Emotion Detection in Early Modern Spanish Correspondence: A Multilabel Transformer Approach

Carlos Díez-Fenoy[1,*]

[1]Machine Learning for Data Science Group (ML4DS), Signal Theory and Communications Department (University Carlos III of Madrid), Leganés, 28911, Madrid, Spain

Abstract

This paper presents a system for multilabel emotion classification over early modern Spanish epistolary correspondence, submitted to the HISEMOTIONS-2026 shared task at IberLEF 2026. The task requires detecting the presence or absence of six emotion categories—*anger*, *fear*, *joy*, *sadness*, *surprise*, and *hope*—in short historical fragments, evaluated with macro F1-score. The proposed system fine-tunes `IsGarrido/roberta-base-bne`, a RoBERTa encoder pre-trained on historical Spanish text, augmented with a custom classification head trained with Focal Loss and per-class positive weights. Training follows a multilabel stratified five-fold cross-validation scheme to produce stable estimates under severe label imbalance. The final submission combines the three best-performing fold models through a validation-F1-weighted ensemble with per-emotion model routing for the two most challenging classes. The system achieves a macro F1 of 0.5656 on the official test set, with precision of 1.00 on *surprise* and recall of 0.93 on *sadness*.

Keywords

multilabel emotion classification, historical Spanish NLP, transformer fine-tuning, ensemble methods, Focal Loss, HISEMOTIONS-2026, IberLEF-2026

1. Introduction

Historical literature and correspondence offer a unique window into how past societies understood and communicated their inner lives. Beyond their cultural value, such texts encode complex emotional dynamics, interpersonal hierarchies, and communicative conventions that differ substantially from those of modern language. Automatic emotion analysis over these materials can therefore support both computational linguistics research and the broader digital humanities agenda of making historical archives accessible and interpretable at scale.

However, applying modern NLP methods to historical text raises a set of well-documented challenges. Orthographic variation, semantic shift, and the scarcity of annotated resources make direct transfer from contemporary corpora unreliable. These difficulties are compounded in short epistolary fragments, where the emotional signal may rest on a single word or formula, and where the label distribution is typically skewed across emotion classes.

HISEMOTIONS-2026 [1] is one of the shared tasks organized within the IberLEF 2026 evaluation campaign [2], addressing multilabel emotion detection in early modern Spanish correspondence from the sixteenth and seventeenth centuries. For each fragment, participating systems must predict the presence or absence of six emotions: *anger*, *fear*, *joy*, *sadness*, *surprise*, and *hope*. The official evaluation metric is macro F1-score, which weights all emotion classes equally regardless of their frequency in the corpus.

1.1. Task Description

The task requires the system to analyze short epistolary fragments from sixteenth- and seventeenth-century Spanish. For each fragment, it must determine the presence or absence of six emotions, based

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

carlosdi@pa.uc3m.es (C. Díez-Fenoy)

0009-0008-5225-3575 (C. Díez-Fenoy)

on the definitions adopted by the shared task.

- ***anger***: direct confrontation, reproach, or accusation, often linked to interpersonal tension and accusatory formulae.
- ***fear***: perceived threat, worry, or danger, frequent in contexts of illness, conflict, or uncertainty.
- ***joy***: satisfaction, relief, or pleasure, typically associated with favorable news or positive events.
- ***sadness***: grief, sorrow, or affliction, characteristic of letters related to loss, distance, or suffering.
- ***surprise***: sudden impact or unexpected discovery, usually expressed briefly and intensely.
- ***hope***: expectation or desire for improvement, closely related to future-oriented, petitionary, or trust-expressing language.

Because the macro F1-score assigns equal importance to all six classes, performance evaluation is especially demanding in a severely imbalanced dataset where some classes are represented by only a handful of documents.

1.2. Contribution

This work proposes a multilabel emotion classifier for early modern Spanish epistolary text, combining a domain-adapted transformer encoder with a training pipeline specifically designed to address severe label imbalance. The encoder `IsGarrido/roberta-base-bne`[1] [3] is fine-tuned with Focal Loss and per-class positive weights under a multilabel stratified five-fold cross-validation scheme. The final system is a validation-F1-weighted ensemble of the three best-performing folds, with asymmetric model routing for *surprise* and *hope*. The source code[2] and fine-tuned model checkpoint[3] are publicly available.

2. Related Work

The proposed approach builds on four strands of prior work: NLP on historical Spanish, emotion detection in non-standard and historical language, multilabel classification with transformers, and cross-validation and ensemble strategies for imbalanced data.

2.1. NLP on Historical Spanish Texts

Historical Spanish exhibits substantial orthographic and lexical variation, making the direct application of standard NLP pipelines unreliable. Work on diachronic corpora has therefore emphasized explicit annotation criteria, layered representations, and resources that preserve both paleographic and linguistic information while still allowing manual correction and reuse in downstream tasks [4, 5].

These efforts are particularly relevant for computational analysis of historical correspondence, where spelling variation, non-standardized language, and corpus-specific characteristics can affect the quality of text representations and downstream classification performance. The development of annotation frameworks, historical language resources, and specialized processing strategies has contributed to establishing methodological foundations for NLP in historical Spanish, supporting tasks that rely on robust textual representations and careful handling of linguistic variation. Such considerations are especially important in emotion detection, where subtle lexical and contextual cues may be influenced by the distinctive characteristics of historical texts. As a result, normalization becomes a key preprocessing strategy for reducing orthographic variation and improving the consistency of textual representations.

