Emotion Detection in Textual Records of Affective Correspondence in Spanish

Detección de emociones en registros textuales de correspondencia afectiva en español

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

Abstract: While addressing emotion detection, a specific textual-domain resource for this task may become a significant issue. There are multiple modern language corpora for emotion recognition, as well as language models trained with it. Nonetheless, how many resources for emotion detection are available for the historical textual domain in Spanish? To address this gap, we introduce a manually annotated corpus of historical correspondence in Spanish. Our work addresses the annotation challenges and the methodology for emotion labelling in this particular textual domain. Furthermore, we describe the results of the performance of existing pretrained LM and LLM in detecting emotions in the Spanish historical correspondence. Finally, we assess the performance of ensemble models based on features measured with TF-IDF relevance score. Our results show 1) moderate inter-annotator agreement; 2) LLM and pretrained models barely outperform a random classifier; 3) an ensemble model with a TF-IDF shows improvement in historical emotion detection.
Keywords: emotions in history, emotion detection, semantic classification, lexical classification.

Resumen: La disponibilidad de distintos dominios textuales para la detección de emociones representa un reto importante para esta tarea. Existen múltiples corpus lingüísticos modernos para el reconocimiento de emociones, así como modelos de lenguaje entrenados con ellos. No obstante, ¿cuántos recursos para la detección de emociones están disponibles en el dominio histórico en español? Para abordar esta brecha, introducimos un corpus de correspondencia histórica en español anotado manualmente. Nuestro trabajo analiza los desafíos de la anotación y la metodología de etiquetado de emociones en este dominio textual. Además, analizamos los resultados de rendimiento de modelos de lenguaje preentrenados y grandes modelos de lenguaje existentes para la detección de emociones en la correspondencia histórica en español. Finalmente, evaluamos el rendimiento de modelos combinados basados en características medidas con la puntuación de relevancia TF-IDF. Nuestros resultados muestran que (1) existe un acuerdo moderado entre anotadores; (2) los modelos de lenguaje existentes apenas superan el rendimiento de un clasificador aleatorio; (3) un modelo combinado con TF-IDF mejora la detección de emociones en textos históricos.
Palabras clave: emociones en historia, detección de emociones, clasificación semántica, clasificación léxica.

1 Introduction

Text-Based Emotion Detection (TBED) bridges affective computing and opinion analysis focusing on the processing, evaluation, and extraction of the emotions conveyed in subjective textual data (Canales and Martínez-Barco, 2014). Traditionally, sentiment analysis has been limited to classifying textual content as positive, negative, or neutral (Zhao, Liu, and Xu, 2016). In contrast, TBED enables the assignment of a broader range of emotions to texts (Troiano, Oberländer, and Klinger, 2023; Greschner and Klinger, 2025), such as those defined in Ekman's framework (Ekman and Friesen, 1978; Ekman, 1992), which identifies six basic emotions (sadness, joy, disgust, surprise, anger, and fear) in addition to the absence of emotion.

In recent years, TBED has attracted significant attention in computational linguistics and NLP (Acheampong, Wenyu, and Nunoo-Mensah, 2020; Acheampong, Nunoo-Mensah, and Chen, 2021; Al Maruf et al., 2024), demonstrating promising results across high- and low-resource languages, as evidenced by evaluations such as SemEval Task 11 (Muhammad et al., 2025b) and LREC (Calzolari et al., 2022; Piperidis et al., 2026). However, most research relies on social media datasets (Twitter, Reddit, Facebook), leaving other domains, particularly historical texts, largely unexplored.

In this context, the present study addresses this gap by examining TBED in the historical domain, focusing on Spanish epistolary texts. Spanish was chosen due to the abundance of English-language resources and the need for comparable datasets in other languages. Based on these motivations, our study contributes to advancing TBED research in the historical Spanish domain through the following contributions:

- Development and implementation of an initial proposal of a methodological framework for the annotation of emotions in historical textual data in Spanish.
- Creation of a new, manually annotated Spanish historical dataset of Textual Records of Affective Correspondence in Spanish, called TRACS for the TBED task.
- A comprehensive evaluation of monolingual, multilingual, large language models (LLMs), and ensemble lexical-based models using zero- and few-shot methodological approaches and our historical dataset to assess how effectively current models handle the historical TBED task. As far as we know, no experiments on this evaluation have been previously conducted.

Moreover, this study contributes a manually annotated corpus independent of social media sources and serves as a valuable resource for the digital humanities, supporting interdisciplinary research on language, emotion, and culture through computational methods (Bostan and Klinger, 2018).

