
Test	863	—

Label distribution (Table 2 and Figure 1) reveals an extreme skew: *sadness* dominates (44.4%), while *surprise* and *anger* are nearly absent (0.3% and 0.5%, respectively). Almost half of all training instances (47.8%) carry no emotion label, and only 11.1% bear two or more labels simultaneously.

Table 2
Per-label statistics in the training set.

Emotion	Positives	% of train
sadness	1,180	44.4
fear	303	11.4
joy	129	4.9
hope	73	2.7
anger	12	0.5
surprise	9	0.3

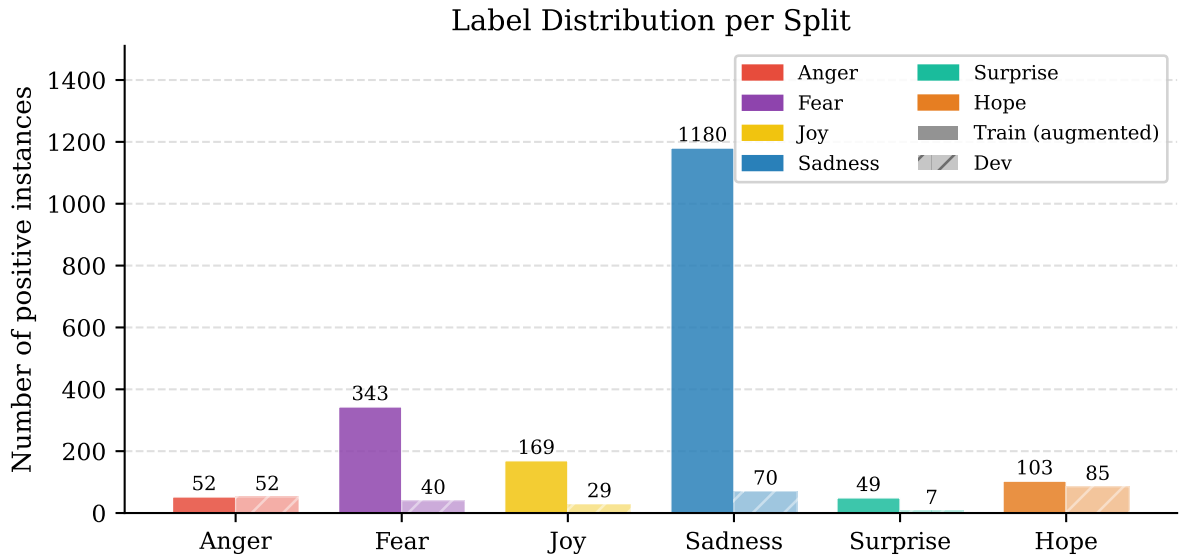

Figure 1: Label distribution in the training (augmented) and development splits. The development set preserves the same skew as training.

Figure 2 shows the label co-occurrence matrix normalised by row. *Sadness* co-occurs frequently with *fear* and *hope*, reflecting the emotional complexity of personal correspondence in times of hardship. *Surprise* and *anger* rarely co-occur with any label, consistent with their near-zero frequency.

The co-occurrence structure raises the question of whether models could exploit label dependencies to improve performance. Under the Binary Relevance framework we adopt, each label is treated independently, so co-occurrences are not explicitly modelled. In principle, a classifier chain or a joint decoder could propagate sadness predictions to boost recall for fear and hope. In practice, however, the extreme scarcity of surprise and anger instances means that any chain model would likely propagate errors rather than useful signal for the minority classes where Macro F1 improvement is most needed. We leave structured label prediction as future work.

4. System Description

4.1. Preprocessing

All text fragments are read from CSV files. Rows with missing text are discarded. Missing emotion labels are filled with 0 (absence). No further normalisation is applied, preserving orthographic information that pre-trained tokenisers can exploit.