In this setting, normalization is not a cosmetic preprocessing step but a core component of the pipeline. Large-scale evaluations of historical text normalization show that mapping variant spellings to a contemporary form can improve downstream processing, while also highlighting that the best strategy depends on language, period, and training data availability. [6]

[1] `IsGarrido/roberta-base-bne` is a community mirror of *PlanTL-GOB-ES/roberta-base-bne*, the original model released by the *Plan de Tecnología del Lenguaje* of the Spanish Government, which is no longer available on the Hugging Face Hub. Model `IsGarrido/roberta-base-bne`:https://huggingface.co/IsGarrido/roberta-base-bne

[2] Source code and demo: https://github.com/CDF51342/HISEMOTIONS-2026-dev

[3] Fine-tuned model checkpoint: https://huggingface.co/chicodelmovil/HISEMOTIONS

2.2. Emotion Detection in Historical and Non-Standard Texts

Emotion analysis in historical language has often relied on adapting affective lexicons to older language stages and validating them against expert judgment. This line of work highlights that emotional meaning is not fully stable over time, so resources designed for contemporary language may not transfer well to historical corpora without adaptation. [7]

When the target text is literary or epistolary, annotation becomes a conceptual problem as well as a technical one. Historical emotion annotation must balance expressive coverage with feasibility, since the taxonomy should capture nuanced affective states while remaining consistent for annotators working with ambiguous passages. [8]

2.3. Multilabel Emotion Classification with Transformers

Emotion detection is often naturally formulated as a multilabel problem, because a short passage may express more than one emotion at the same time. Recent shared-task systems treat this explicitly and fine-tune transformer encoders for multilabel prediction instead of forcing a single label per instance. [9, 10]

These systems usually combine transformer representations with practical mechanisms for imbalance, such as oversampling, threshold tuning, or other decision-calibration strategies. The main lesson is that strong contextual representations are necessary, but they are not sufficient on their own when the label distribution is skewed.

2.4. Cross-Validation and Ensemble Strategies

With limited data, cross-validation provides a more stable estimate of model behaviour than a single train-dev split. This is especially useful in emotion classification, where class imbalance can make individual partitions misleading and can produce unstable performance estimates. [11]

Ensembles are a natural complement to cross-validation: averaging or weighting the predictions of several models can reduce variance and improve robustness. Recent emotion systems have used weighted voting layers or threshold-based ensembles, showing that combining complementary models often outperforms relying on a single checkpoint. [12, 9]

3. Data Analysis and Preprocessing

This section analyzes the properties of the HISEMOTIONS-2026 corpus that most directly condition the modeling choices described in Section 4. Document length and label distribution are first analyzed, followed by an analysis of label co-occurrence, the prevalence of unannotated samples and very short texts, and finally the discriminative vocabulary of each emotion class. The resulting findings motivate the use of Focal Loss, per-class threshold tuning, no-emotion subsampling, and the post-processing rules introduced in Section 4.

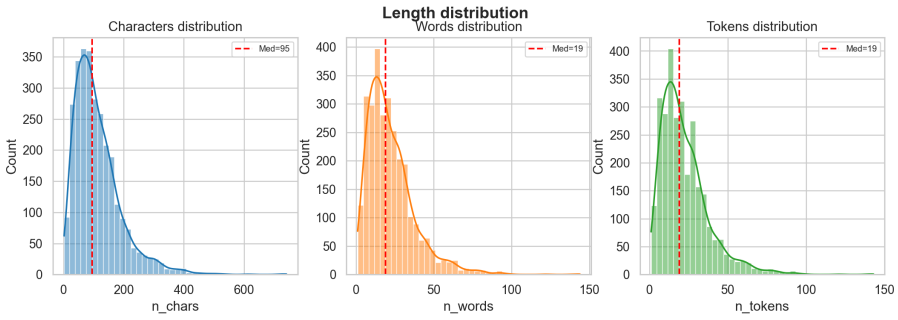

Figure 1: Global word-length distribution of the corpus. The dashed line marks the median (19 words). Over 90% of documents fall below 55 words.

3.1. Dataset Description

The corpus consists of fragments of Spanish epistolary correspondence from the sixteenth, seventeenth and early eighteenth centuries. Documents are uniformly short: the global median length is **19 words**, and more than 90% of texts fall below 55 words. This range is well within the context window of any modern transformer (512 tokens), so truncation is not an issue in practice. However, brevity has a direct implication for modeling: individual words carry disproportionate weight, and a single verb or emotionally charged noun may be the only evidence available to the classifier.

Table 1
Descriptive statistics of document length by emotion class (merged train + dev split, duplicates removed).

Emotion	Docs	Mean (w)	Median (w)	SD (w)	Median (ch)
anger	63	22.4	19.0	15.1	103
fear	333	27.8	22.0	21.3	117
joy	150	24.6	20.0	16.8	108
sadness	1206	27.1	22.0	20.9	118
surprise	14	31.4	26.5	22.7	143
hope	157	24.3	20.0	17.2	107

Table 1 reports descriptive statistics per emotion class. Two clusters emerge from the standard-deviation column. *Sadness* and *fear* show higher intra-class variance, mixing very short fragments with longer, more elaborate texts; *anger*, *joy* and *hope* are more homogeneous in length. As discussed in Section 3.5, this length clustering anticipates the lexical overlap observed between *sadness* and *fear*. Crucially, length does not act as a proxy for any emotion class, so no length-based feature is added to the model.