The paper is organized as follows: Section 2 reviews the state of the art in text-based emotion detection, including historical texts; Section 3 details the TRACS dataset and annotation methodology; Section 4 presents experiments with LLMs, pretrained models, and lexical-based models; Section 5 offers qualitative analysis on best-performing model outputs; and Section 6 presents conclusions and future challenges.

2 Related Work

The analysis of affective states, such as sentiments, emotions, and opinions, has gained broad interest across fields like NLP, sociolinguistics, cognition studies, and machine learning (Weber, Greschner, and Klinger, 2026; Schäfer, Wagner, and Klinger, 2026; Greschner and Klinger, 2025). This growing focus has highlighted key methodological challenges in dataset annotation and validation. Consequently, several open-access annotated datasets have been developed to facilitate fine-tuning and evaluation of TBED models.

Among these labeled textual corpora, many adopt Ekman's basic emotion classification (Ekman and Friesen, 1978; Ekman, 1992). Examples include EmotionLines (Hsu et al., 2018) and MELD (Poria et al., 2019), both derived from Friends dialogues; DailyDialog (Li et al., 2017), collected from everyday conversations; and the SemEval 2025 BRIGHTER dataset (Muhammad et al., 2025a), based on Twitter. Acheampong, Wenyu, and Nunoo-Mensah (2020) provide a comprehensive overview of such "discrete" datasets, which categorize emotions into distinct classes. In contrast, dimensional approaches represent emotions in

a continuous, interdependent space, as in EmoBank (Buechel and Hahn, 2017), sourced from news headlines, essays, blogs, letters, and travel guides, the Valence and Arousal Dataset (Preoţiuc-Pietro et al., 2016), based on Facebook posts, and datasets annotated according to the cognitive appraisals theory (Troiano, Oberländer, and Klinger, 2023; Greschner, Weber, and Klinger, 2025). Overall, these corpora are mainly English and focus on similar domains such as social media, scripted dialogue, and mass media.

Regarding historical-domain resources in multiple languages, the availability of open-source, digitized datasets remains limited. Some released initiatives include the CLARIN project (Hinrichs and Krauwer, 2014), as well as resources such as DARIAH (2012), CORDE (2007), and CODEA (2015). The main advantage of these initiatives is their extensive collections of digitized texts in multiple languages. However, they lack annotations for emotion detection tasks.

The textual datasets and models employed for TBED in historical domains include the work of Schmidt, Dennerlein, and Wolff (2021), who released an annotated dataset of German literary texts from past centuries, using 13 emotion tags (desire, love, friendship, adoration, joy, gloating, fear, despair, suffering, compassion, anger, hate, and emotional movement). This resource comprises 6.500 emotion annotations generated through a semi-automatic procedure using CATMA (2020), a lexical-based annotation and analysis tool, achieving an inter-annotator agreement (IAA) of 0.333 (Cohen's κ). We treat this IAA coefficient as a low baseline, against which we aim to demonstrate an improvement in the annotation process carried out in our work.

Another contribution to TBED for historical texts is provided by Ehrlicher et al. (2019), who investigated emotions in Spanish travelogues and reports spanning the early 16th to the late 20th century. Their study yielded 1.708 manually annotated sentences with Ekman's six emotions. Given the absence of embedding representations specifically tailored to historical Spanish corpora, as mentioned in the study, the authors conducted their experiments using bag-of-words representations, achieving an F1 score of 0.36. We rely on this F1 score as a low baseline, against which we assess the performance of the automatic annotation systems evaluated in our work.

Building on these insights and the contributions outlined in Section 1, our objectives are to create annotated datasets for emotion detection in historical Spanish texts, evaluate the performance of existing LMs and LLMs on this task, and assess ensemble lexical-based models.

3 The TRACS Dataset

As outlined in the previous section, emotion detection in the historical domain remains underexplored. Our interest in emotion detection in historical texts is conditioned by the hypothesis situated at the intersection of historical research and digital humanities: that the emotion extraction validates historical arguments and helps to reconstruct past identities (Ehrlicher et al., 2019; Ortega-Sánchez, Pagès Blanch, and Pérez-González, 2020; Eustace et al., 2012). We therefore consider the development of resources for TBED in historical texts crucial, as it contributes to both historical studies and the digital humanities by facilitating and automating the identification of emotions. In the following subsections, we describe the data compilation process and the annotation methodology used to develop these resources.

3.1 Data Collection

One of the primary challenges in our study was selecting historical texts based on their potential to convey emotion content. As discussed in Section 2, previous studies on TBED in the historical domain have predominantly focused on travelogues and literary works, textual genres typically composed with a specific audience or readership in focus. In contrast, we aimed to explore a more intimate emotion context, as expressed through personal affective correspondence.