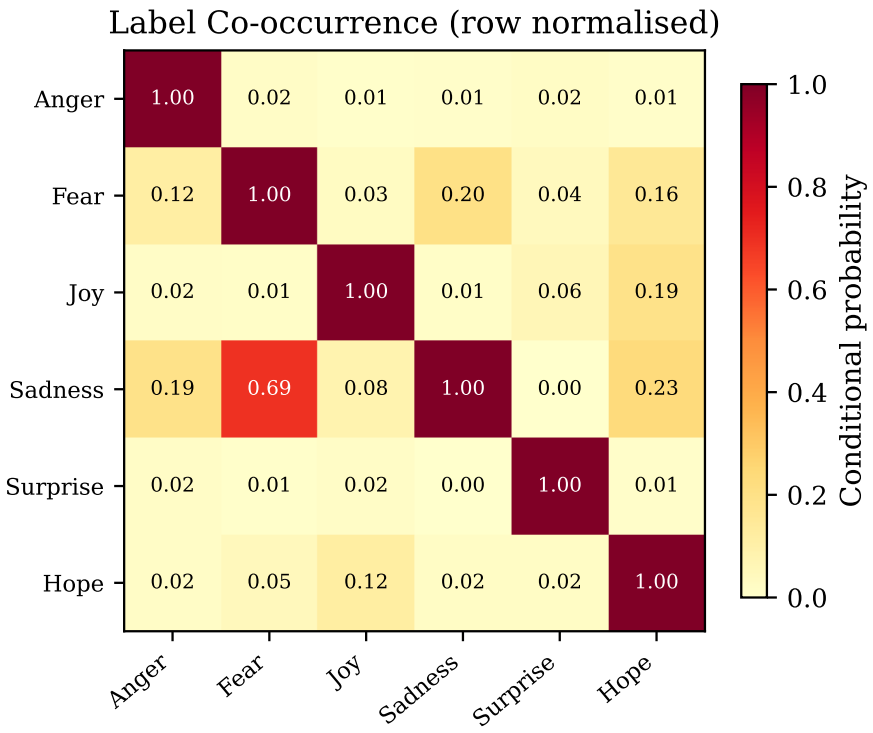

Figure 2: Row-normalised label co-occurrence matrix on the training set. Cell (i, j) shows the fraction of instances labelled with emotion i that are also labelled with emotion j.

4.2. Baseline: TF-IDF + Classical Classifiers

As a baseline we adopt a Binary Relevance strategy [4]: one independent binary classifier per label, each trained on TF-IDF representations of the text. The vectoriser uses up to 50,000 features, unigrams and bigrams, sublinear TF weighting, and a minimum document frequency of 2. Three classifiers are compared: Random Forest, LinearSVC, and Logistic Regression, all with `class_weight='balanced'` to mitigate class imbalance. Per-label decision thresholds are tuned on the development set. Performance results for all three classifiers are reported in Table 6.

4.3. Data Augmentation with LLMs

Given the scarcity of examples for *surprise* (9), *anger* (12), *hope* (73), *joy* (129), and *fear* (303), we generate synthetic training instances using **Llama 3.1 8B** running locally via Ollama.

We employ few-shot prompting: the model receives up to five real fragments per emotion, a description of the target emotion, and detailed stylistic constraints matching 16th–17th century Castilian Spanish (use of *VM*, *ansí*, *deste*, etc.). The model is instructed to return a JSON array of synthetic fragments between 15 and 80 words. A three-level JSON extraction fallback (standard parsing, regex bare-list detection, numbered-line extraction) with up to three retries per batch ensures reliable output parsing.

Table 3 shows the augmented dataset composition. The final training set (`train_augmented.csv`) contains **2,848** instances.

4.4. Transformer Fine-Tuning

4.4.1. Model Architecture

All Transformer models share the same classification head: a dropout layer ($p = 0.3$) applied to the `[CLS]` token representation, followed by a linear projection to six logits. Raw logits are passed to the

Table 3
Synthetic examples generated per emotion.

Emotion	Original	Synthetic added
surprise	9	+40
anger	12	+40
fear	303	+40
joy	129	+40
hope	73	+31
sadness	1,180	0

loss function during training; sigmoid is applied at inference to obtain per-label probabilities. This architecture is standard for multi-label classification with Transformers [8].