3.2. Label Distribution

Figure 2 shows the frequency of each emotion in the corpus and the distribution of the number of labels per document (Table 2). The imbalance is severe and operates at three distinct levels.

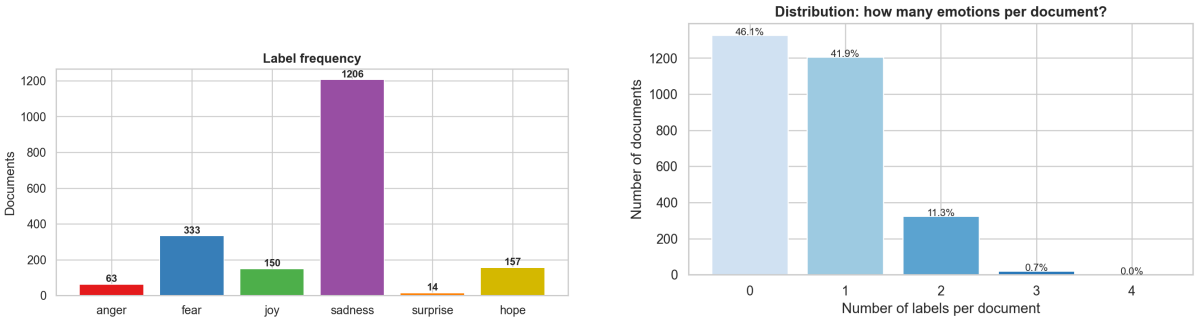

Figure 2: Left: absolute frequency of each emotion class. Right: distribution of the number of emotion labels per document.

Sadness is the dominant class, present in 1,206 documents (41.9% of labeled instances). With standard binary cross-entropy this creates a silent training bias: the loss is minimized by predicting *sadness* aggressively, yet F1-macro penalizes the resulting false positives equally to errors on minority classes. At the opposite extreme, *surprise* appears in only 14 documents (0.5%), a volume insufficient for any supervised classifier to generalize. *Anger* (63 documents, 2.2%) is manageable with re-weighting techniques.

Regarding label multiplicity, 46.1% of documents carry no emotion label, 41.9% carry exactly one, and only 12% carry two or more. This near-binary structure justifies an independent sigmoid output per emotion (multilabel head) rather than a softmax over mutually exclusive classes. The large proportion

of unannotated documents also biases decision thresholds toward non-prediction, motivating the no-emotion subsampling strategy described in Subsection 3.6.

Table 2
Distribution of the number of emotion labels per document.

No. labels	Documents	%
0	1,326	46.1
1	1,206	41.9
2	325	11.3
3	21	0.7
4	1	<0.1

3.3. Label Co-occurrence

Figure 3 shows the co-occurrence matrix and the Pearson correlation between emotion pairs. The most relevant pair is *sadness–fear*, sharing 240 documents and a correlation of 0.22. This affinity is not purely lexical: in early modern Spanish epistolary writing, suffering frequently blends fear and sadness within the same thematic field—illness, death, loss, and conflict—producing systematic ambiguity that neither lexical features nor transformer representations can fully resolve. The *joy–hope* pair (correlation 0.16) is less problematic because a temporal distinction separates the two: *joy* is reactive ("contento", "recibí", "placer") while *hope* is prospective ("espero", "venida", "plega"). This distinction is recoverable from syntactic context, making it more tractable for a contextual encoder.

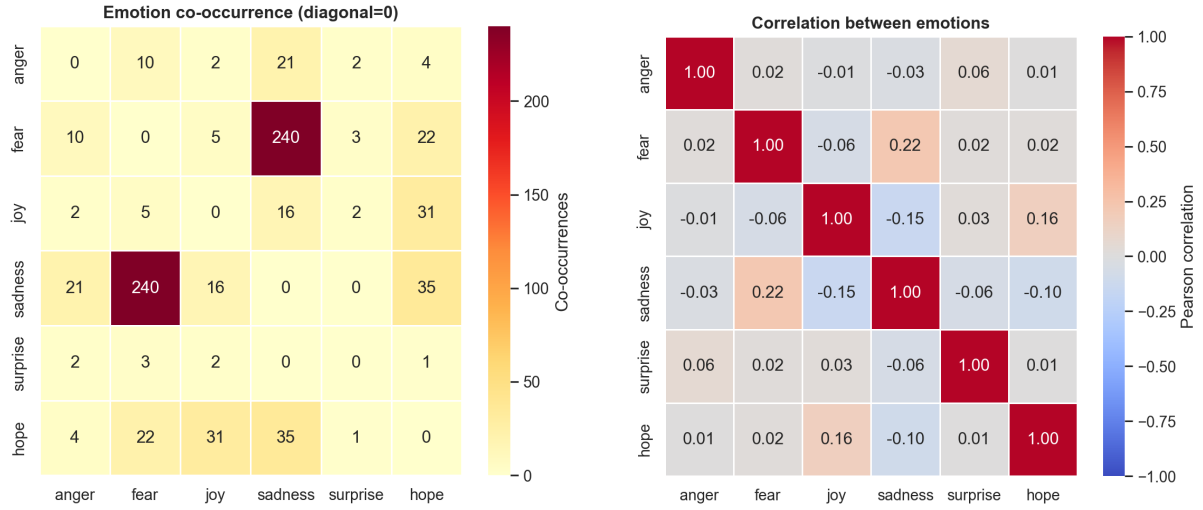

Figure 3: Left: label co-occurrence counts. Right: Pearson correlation between emotion pairs. The *sadness–fear* pair shows the strongest overlap (240 co-occurrences, $r = 0.22$).