To this end, we selected and processed 40 letters (totaling 753 instances) in Spanish from the corpus Carter@s,[1] which comprises personal trans-Atlantic correspondence written and/or received by women in the sixteenth century. This particular textual domain was chosen due to its potential richness

[1]This corpus is part of the project "Fastos, archivo y cultura femenina en la América virreinal" (PID2024-156258N8-100), subsidized by the Ministry of Science, Research and Universities and the Spanish Government's State Research Agency. `https://www.archivocolonial.csic.es/hdlab/`

in emotion content. The content of the letters underwent normalization, correction and preprocessing by the project team members.

After collecting the letters, we conducted a manual review of their content to ensure their suitability for the TBED task and to establish the methodology for emotion annotation, which is outlined in the following section.

3.2 Methodology

Looking to fulfill the contribution one outlined in Section 1, we aim to develop the initial methodology and annotation guidelines, which poses a challenge as no established system exists for emotion annotation in historical epistolary corpora. We adopted an experimental approach to define an emotion tagset and guidelines, following Ehrlicher et al. (2019).

A significant challenge in annotating a historical epistolary corpus is creating a reliable gold-standard tagset. Since both the authors and contemporaneous witnesses are long deceased, the true nature of the emotions expressed cannot be directly verified. To address this, we assembled an interdisciplinary team of experts who conducted consecutive annotation rounds, with a moderator, excluded from the initial annotation, overseeing the final validation. The backgrounds of the ten experts are shown in Appendix A. Since the annotation team was drawn from two research centers (IEGPS-CSIC and ILLA-CCHS-CSIC), disagreement cases between the two groups of annotators were resolved by the moderator.

The annotation scheme was initially built upon standard NLP approaches to emotion detection, as discussed in Section 2. To evaluate the suitability of these approaches for our corpus, we annotated 20 letters within Round 1 (see Table 1) from the corpus Carter@s using Ekman's six basic emotions framework (Ekman and Friesen, 1978; Ekman, 1992). Additionally, we introduced an emotion category of "nostalgia" in this Round 1, following the definition proposed by Gilbert (2019) and van Tilburg (2023), based on the prior manual review of the content.

Following Round 1 of annotations, we observed that our initial tagset did not capture the full emotion diversity present in the letters. In particular, several annotators identified the category of "hope". Consequently, in Round 2, we annotated different 20 Spanish letters, not included in Round 1 (see Table 1), and incorporated this class, adopting the definition proposed by Feldman and Jazaieri (2024).

For the annotation procedure, we adopt the SemEval 2025 multi-label emotion detection framework (Muhammad et al., 2025b), which captures the complexity of emotions expressed in text. Regarding emotion perception, we follow Mohammad (2022), focusing on how most annotators interpret the author's emotions. However, this perceptual approach poses challenges due to annotators' cultural and individual biases (Wakefield, 2021). To mitigate this, our guidelines include before-annotation instructions to standardize emotion interpretation, as outlined below.

- Annotation focuses on the emotions expressed by the authors of the letters themselves, rather than those of third parties mentioned within the texts.
- Emotion detection should prioritize the identification of emotions currently expressed in the text, rather than those referenced in past events within the narrative.
- The non-annotated instances (assigned a value of 0) in the corpus included formulas of courtesy and politeness, as these constitute structural elements of letters and do not convey any emotion content.

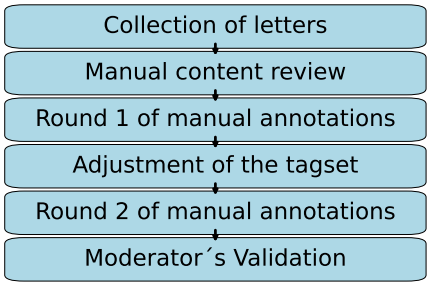

Figure 1: Annotation flow.

Figure 1 illustrates our adaptive methodology workflow, providing the flexibility needed to develop a robustly annotated corpus. The results and detailed description of the definitive dataset are presented in the following subsections.

3.3 Inter-Annotators Agreement

The IAA was computed for all instances in the historical Spanish dataset, each annotated by nine annotators. We employed the framework proposed by Artstein and Poesio (2008) to compute the weighted Fleiss's κ (Fleiss, 1971) for both Round 1 and 2, as shown in Table 1.