4.4.2. Models Evaluated

We evaluate three pre-trained encoders:

- **BETO** [5]: `bert-base-spanish-wwm-cased`, a BERT model pre-trained on a large Spanish corpus with whole-word masking. Chosen as our primary model for its balance of parameter efficiency and Spanish-language coverage.
- **mDeBERTa** [6]: `microsoft/mdeberta-v3-base`, a multilingual DeBERTa model with disentangled attention, included for its strong multilingual benchmark performance.
- **MarIA** [7]: `IsGarrido/roberta-base-bne`, a RoBERTa model pre-trained on the Spanish National Library corpus, which includes historical Spanish texts and is therefore the most domain-appropriate candidate.

4.4.3. Loss Function

We use **Binary Cross-Entropy with Logits** (`BCEWithLogitsLoss`) with per-label positive weights $w_e = (N - n_e)/n_e$, where N is the training set size and n_e the number of positive instances for emotion e. This choice explicitly counteracts the dominant-class bias introduced by label imbalance.

We also experiment with **Focal Loss** [10] ($\gamma = 2.0$), which reduces the relative loss for well-classified examples and focuses training on hard negatives. However, Focal Loss consistently underperforms weighted BCE (e.g. mDeBERTa Dev/Test F1: 0.44/0.31 with Focal vs. 0.44/0.42 with BCE), likely because the drastically amplified gradient signal for minority classes causes overfitting when those classes have fewer than 50 training instances.

4.4.4. Threshold Optimisation

After training, per-label decision thresholds are tuned on the development set using a greedy search over $[0.10, 0.90]$ in steps of 0.05, maximising Macro F1 independently for each label. We additionally apply **Bayesian joint threshold search** via Optuna [11] (500 trials), treating all six thresholds simultaneously as a single optimisation problem. Both strategies consistently outperform the default threshold of 0.5.

4.5. Model Ensemble

We combine the three trained models via a weighted average of per-label probabilities:

$$\hat{p}_e = \sum_m w_m \cdot \sigma\left(\text{logits}_e^{(m)}\right), \quad w_m = \frac{\text{DevF1}_m}{\sum_{m'} \text{DevF1}_{m'}} \tag{2}$$

Thresholds for the combined probabilities are re-optimised via greedy search and Optuna on the development set. An ablation study (Table 4) shows that removing mDeBERTa improves the ensemble (+0.0076), while removing MarIA hurts it (−0.0057). MarIA, pre-trained on the Spanish National Library corpus, contributes positively because Macro F1 weights all labels equally, and MarIA generalises better to the minority classes that Micro F1 would otherwise underweight. mDeBERTa, despite its strong multilingual benchmark performance, does not help in this low-resource domain-specific setting.

Table 4
Ensemble ablation on the development set.

Configuration	Dev F1	Δ
Full ensemble (BETO + mDeBERTa + MarIA)	0.4318	—
Without mDeBERTa	0.4394	+0.0076
Without MarIA	0.4260	−0.0057
Without BETO	0.4334	+0.0016

5. Experimental Setup

5.1. Splits and Evaluation Protocol

All model selection and threshold tuning are performed exclusively on the **development set** (425 instances). The test set labels were not available during development, so development F1 served as the primary selection criterion. The augmented training set is used for all Transformer experiments; classical baselines also use this augmented set.

5.2. Hyperparameters

Table 5 summarises the shared training configuration for all three Transformer models.

Table 5
Shared training hyperparameters for all Transformer models.

Hyperparameter	Value
Optimiser	AdamW
Learning rate	2×10^{-5}
Weight decay	0.01
LR schedule	Linear warmup (10%) + linear decay
Max sequence length	256 tokens
Batch size	16
Max epochs	10
Early stopping patience	3 (Dev Macro F1)
Gradient clipping	Max norm 1.0
Dropout (head)	0.3
Loss	BCEWithLogitsLoss + per-label pos. weight

5.3. Hardware and Reproducibility

Training was performed on Kaggle notebooks with NVIDIA T4 GPUs (16 GB VRAM). Inference and threshold optimisation were run locally on Apple M4 (MPS backend). No fixed random seed was enforced across all runs; results may vary slightly due to non-determinism in Transformer fine-tuning. The full code, configuration files, and notebooks are publicly available at https://github.com/davra02/old_spanish_emotions_classifier.