3.4. Short Documents and No-Emotion Samples

Documents without any emotion label account for 46.1% of the corpus (1,326 instances) and are systematically shorter than labeled documents: their median length is 15 words versus 23 words for labeled texts. This difference suggests that very short fragments—letter headers, closing formulas, or administrative notations—are often captured without the surrounding context that would make their emotional content recoverable.

Table 3 reports the number of labeled documents with five words or fewer per emotion class. The 21 very short *sadness* documents are the most critical: they are sufficiently numerous to influence training, yet too short for any reliable prediction. *Surprise* is the only class with no document of five words or

Table 3
Number of labeled documents with five words or fewer, by emotion.

Emotion	Docs (≤5 words)
sadness	21
fear	4
hope	4
joy	3
anger	3
surprise	0

fewer, suggesting that the emotion of surprise in this corpus requires a minimum syntactic elaboration to be expressed.

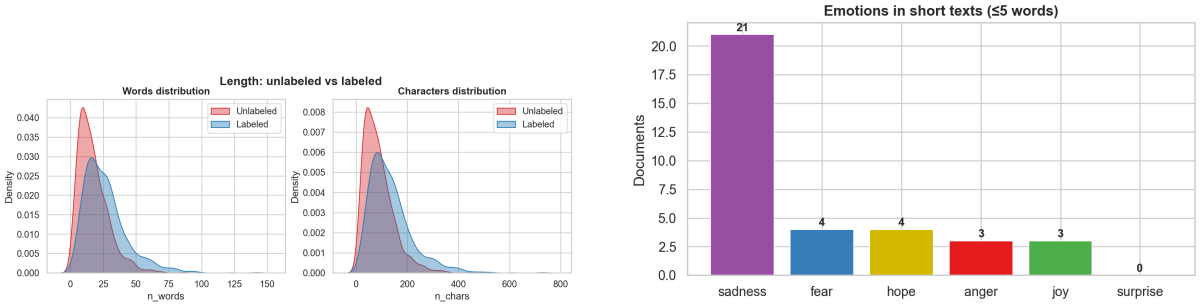

Figure 4: Left: word-length distributions for unannotated vs. labeled documents (median of 15 and 23 words respectively). Right: count of very short labeled documents (≤5 words) per emotion.

3.5. Label Co-occurrence and Lexical Exploration

TF-IDF analysis was conducted in two passes to disentangle discriminative vocabulary from generic historical corpus noise. The first pass used unigrams with a standard Spanish stop-word list; the second added bigrams and an enriched stop-word list removing high-frequency terms shared by three or more emotion classes (primarily `Dios` and `amor`). Figure 5 shows the top-scoring terms per emotion in the cleaned second pass.

- ***Anger*** exhibits the strongest and most specific lexical signal: the term `vos` scores markedly above the next candidate, and the bigrams `mala mujer` and `vos hecho` point to a register of direct accusatory address characteristic of seventeenth-century Spanish.
- ***Joy*** is the cleanest emotion lexically, dominated by reactive terms (`contento`, `gusto`, `placer`) without relevant overlap with other classes.
- ***Sadness*** lacks a dominant term after cleaning; its signal is distributed across vocabulary of diffuse suffering (`pena`, `menester`, `remedio`), which is what generates its lexical overlap with *fear*.
- ***Surprise*** cannot be reliably characterized with TF-IDF given its 14-document sample; its top terms include proper nouns specific to those texts and should not be interpreted as generalizable signals.
- ***Fear*** initially shares 9 of its top-20 unigrams with *sadness*, but this overlap falls to 1 after the second pass (Table 4), confirming that most of the confusion is attributable to generic historical-religious vocabulary rather than genuine semantic ambiguity. After cleaning, *fear* is distinguished by vocabulary of physical suffering and illness (`trabajos`, `cama`, `padezco`).
- ***Hope*** is distinguished by prospective bigrams (`soy libre`, `espero vos`) not recoverable from unigrams alone, marking a temporal contrast with *joy* that a contextual encoder should be able to exploit.

Table 4
Number of shared terms in the top-20 TF-IDF lists across three analysis passes.