Round 1, which involved the manual emotion categorization of 20 letters (or 432 instances) from the corpus Carter@s, the IAA reached 0.234. This score is fair (Landis and Koch, 1977; Artstein and Poesio, 2008), considering that it represented the initial phase of annotation using an experimental methodology, with definitive guidelines still in development, and the involvement of a notably large, heterogeneous group of nine annotators. Round 2 consisted of the manual annotation of 20 additional letters (321 instances), distinct from those annotated in Round 1, but drawn from the same corpus, and the IAA increased to moderate $\kappa = 0.590$. This score reflects a positive dynamic in the annotation process. It supports the ongoing establishment of a robust methodology and comprehensive guidelines for future TBED in Spanish epistolary corpora.

Overall, we observed a moderate IAA, reflecting both the consistency among the nine annotators, and the effectiveness of review meetings during the annotation and curation phases. Discussions of disagreement cases and careful review during curation revealed inconsistencies in specific emotion categories (Table 1), with lower IAA in specific categories discussed in subsection 3.4. This process enabled a thorough revision of the Spanish epistolary corpus by a multidisciplinary team from diverse backgrounds.

To address the cases of disagreement and facilitate the moderators' validation of definitive labels, an initial small subset of lexicon was elaborated during the curation phase by our interdisciplinary group of experts as a part of our experimental methodology. This lexicon (see Appendix B) presents some textual distinctive characteristics for emotion conveying in our dataset, which also contains diachronic traits for the correct interpretation of emotion bearing units.

[2]The hope annotation was added to the 432 previously annotated instances from Round 1 later. This annotation was conducted by a smaller team of experts and reviewed by the moderator, yielding a Fleiss's κ of 0.969. The overall IAA does not account for this category in Round 1, as it was not included initially in this Round.

Features	Round 1	Round 2
letters	20	20
instances	432	321
anger	0.445	0.533
disgust	0.022	0.605
fear	0.188	0.480
joy	0.307	0.673
sadness	0.398	0.733
surprise	0.224	0.732
nostalgia	0.057	0.246
hope	N/A[2]	0.720
avg κ	**0.234**	**0.590**

Table 1: Fleiss's IAA values by emotion category. IAA values were computed across all annotators, and disagreement cases were subsequently resolved by a moderator.

Both the result of IAA and the elaboration of initial lexicon for emotion interpretation informed the development of a robust methodology and guidelines for future annotations.

3.4 Results

In pursuit of contribution two outlined in Section 1, we compiled a total of 753 instances[3] from a total 40 letters from the corpus Carter@s to create Textual Records of Affective Correspondence in Spanish or TRACS[4] (Sarymsakova et al., 2026). The statistics for each emotion category are presented in Figure 2.

Overall, 61.62% of the dataset contains one or multiple emotion categories. The distribution of the eight emotion classes is notably imbalanced, reflecting the authenticity of the original, non-synthetic texts. Despite this, the dataset is considered robust due to its rigorous annotation process: (i) a heterogeneous group of annotators, (ii) two annotation rounds with moderate IAA (0.234 and 0.590), (iii) refinement of the tagset, and (iv) final validation by a moderator, excluded from the initial annotation. In total, 464 instances contain one or more emotions, while the rest are neutral. Sadness is the most fre-

[3]We provide examples of annotated fragments of our corpus in the Appendix B.

[4]The TRACS dataset is released and made available for non-commercial academic research purposes only: `https://doi.org/10.20350/digitalCSIC/18344`

Model	1			2			3			4			5		
Class	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
anger	0.24	0.05	0.08	0.50	0.03	0.05	0.29	0.08	0.12	0.23	0.20	0.21	0.53	0.24	0.33
disgust	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fear	0.25	0.09	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.09	0.15	1.00	0.09	0.17
joy	0.30	0.04	0.07	0.21	0.56	0.30	0.39	0.32	0.35	0.19	0.44	0.27	0.34	0.26	0.29
sadness	0.32	0.05	0.09	0.51	0.14	0.22	0.59	0.59	0.59	0.47	0.80	0.59	0.79	0.41	0.54
surprise	0.50	0.04	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mac_avg	0.24	0.05	**0.08**	0.24	0.15	**0.11**	0.21	0.16	**0.18**	0.23	0.26	**0.20**	0.44	0.17	**0.22**
mic_avg	0.35	0.08	**0.14**	0.26	0.20	**0.22**	0.51	0.35	**0.42**	0.32	0.51	**0.39**	0.59	0.30	**0.40**

Table 2: The scores on multi-label emotion detection task with BERT-based multilingual (1 - multilingual_go_emotions_V1.2, 2 - distilbert-base-multilingual-cased-finetuned-emotion) and monolingual pretrained models (3 - Vera, Araque, and Iglesias (2021), 4 - Pérez et al. (2022), 5 - del Arco et al. (2020)).