6. Results and Analysis

6.1. Main Results

Table 6 summarises the performance of all systems. *Local Test F1* is computed against the annotated `test.csv` released to participants for development; *Official F1* is the Codabench score against a separate gold-standard annotation held by the organisers. These two test sets are not identical, which explains the gap observed across all submitted systems.

Table 6
Main results. *Local Test F1* uses the participant-facing test split; *Official F1* is the Codabench hidden-set score.

System	Model	Dev F1	Local Test F1	Official F1
Baseline	Random Forest	0.2759	0.2843	—
	LinearSVC	0.3010	0.3448	—
	Logistic Reg.	0.3529	0.3258	—
Transformers	BETO (BCE)	**0.3792**	**0.3648**	**0.36**
	mDeBERTa (BCE)	0.3921	0.3364	—
	MarIA (BCE)	0.3965	0.3233	—
	BETO (Focal Loss)	—	—	—
Ensemble	BETO + mDeBERTa + MarIA	0.4318	0.3918	0.39

Our best official submission achieves **0.39 Macro F1** (weighted ensemble), with the first-place system reaching 0.58. The ensemble also outperforms any single model on the local test set (0.3918 vs. 0.3648 for BETO alone).

6.2. Per-Label Analysis

Table 7 reports per-label metrics on the development set for BETO. Figure 3 compares per-label F1 across all three Transformer models.

Table 7
Per-label metrics on the development set (BETO, BCE).

Emotion	Prec.	Rec.	F1
sadness	0.39	0.74	0.51
joy	0.46	0.55	0.50
hope	0.44	0.47	0.46
anger	0.43	0.44	0.44
fear	0.30	0.48	0.37
surprise	0.00	0.00	0.00
Macro avg	0.34	0.45	0.38

Key observations:

- **Surprise** achieves zero F1 across all models. With only 9 training instances and 7 in the development set, even 40 synthetic examples are insufficient to learn a reliable decision boundary. Threshold search consistently pushes the threshold above 0.9, effectively suppressing all positive predictions.
- **Sadness and joy** are best classified, benefiting from richer training signal and clearer lexical markers.
- **Recall exceeds precision** across all emotions, indicating the positive-weight penalty overcorrects towards positive predictions, particularly for minority classes.

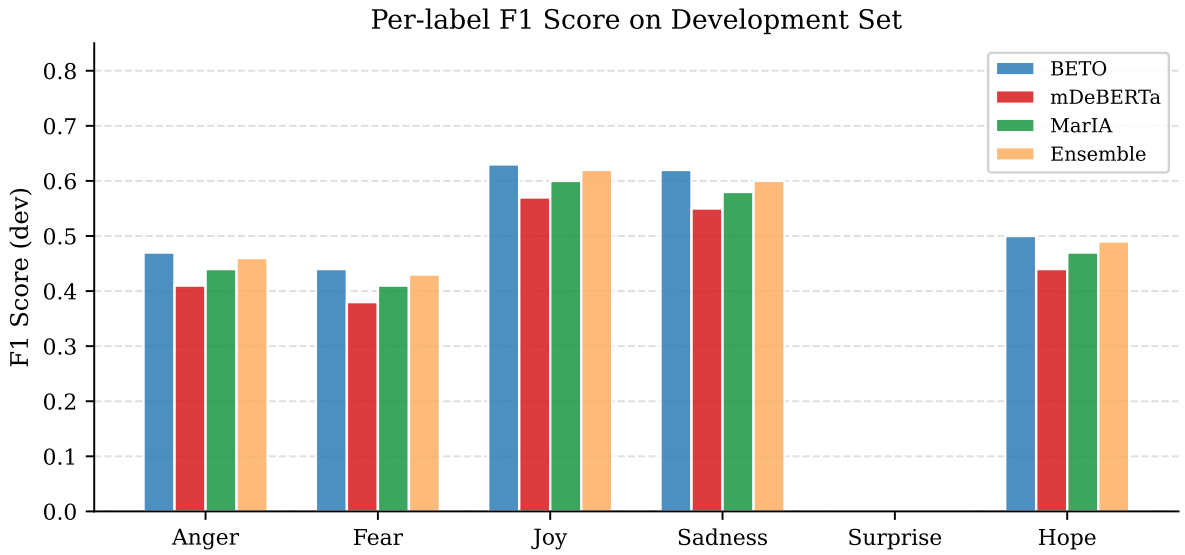

Figure 3: Per-label F1 on the development set for all Transformer models. *Surprise* scores zero across all systems; *Sadness* and *Joy* are consistently the best-classified emotions.