Emotion pair	Original (bi)	Pass 1 (uni)	Pass 2 (bi+clean)
fear–sadness	9	2	1
joy–hope	3	3	2
sadness–hope	4	3	1
fear–joy	2	2	2
anger–fear	1	2	2
surprise–hope	1	0	2

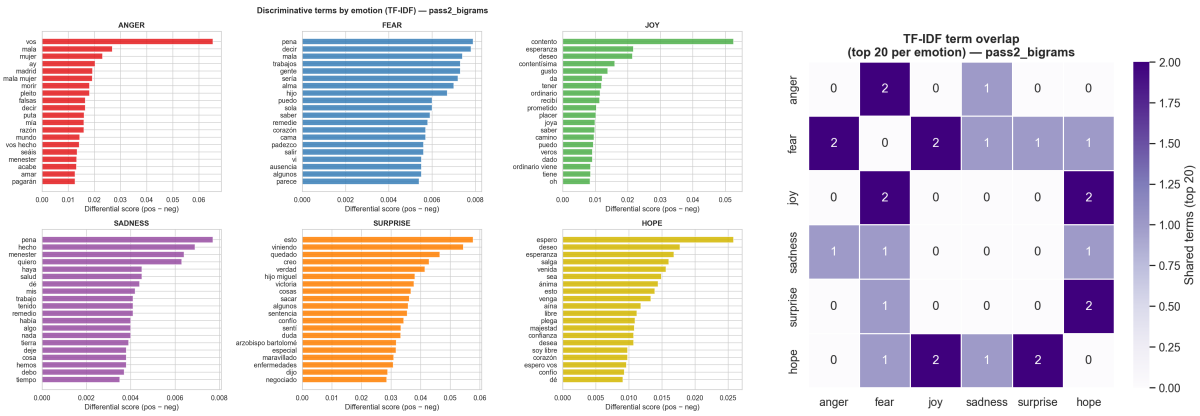

Figure 5: Left: top TF-IDF terms per emotion after the second pass (unigrams + bigrams, enriched stop-word list). Right: pairwise term overlap matrix after cleaning.

These findings collectively motivate the use of a deep contextual encoder rather than a bag-of-words approach, and the adoption of Focal Loss to concentrate gradient updates on the harder examples where lexical overlap creates ambiguity.

3.6. Preprocessing Pipeline

The preprocessing pipeline operates on the merged training and development sets (duplicates removed) and applies four steps.

- **No-emotion subsampling.** Unannotated documents are downsampled to a maximum of 400 instances per training fold, drawn uniformly at random with a fixed seed. This cap reduces the bias toward non-prediction while preserving enough negative examples for the model to learn to abstain when no emotion is present. The threshold of 400 was selected empirically through cross-validation.
- **Data augmentation.** To mitigate the scarcity of *anger*, *fear*, and *hope* examples, the training data is expanded with synthetic monolabel samples generated with Sonnet 4.6 by paraphrasing existing instances while preserving the historical register of the corpus. The generated texts were manually reviewed and curated before being incorporated into training. Augmentation is applied preferentially to the development partition, since its annotations were produced by domain experts, whereas the automatically generated training labels are more susceptible to noise. The strategy is further guided by 22 near-duplicate document pairs identified through Jaccard similarity (threshold 0.6), whose lexical differences are treated as candidate emotion-bearing cues.
- **Tokenization.** Texts are tokenized with the `IsGarrido/roberta-base-bne` tokenizer, applying padding to a fixed maximum length of 128 tokens. No text-level normalization (spelling standardization, accent removal) is applied, since the pre-trained encoder was exposed to historical Spanish orthography and normalization would discard potentially discriminative graphemic variation.

- **Positive-weight computation.** Per-class positive weights (`pos_weight`) are recomputed from the class distribution of each training fold rather than from the full dataset, ensuring that the loss function reflects the actual imbalance seen by the model during training.

4. Methodology

This section describes the architecture, training procedure, loss function, and decision-calibration strategy of the proposed system. All components are motivated by the data properties identified in Section 3.

4.1. Model Architecture

The base encoder is `IsGarrido/roberta-base-bne`[1]. This selection is motivated by domain proximity: pre-training on historical Spanish text provides lexical representations better suited to the graphemic variation and vocabulary of the early modern period than encoders trained exclusively on contemporary corpora.

A custom classification head is added on top of the encoder, consisting of a layer normalization (LayerNorm) applied to the `[CLS]` token representation, followed by two linear layers with GELU activation and dropout regularization at two points. The intermediate dimension is 256, with dropout rates of 0.3 before the first linear layer and 0.1 before the second. The final output is a vector of six logits, one per emotion class, passed through independent sigmoid activations to enable multilabel prediction.

4.2. Training Strategy

The classifier is trained under a multilabel stratified k-fold cross-validation scheme using `MultilabelStratifiedKFold` from the *iterative-stratification* library [13, 14]. Standard random partitioning is inadequate for this corpus: with $k = 5$, *surprise* (14 total instances) would be absent from at least one validation fold with high probability, producing unreliable F1-macro estimates. Stratification addresses this by iteratively assigning each instance to the fold that minimizes the deviation from the target label distribution across all six classes simultaneously, ensuring that fold diversity reflects variation in training data composition rather than in class availability—the property that makes per-fold models complementary and worth combining in an ensemble.

This concern is compounded by label quality: training annotations were produced automatically by a generative language model, whereas development annotations carry expert validation. Stratified partitioning ensures that the validation fold consistently reflects the higher-quality labeling of the development set.

Within each fold, the model is optimized with AdamW [15] and a linear warmup scheduler. Training runs for up to 20 epochs with early stopping (patience = 3) on the validation F1-macro, and the best checkpoint of each fold is saved for subsequent ensemble construction. Full hyperparameter values are reported in Table 5.