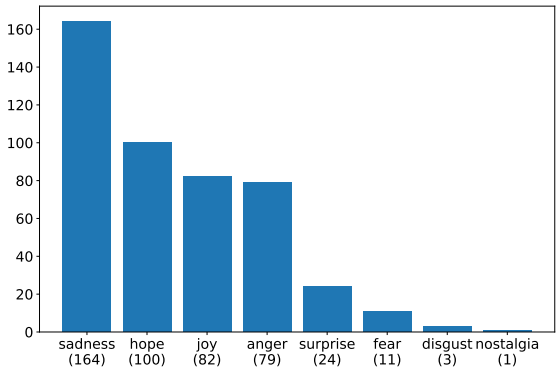

Figure 2: Distribution of emotion classes in TRACS.

quent emotion (164 instances), followed by hope (100 instances), joy (82), and anger (79). Less represented emotions include surprise (24 instances), fear (11 instances), disgust (3), and nostalgia (1). Since the corpus was annotated using a multi-label methodology, achieving a relatively balanced distribution is challenging, particularly when working with authentic texts from personal correspondence. This imbalance may be beneficial for evaluating existing models, as it allows us to examine whether they can capture contextual relationships and detect multiple emotions within a single instance, as described in the following Section 4.

4 Experiments

Aiming to fulfil contribution three from Section 1, we present experiments with pretrained BERT-based models and LLMs using zero- and few-shot evaluation techniques. Notably, zero-shot methods have shown promising results for social media TBED tasks (del Arco et al., 2020), motivating our study of model behavior in a specific historical TBED context. We also report results from evaluating these models on our dataset, providing insights into the potential for automating the emotion annotation of historical Spanish texts.

4.1 Evaluation of Pretrained Models

To evaluate the performance of pretrained models on the TBED task for historical epistolary corpora, we tested them on the TRACS dataset, comprising 40 letters or 753 instances. The results of this evaluation are presented in Table 2. Since the pretrained models for the TBED task do not fully align with our tagset, we reduced it to Ekman's six basic emotions, which are the categories shared between our tagset and the models' outputs. Although the hope category, one of the most frequent categories in the dataset, had to be excluded from this part of the evaluation due to its absence in the pretrained models' label space, the evaluation remains valid as it provides a meaningful comparison of model performance on the shared labels, and the results offer a reliable indication of their overall classification capabilities.

We selected two of the existing best-performing multilingual models and three monolingual models trained specifically for Spanish TBED. Our experiments focused on automatic emotion annotation in text, followed by evaluation against manually created ground-truth labels provided by our annotation team. The best-performing models are the monolingual Spanish ones, specifically those developed by Pérez et al. (2022) and del Arco et al. (2020), which achieved

Model	Gemma						Qwen						LLama					
Method	**Zero**			**Few**			**Zero**			**Few**			**Zero**			**Few**		
Class	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
anger	0.29	0.08	0.12	0.28	0.06	0.10	0.05	0.01	0.02	0.13	0.08	0.10	0.33	0.14	0.20	0.22	0.65	0.33
disgust	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fear	0.06	0.18	0.09	0.07	0.36	0.12	0.17	0.36	0.23	0.06	0.18	0.09	0.00	0.00	0.00	0.03	0.09	0.05
joy	0.27	0.15	0.19	0.24	0.34	0.28	0.22	0.24	0.23	0.21	0.27	0.23	0.16	0.43	0.24	0.17	0.67	0.27
sadness	0.38	0.18	0.24	0.46	0.32	0.38	0.37	0.41	0.39	0.29	0.32	0.30	0.32	0.40	0.36	0.40	0.67	0.50
surprise	0.12	0.04	0.06	0.10	0.12	0.11	0.08	0.12	0.10	0.25	0.29	0.27	0.10	0.04	0.06	0.00	0.00	0.00
hope	0.18	0.11	0.14	0.26	0.16	0.20	0.27	0.22	0.24	0.22	0.08	0.12	0.16	0.58	0.25	0.16	0.53	0.25
mac_avg	0.18	0.10	**0.12**	0.20	0.20	**0.17**	0.17	0.20	**0.17**	0.17	0.17	**0.16**	0.15	0.23	**0.16**	0.14	0.37	**0.20**
mic_avg	0.23	0.13	**0.17**	0.25	0.24	**0.24**	0.25	0.25	**0.25**	0.22	0.21	**0.21**	0.20	0.37	**0.26**	0.23	0.58	**0.32**

Table 3: Evaluation of LLMs using zero- and few-shot techniques.

a macro average F1 score of 0.20 and 0.22. The notable gap between macro and micro average F1 scores suggests that the models are able to classify certain categories, such as joy and sadness, while struggling with the remaining emotion classes.