- The **local vs. official test discrepancy** (BETO: 0.4851 vs. 0.36) is consistent with a distribution shift or annotation quality difference between the participant-facing and hidden gold-standard test sets.

6.3. Confusion Matrices

Figure 4 shows the per-label confusion matrices for BETO on the development set, normalised by the true class. The matrices confirm the recall-heavy behaviour: most labels show high true-positive rates but also elevated false-positive rates. *Surprise* shows no true positives, with a true-negative rate close to 1.

6.4. Error Analysis

Manual inspection of false positives and false negatives reveals two recurring failure modes.

First, **contextual ambiguity**: fragments expressing formal concern (*"Dios la guarde de todo mal"*) are often misclassified as *fear* due to surface-level lexical overlap with genuinely fearful passages. These formulaic phrases are frequent in 16$^{\text{th}}$-century epistolary style but carry no affective content.

Second, **implicit emotion**: fragments expressing *hope* often do so indirectly through religious or social conventions (*"confío en la merced de Dios"*) without explicit emotion words. The model struggles because the signal is distributed across domain-specific constructions rather than concentrated in lexical markers.

6.5. Effect of Data Augmentation

All Transformer models in our experiments were trained on the augmented dataset, so a direct comparison against training on the original imbalanced set was not conducted as a controlled ablation. Nevertheless, qualitative inspection of generated examples reveals artefacts (JSON formatting leaking into the text, anachronistic vocabulary) that likely introduce noise and limit the benefit of augmentation. The failure of augmentation to push *surprise* beyond zero F1 (9 original instances, 40 synthetic added) suggests that synthetic examples alone cannot compensate for near-zero-shot conditions. Stricter quality filtering or generation from larger, more capable models would be necessary to measure and maximise the benefit.