4.3. Loss Function and Class Imbalance Handling

Given the severe between-class imbalance, binary cross-entropy is suboptimal, since it assigns equal weight to all examples regardless of prediction confidence, which leads the model to prioritize the dominant class. Binary Focal Loss [16] is used instead, incorporating a modulating factor $(1 - p_t)^\gamma$ that down-weights easy examples and concentrates the gradient on hard, misclassified instances. The focusing parameter is fixed at $\gamma = 2.0$, following the original parameterization of [16].

In addition, per-class positive weights (*pos_weight*) are computed as the ratio of negative to positive instances per class, explicitly penalizing false negatives on minority classes. Per-class alpha weights are derived from inverse class frequency, normalized to the range $[0.25, 0.75]$.

4.4. Threshold Tuning

Rather than applying a uniform decision threshold of 0.5 across all emotion classes, a per-class threshold is optimized by grid search over the validation set. The search covers 50 evenly spaced values in $[0.1, 0.9]$, maximizing per-class F1-score. To prevent the search from learning overfitted thresholds, per-class bounds are imposed based on the probability distribution observed on the development set:

- ***anger***: $[0.25, 0.45]$—a reduced upper bound compensates for the model's tendency to under-predict this class.
- ***fear***: $[0.45, 0.65]$—an elevated lower bound controls false positives generated by the lexical overlap with *sadness*.
- ***joy***: $[0.30, 0.60]$—no systematic prediction bias was observed for this class; the standard range is retained.
- ***sadness***: $[0.50, 0.70]$—a high lower bound limits over-prediction of the dominant class.
- ***hope***: fixed initial value of 0.15, as the model consistently under-estimates this class.
- ***surprise***: fixed at 0.50.

4.5. Post-processing Rules

Two post-processing rules are applied to the binarized predictions. The first limits the number of active labels per document to two: if more than two emotions are predicted, only the two with the highest probability are retained. This constraint is justified by the corpus distribution, in which 99% of documents carry zero, one, or two emotion labels. The second rule silences all predictions for documents of fewer than four words, assigning them an empty label regardless of model output, since the available signal is insufficient for any reliable prediction.

5. Ensemble Construction

This section describes how the per-fold models produced by the cross-validation procedure are combined into a single prediction system.

5.1. Fold Selection

Once cross-validation training is complete, not all fold checkpoints contribute equally to the final system. Rather than aggregating all trained models indiscriminately, three folds are selected on the basis of their individual validation F1-macro, prioritizing models that generalize best to the held-out partition of their respective fold. The selected folds were trained under progressively enriched augmentation configurations targeting the most underrepresented classes, which introduces complementary inductive biases and increases ensemble diversity without additional inference cost. Selection is based exclusively on performance measured on the validation partition; no information from the test set is used at any stage of fold selection or ensemble construction.

5.2. Weighted Ensemble

The three selected fold models are combined through a validation-F1-weighted probability aggregation scheme. For each fold, the probability outputs saved at the best validation checkpoint are loaded, together with the corresponding validation F1-macro extracted from the per-epoch training log. Each fold receives a weight proportional to its validation F1-macro after positive normalization, so that better-validated models contribute more strongly to the final prediction. Aggregation consists of a weighted average of per-class probability vectors across the three folds.

6. Experimental Setup

This section documents the training configuration, the evaluation protocol, and the implementation details of the proposed system.

6.1. Training Configuration

Table 5 summarizes the hyperparameters used across all training runs. The configuration was kept fixed across folds and experiments to ensure that observed differences in validation F1 reflect fold composition rather than hyperparameter variation. The sole exception concerns the positive-weight multiplier for *anger*, which was increased across the augmented-data experimental series to amplify the gradient signal on this minority class.

Table 5
Training hyperparameters.

Hyperparameter	Value
Encoder	`IsGarrido/roberta-base-bne`
Hidden dim (head)	256
Dropout 1 / 2	0.3 / 0.1
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0.01
Warmup ratio	0.10
Batch size	16
Max sequence length	128 tokens
Max epochs	20
Early stopping patience	3
k-folds	5
Focal loss γ	2.0
No-emotion cap	400
Threshold grid	50 steps in $[0.1, 0.9]$
Max labels per doc	2
Seed	42

6.2. Evaluation Metric

The official evaluation metric of HISEMOTIONS-2026 is **macro F1-score**, computed as the unweighted average of per-class F1 scores:

$$\mathrm{F1}_{\text{macro}} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \mathrm{F1}_{\ell}$$

where $\mathcal{L} = \{\textit{anger, fear, joy, sadness, surprise, hope}\}$. This formulation treats all classes equally regardless of frequency, meaning that errors on *surprise* (14 instances) have the same weight as errors on *sadness* (1,206 instances). All model selection decisions—early stopping, fold selection for the ensemble, and threshold tuning—optimize this metric.

7. Results

Results are reported at three levels of granularity: per-fold validation performance, overall ensemble performance on the official test set, and per-class breakdown to connect outcomes with the data properties identified in Section 3.

7.1. Cross-Validation Results

The validation macro F1-scores of the folds selected for the final ensemble are summarized in Table 6. The inclusion of folds derived from three distinct training configurations contributes to the overall diversity of the ensemble.

Table 6
Validation F1-macro of the three folds selected for the final ensemble. Each experiment introduces a new augmentation layer over the previous one.