4.2 Evaluation of LLMs

We evaluated LLM performance on the TBED task using the entire TRACS dataset,[5] the same one used for BERT-based models evaluation. Two common prompting strategies were employed: zero-shot and few-shot. Table 9 in Appendix C shows the prompting structure for both zero-shot and few-shot techniques we used for LLM evaluation, adapted from Bagdon et al. (2024).

We evaluate three LLMs suited for local deployment, Gemma-3-4b (Kamath et al., 2025), Llama-3.3-70b (Grattafiori et al., 2024), and Qwen2.5-7b (Hui et al., 2024), selected on the basis of their established performance in text-based emotion classification tasks (Muhammad et al., 2025b; Soligo, Mikulik, and Saunders, 2026; Yacoubi et al., 2025; Huang et al., 2025; Zhang and Zhong, 2025). All models are run locally via the ollama library (v0.12.9).[6]

As shown in Table 3, Llama shows stronger overall performance, improved moderately from the zero-shot to few-shot setting, with macro-F1 rising from 0.16 to 0.20 and micro-F1 from 0.26 to 0.32. Most gains came from frequent classes like sadness (0.50) and anger (0.33), suggesting improved handling of dominant emotions (Figure 2). However, performance remained weak for low-frequency classes such as disgust and surprise, both with zero F1 scores, suggesting difficulty in generalizing to less frequent and more context-dependent emotions in historical language.

In contrast, Gemma and Qwen exhibit lower overall scores, with peak macro-F1 of 0.17 and micro-F1 of 0.25. While Gemma improves from zero-shot to few-shot settings, Qwen shows the opposite trend, with slight declines in both macro-F1 (0.17 to 0.16) and micro-F1 (0.25 to 0.21). Regarding per-class emotion detection, both models still struggle with the least represented categories, such as disgust (F1 0.00), while performing comparatively better on more frequent classes, such as hope (best F1 of 0.20 and 0.24 for Gemma and Qwen, respectively) and sadness (F1 of 0.38 and 0.39, respectively).

The persistent class imbalance and limited recognition of low-frequency emotions highlight the need for further refinement for tested LLMs to enhance robustness across all categories. These results also suggest that LLMs often may fail to account for historical context, since those are trained on modern data, or fully interpret how emotions are conveyed within 16th-century Spanish texts.

The results presented in Sections 4.1 and 4.2 demonstrate that neither BERT-based emotion classification models nor LLMs are capable of accurately analyzing emotions in 16th-century Spanish historical correspondence.

4.3 Evaluation of the Lexical Model and Its Ensemble Variants

To improve the performance in the historical TBED task, we use TF-IDF relevance score and logistic regression (LR) as a classification method, and its multiple ensemble settings.

[5]The nostalgia category was excluded from evaluation, as it contains only a single instance in the entire corpus.

[6]`https://github.com/ollama/ollama/releases/tag/v0.12.9`

Split	Train	Test
1	800 P.S.	
2	1.100 P.S. (+ 300)	753 TRACS
3	1.288 P.S. (+ 188)	

Table 4: Data splits for train and test sets originated from TRACS and Post Scriptum corpus (P.S.) instances.

4.3.1 Use of Data Augmentation and TF-IDF with LR

We represent the instances of the corpus as unigrams with TF-IDF with LR, which has proven effective for information retrieval and text classification (Martínez-Cámara et al., 2011), as well as for multi-label emotion detection (Linardatou and Platanou, 2025). To address the imbalanced class distribution, we applied data augmentation strategy. The additional instances were obtained from an automatically annotated Spanish subset of the Post Scriptum (P.S.) corpus[7] (Vaamonde, 2015), annotated using Gemma-3-4B-it and subsequently revised by human experts.[8] This allowed us to avoid the need for synthetic data generation.

Table 4 presents three progressively larger training splits, all paired with the same test set of 753 TRACS instances to ensure comparability with the results reported in Sections 4.1 and 4.2. Split 1 consists of 800 manually reviewed instances from the P.S. test and development sets; Split 2 expands this to 1.100 instances by adding 300 more from the same sets; and Split 3 further increases the total to 1.288 instances with the addition of 188 more from the same sets.

Using TF-IDF as relevance measure and logistic regression as classification method across our three data splits, we observed macro-average scores ranging from 0.284 to 0.301 and micro-average scores from 0.414 to 0.421, as shown in Figure 3. Notably, even the lowest performance on the first split is comparable to the best few-shot LLM result in Table 3, which achieved a macro-average of 0.20. Nevertheless, data augmentation does not yield a significant improvement in TF-IDF + LR performance and even causes a decline starting from the third split.