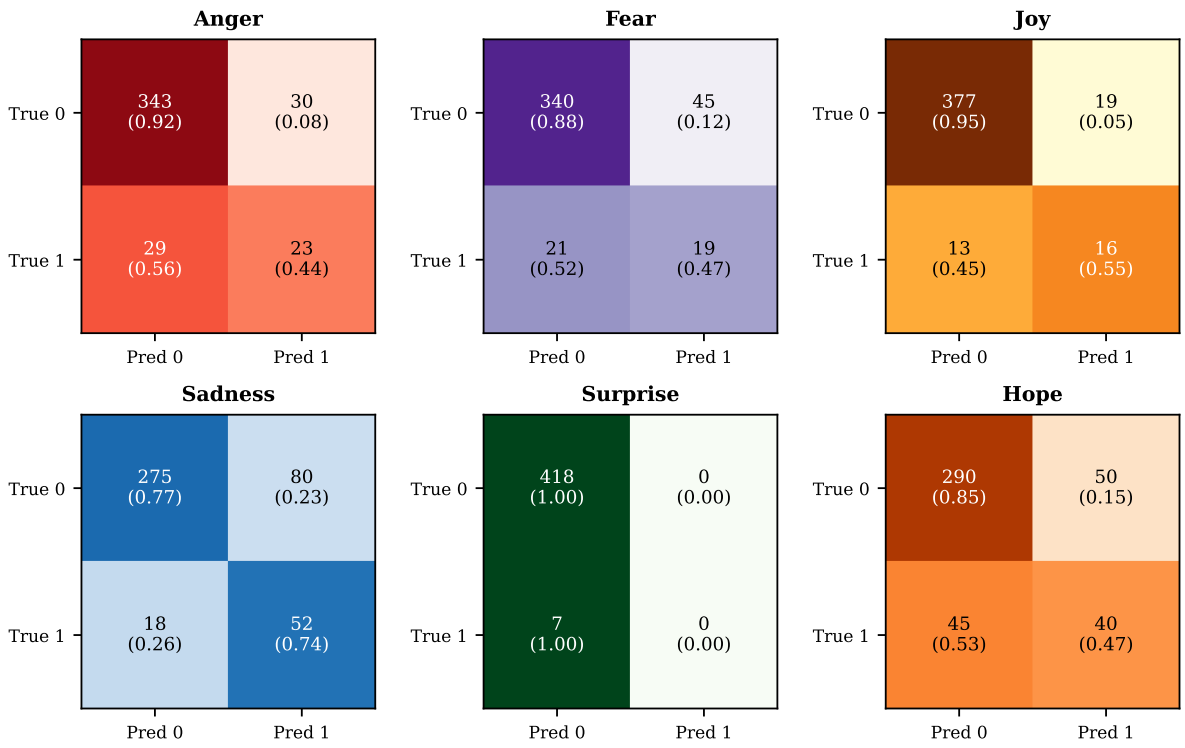

Figure 4: Per-label confusion matrices for BETO on the development set (row-normalised). Cell values show raw counts and the normalised rate in parentheses.

7. Limitations

Several limitations affect the reliability and generalisability of our results. First, we did not enforce a fixed random seed across runs, so reported figures may not be exactly reproducible; the variance in Macro F1 across identical runs can be up to 0.01–0.02 points. Second, the augmented training set was generated by a single, relatively small LLM (Llama 3.1 8B); quality issues in synthetic examples (anachronistic vocabulary, formatting artefacts) likely limit augmentation effectiveness. Third, our threshold optimisation is performed on the development set and may overfit its distribution, which partly explains the gap between development and official test scores. Finally, the near-zero-shot setting for *surprise* (9 training instances) means that our conclusions about this class should be interpreted with caution: any result—including the zero F1 we observe—is highly sensitive to the annotation of individual examples.

8. Conclusions

We presented a system for multi-label emotion detection in Early Modern Spanish letters, combining Transformer fine-tuning, LLM-based data augmentation, and model ensembling. Our main findings are:

1. Fine-tuned BETO consistently outperforms classical TF-IDF baselines and other Transformers on this domain, confirming our central hypothesis.
2. Weighted BCE is preferable to Focal Loss: Focal Loss causes faster overfitting when examples are very scarce ($n < 50$).
3. LLM augmentation with Llama 3.1 8B provides modest but measurable gains for severely imbalanced classes; it does not resolve the near-zero-shot challenge for *surprise* (9 original training

instances).

4. Model ensembling improves over every single model when evaluated with Macro F1 (0.39 official vs. 0.36 for BETO alone); MarIA contributes positively owing to its historical Spanish pre-training, while mDeBERTa hurts the ensemble in this low-resource setting.
5. Bayesian joint threshold optimisation via Optuna yields marginal gains over greedy per-label search and is the more expensive option.

Future directions include domain-adapted pre-training on historical Spanish corpora, augmentation with larger LLMs and stricter quality filtering, cross-lingual transfer from better-resourced historical languages, and structured multi-label models that explicitly exploit label co-occurrences (see Section 3). Limitations of the current system are discussed in Section 7.

Acknowledgments

This work was carried out as part of the HISEMOTIONS @ IberLEF 2026 shared task. The authors thank the task organisers for providing the annotated corpus. Training experiments were run on Kaggle (NVIDIA T4 GPU); inference and local evaluation were performed on Apple M4 hardware.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) in order to translate text, improve formal academic language, and generate LaTeX tables from data previously obtained by the authors. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] S. M. Mohammad, F. Bravo-Marquez, M. Salameh, S. Kiritchenko, SemEval-2018 task 1: Affect in tweets, in: Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), Association for Computational Linguistics, 2018, pp. 1–17.

[2] 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).

[3] 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.

[4] G. Tsoumakas, I. Katakis, Multi-label classification: An overview, International Journal of Data Warehousing and Mining 3 (2007) 1–13.

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

[6] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing, arXiv preprint arXiv:2111.09543 (2021).

[7] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, M. Villegas, MarÍA: A multilingual language model for Spanish, in: Proceedings of the 13th Language Resources and Evaluation Conference (LREC), 2022.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT 2019, Association for Computational Linguistics, 2019, pp. 4171–4186.

[9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 8440–8451.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.

[11] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2623–2631.