Fold	Augmentation configuration	Val F1-macro
Fold 1	*anger* samples (dev-based) and pos_weight boost	0.4717
Fold 2	+ *fear* samples (dev-based)	0.5188
Fold 3	+ *hope* samples (dev-based)	0.5229

The variability across folds reflects the inherent instability of training on a small, imbalanced corpus, where the composition of each partition directly affects the validation estimate. Folds whose validation partition contains a higher proportion of *surprise* and *anger* instances tend to produce more conservative F1-macro estimates, since errors on these minority classes are penalized equally to errors on the dominant class.

7.2. Ensemble Results

The per-class breakdown in Table 7 reveals the specific contributions and failure modes of this aggregate result. The validation-F1-weighted ensemble of the three selected folds achieves a macro F1 of **0.5656** on the official test set, exceeding the validation F1 of any individual fold model. This indicates that combining complementary models reduces prediction variance and improves robustness on the most challenging classes.

7.3. Per-Class Performance

Table 7 reports precision, recall and F1-score per emotion on the official test set.

Table 7
Per-class precision, recall and F1-score of the final ensemble on the official test set. Macro F1 = 0.5656.

Emotion	Precision	Recall	F1
anger	0.6341	0.4906	0.5532
fear	0.3824	0.7222	0.5000
joy	0.6286	0.6111	0.6197
sadness	0.3610	0.9325	0.5205
surprise	1.0000	0.5000	0.6667
hope	0.5957	0.4828	0.5333
Macro			**0.5656**

The per-class results are consistent with the exploratory analysis of Section 3.

Joy achieves the highest F1 among the frequent classes (0.6197), in line with its clean lexical signal and the absence of relevant overlap with other classes.

Surprise attains the highest system-level F1 (0.6667) with perfect precision (1.00), a direct consequence of the emotion-specific routing strategy that restricts its prediction to the two most sensitive fold models; a recall of 0.50 indicates that half of the true *surprise* instances are still missed.

Sadness shows the inverse pattern: very high recall (0.9325) but low precision (0.3610), reflecting the persistent dominance of this class during training.

Fear behaves similarly to *sadness*—high recall (0.7222) against low precision (0.3824)—suggesting that the residual semantic overlap between both classes generates systematic false positives in *fear* whenever the model detects a signal associated with *sadness*.

8. Discussion

This section discusses three cross-cutting themes that emerge from the experimental results: the role of class imbalance as the primary source of error, the specific challenge posed by very short and unannotated documents, and the contribution of the ensemble strategy to final performance. Each theme is connected to the data properties characterized in Section 3 and to the modeling decisions described in Section 4.

8.1. Effect of Class Imbalance

Class imbalance is the dominant source of error in the system. *Sadness*, present in 41.9% of labeled documents, exerts continuous pressure during training that manifests as excessively high recall at the cost of precision. Despite Focal Loss and per-class positive weights, the model learns that predicting *sadness* is a low-risk strategy, generating systematic false positives in documents that actually express *fear* or other classes with overlapping vocabulary. Raising the *sadness* decision threshold to the range $[0.50, 0.70]$ partially mitigates this effect but does not eliminate it.

A targeted manual review of *sadness*-labeled instances in the automatically annotated training partition revealed a substantial number of cases whose labels appeared inconsistent with the emotional content of the text. This observation suggests that annotation noise may contribute to the overprediction pattern observed for this class. However, alternative *sadness*-specific calibration strategies beyond the reported threshold optimization were not systematically evaluated and remain a direction for future work.

At the opposite end of the distribution, *surprise* with 14 training instances poses a qualitatively different challenge: it is not a matter of threshold calibration but of insufficient signal for generalization. The adopted strategy—restricting *surprise* predictions to the two most sensitive fold models—produces a surprisingly strong precision result, though the recall of 0.50 confirms that half of the true instances remain undetected.

8.2. Impact of Short and Empty Documents

Documents of fewer than four words represent a qualitatively different problem from class imbalance. In these texts, the emotional signal is effectively absent, and any model prediction reflects noise rather than genuine content. The silencing rule applied during post-processing eliminates this noise at the cost of losing the small number of very short documents that do carry an emotion label. The high proportion of unannotated documents (46.1%) compounds this problem by biasing decision thresholds towards non-prediction, since the model learns that abstaining is frequently the correct response. Capping no-emotion documents at 400 during training reduces this bias, but does not eliminate it entirely.

8.3. Why the Ensemble Helps

The three-fold ensemble improves upon any individual fold model for two complementary reasons. First, it averages predictions in regions of the probability space where a single model is unstable, reducing the variance introduced by the random composition of cross-validation partitions. Second, the asymmetric emotion routing—whereby *surprise* and *hope* are restricted to the fold models with the highest sensitivity to each class—ensures that minority-class predictions are not diluted by less specialized models, without sacrificing global performance in favor of any specific class.

8.4. Limitations

Despite the overall effectiveness of the proposed system, several methodological and data-related constraints remain that may affect the stability and generalizability of the reported results.