[7] http://teitok.clul.ul.pt/postscriptum/es/index.php?action=downloads

[8] https://github.com/albinasarymsakova/HISEMOTIONS_2026

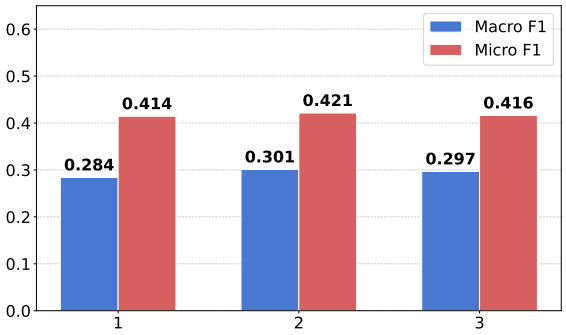

Figure 3: F1 scores for TF-IDF relevance score and logistic regression across three splits.

4.3.2 Use of Ensemble Settings

To enhance TF-IDF with LR performance, particularly for low-frequency emotions, we applied stacked generalization ensemble methods (Chatzimparmpas et al., 2021; Garouani, Barhrhouj, and Teste, 2025). We opted for the following ensemble settings as they have shown promising results in multilabel emotion classification (Haralabopoulos, Anagnostopoulos, and McAuley, 2020; Kusal et al., 2026; Adamu, Azmi Murad, and Nasharuddin, 2026): (1) TF-IDF + LR, (2) TF-IDF + LinearSVC, (3) Sentence Embeddings (Reimers and Gurevych, 2019) + LR, and (4) Stack of three previous models, each outputting a probability per label, and then averaging those probabilities before deciding the final label. All these methods also use per-label threshold optimization; that's why the results for the previously used TF-IDF + LR method differ from those reported in Section 4.3.1. We utilized the second training set (see Table 4), which includes 1.100 instance from the P.S. corpus[9] and tested, and evaluated the predictions against 753 instances from the TRACS datasets, since this setup leveraged the best results for TF-IDF + LR with augmented data in Section 4.3.1. In this way, we obtain the comparable results for evaluation to contrast with those from Sections 4.1, 4.2, and 4.3.1.

TF-IDF + LR model counts how often each word appears in the text, uses those counts to compute a score for each emotion, and predicts an emotion as present if that score exceeds a learned threshold. The

[9] The P.S. label distribution is included in the Appendix D.

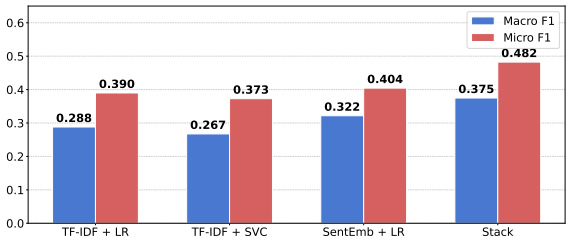

Figure 4: F1 scores for TF-IDF ensemble models.

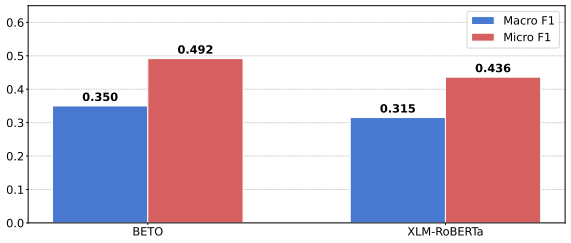

Figure 5: F1 scores for fine-tuned models.

pipeline is: Raw text → TF-IDF vectorization → linear score → sigmoid (logistic function) → per-label threshold → binary predictions. TF-IDF + LinearSVC setting uses the same word-count representation, but uses a different statistical approach to decide which side of the boundary (if emotion is present or not) a text falls on, and is represented in the following pipeline: Raw text → TF-IDF vectorization → linear decision score → sigmoid normalization → per-label threshold → binary predictions. Sentence Embeddings + LR configuration, instead of counting words, converts the whole text into a numerical representation that captures meaning, so words with similar meanings are treated as related, and a score is then computed for each emotion from that representation, as the following pipeline shows: Raw text → sentence transformer encoding → embeddings → linear score → sigmoid (logistic function) → per-label threshold → binary predictions. Finally, the Stack combines the interaction of all three aforementioned models: each model independently produces a probability for every emotion, and those three probabilities are averaged; the final yes/no decision is made from the average, so no single model can dominate the outcome: Raw text → 3 models' probabilities → element-wise average → per-label threshold → binary predictions.