1. **Scarcity of the *surprise* class.** The *surprise* category contains only 14 documents, making performance estimates for this class highly unstable. Consequently, a single misclassified document in the test set can substantially affect its F1 score and lead to unreliable conclusions regarding the model's true performance.
2. **Semantic overlap between *fear* and *sadness*.** Despite the threshold optimization strategy, these two emotions share linguistic and contextual characteristics that generate systematic false positives. As a result, further reductions in misclassification would likely require more sophisticated modeling approaches rather than simple threshold adjustments, which would otherwise reduce recall.
3. **Large proportion of unannotated documents.** The dataset contains a considerable number of texts without emotion labels, limiting the model's ability to differentiate between genuine emotional absence and weak emotional expression. This ambiguity may negatively affect classification performance, particularly in borderline cases.
4. **Limited size of the fine-tuning corpus.** The relatively small training corpus constrains the effectiveness of the fine-tuning process and increases sensitivity to random initialization and dataset partitioning. Consequently, performance may vary depending on the random seed and the composition of the cross-validation folds.

9. Conclusions

9.1. Summary of the Proposed Approach

This paper has presented a multilabel emotion classifier for early modern Spanish epistolary texts based on four core components. First, the encoder `IsGarrido/roberta-base-bne` provides contextual representations adapted to the orthographic and lexical conventions of the target period. Second, Focal Loss with per-class positive weights addresses the severe label imbalance without resorting to aggressive oversampling that would distort the training distribution. Third, multilabel stratified k-fold cross-validation produces stable, representative validation estimates and a set of complementary fold models suitable for ensembling. A validation-F1-weighted ensemble combining the three best-performing folds—with emotion-specific model routing for *surprise* and *hope*—reduces prediction variance and improves robustness on minority classes. The final system achieves a macro F1 of 0.5656 on the official HISEMOTIONS-2026 test set.

9.2. Future Work

Three directions within the same methodological line offer the most promising avenues for improvement:

1. **Refined Probability Calibration:** Applying techniques such as per-class Platt scaling fitted on the development set could improve threshold quality without relying exclusively on grid search, which is prone to overfitting the validation partition.
2. **Systematic Subsampling Study:** Evaluating the no-emotion subsampling ratio through cross-validation, rather than setting it empirically, could further reduce the bias towards non-prediction.
3. **Exhaustive Combinatorial Search:** Replacing the manual fold selection procedure with an exhaustive combinatorial search over fold subsets and augmentation configurations would allow a more principled identification of complementary models that maximize the global macro F_1 score.

Declaration on Generative AI

During the preparation of this work, the author used Claude (Anthropic) in order to: grammar and spelling checking, and text revision. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

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

[2] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS. org, 2026.

[3] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0. doi:10.26342/2022-68-3.

[4] C. Sánchez-Marco, G. Boleda, J. M. Fontana, J. Domingo, Annotation and representation of a diachronic corpus of spanish, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010. URL: https://aclanthology.org/L10-1368/.

[5] M. Janssen, J. Ausensi, J. M. Fontana, Improving pos tagging in old spanish using teitok, in: Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, Linköping University Electronic Press, 2017, pp. 2–6. URL: https://aclanthology.org/W17-0502/.

[6] M. Bollmann, A large-scale comparison of historical text normalization systems, in: Proceedings of NAACL-HLT 2019, 2019, pp. 3885–3898. URL: https://aclanthology.org/N19-1389/.

[7] S. Buechel, J. Hellrich, U. Hahn, Feelings from the past—adapting affective lexicons for historical emotion analysis, in: Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), Osaka, Japan, 2016, pp. 54–61. URL: https://aclanthology.org/W16-4008/.

[8] R. Sprugnoli, A. Redaelli, How to annotate emotions in historical italian novels: A case study on i promessi sposi, in: Proceedings of LT4HALA 2024@LREC-COLING-2024, 2024, pp. 105–115. URL: https://aclanthology.org/anthology-files/pdf/lt4hala/2024.lt4hala-1.13.pdf.

[9] A. Paranjape, G. Kolhatkar, Y. Patwardhan, O. Gokhale, S. Dharmadhikari, Converge at wassa 2023 empathy, emotion and personality shared task: A transformer-based approach for multi-label emotion classification, in: Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, 2023, pp. 558–563. URL: https://aclanthology.org/2023.wassa-1.51/.

[10] O. J. Abiola, T. O. Abiola, O. E. Ojo, O. Kolesnikova, G. Sidorov, H. Calvo, Cic-ipn at semeval-2025 task 11: Transformer-based approach to multi-class emotion detection, in: Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), 2025, pp. 1609–1615. URL: https://aclanthology.org/2025.semeval-1.212/.

[11] F. Vaassen, W. Daelemans, Automatic emotion classification for interpersonal communication, in: Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011), Association for Computational Linguistics, Portland, Oregon, 2011, pp. 104–110. URL: https://aclanthology.org/W11-1713/.

[12] J. Han, M. Horikawa, S. Lehtosalo, Tue-jms at semeval-2025 task 11: Krelax: An ensemble-based approach for multilingual emotion detection and addressing data imbalance, in: Proceedings of

the 19th International Workshop on Semantic Evaluation (SemEval-2025), 2025, pp. 823–833. URL: https://aclanthology.org/2025.semeval-1.113/.

[13] K. Sechidis, G. Tsoumakas, I. Vlahavas, On the stratification of multi-label data, in: Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Springer, 2011, pp. 145–158. doi:10.1007/978-3-642-23808-6_10.

[14] P. Szymański, T. Kajdanowicz, A network perspective on stratification of multi-label data, Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications 74 (2017) 22–35.

[15] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations (ICLR), 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[16] 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. doi:10.1109/ICCV.2017.324.