Figure 4 shows that the Stack of all three models performs best,[10] achieving a macro F1 of 0.375 and micro F1 of 0.482.

4.4 Fine-tuning of Pretrained Models

The tested BERT-based models from Section 4.1) have demonstrated relatively low performance on historical Spanish TBED. To improve it, we fine-tuned two models: 1) XLM-RoBERTa base (Conneau et al., 2020), the base model for del Arco et al. (2020), the best-performing BERT-based model from Section 4.1, and BETO (Cañete et al., 2020), a monolingual Spanish BERT-based model. We employed the second training set (Table 4), consisting of 1.100 instances from the P.S. corpus, and evaluated models' predictions against 753 TRACS instances, as this configuration yielded the best results for TF-IDF + LR with augmented data (Section 4.3.1). We used a standard approach for pretrained transformer models, leveraging the *transformers* library[11] from HuggingFace (Wolf et al., 2020).

The best checkpoint based on F1 score was obtained using the following hyperparameters: 15 epochs with a batch size of 32, using a learning rate of 2×10^{-5}, and early stopping with a patience of 4 epochs; the bottom 50% of encoder layers were frozen during training, and all experiments used a fixed random seed of 42 for reproducibility.

Figure 5 shows the monolingual Spanish BETO achieves the best performance(macro-micro F1 of 0.350 and 0.492). Nonetheless, this fine-tuned model doesn't outperform the TF-IDF ensemble stack of models (Section 4.3.2) with the best result for macro F1 of 0.375.

5 Discussion

To better understand the label predictions of the Stack of (1) TF-IDF + LR, (2) TF-IDF + LinearSVC, and (3) Sentence Embeddings + LR, the best-performing model, we conducted a qualitative analysis of its outputs. The Stack integrates the outputs of all three models by averaging their independently produced emotion probabilities; the final classification decision is derived from this average,

[10] The code is released and publicly available: `https://doi.org/10.5281/zenodo.20612180`

[11] `https://github.com/huggingface/transformers`

Tag	TP	TN	FP	FN	Posit	Neg
anger	41	584	90	38	79	674
fear	5	664	78	6	11	742
joy	39	637	34	43	82	671
sadness	139	457	132	25	164	589
surprise	3	709	20	21	24	729
hope	79	504	149	21	100	653

Table 5: Classification results per emotion tag for Stack ensemble model.

Emotion	**P**	**R**	**F1**
anger	0.31	0.52	0.39
fear	0.06	0.45	0.11
joy	0.53	0.48	0.50
sadness	0.51	0.85	0.64
surprise	0.13	0.12	0.13
hope	0.35	0.79	0.48
macro avg	0.32	0.54	**0.37**
micro avg	0.38	0.67	**0.48**

Table 6: Metrics of the Stack of (1) TF-IDF + LR, (2) TF-IDF + LinearSVC, and (3) Sentence Embeddings + LR, the best-performing model, evaluated on TRACS corpus.[12]

which we believe is the model's main advantage over the other evaluated models in our study.

Table 5 shows that the model has a stronger tendency to predict correctly True Negatives (TN): for instance, for anger out of 674 of total negatives, it has correctly identified 584 cases, while for True Positives, it detected 41 out of 79. For True Positives (TP), the overall tendency in the prediction of high-frequency classes is positive: the hope category was correctly identified 79 times out of 100, and sadness was detected correctly 139 out of 164. Also, the False Positives (FP) rate is high for these classes (132 FP vs 139 TP for sadness and 149 FP vs 79 TP for hope). Moreover, low-frequency label prediction remains an issue: fear was correctly predicted 5 times out of 11, and surprise only 3 times out of 24, indicating the model still struggles to detect low-frequency emotion categories accurately.

Table 6 exhibits that categories like fear and surprise show very low F1 scores (0.11 and 0.13), while sadness, joy and hope provide relatively high F1 (0.64, 0.50, and 0.48), confirming our previous observation on issues with handling low-classes predictions, but also, indicating possible confusion in those classes due to diachronic shifting in the context of XVI century Spanish emotion expression. This suggestion is also confirmed by the overall high rates of FP across all emotion classes, as Table 5 shows.

6 Conclusion

To the best of our knowledge, this study presents several novel contributions to the field of TBED in the historical domain of Spanish, not examined in prior research:

First, we developed an initial methodological framework for annotating emotions in historical texts, including a small lexicon to resolve disagreements, and implemented it with nine annotators and a moderator for the TBED annotation of a TRACS corpus.

Second, leveraging this framework, we created a novel, manually annotated TBED corpus based on 16th-century Spanish personal correspondence (TRACS), providing a valuable resource for future historical TBED research.

Third, we evaluated monolingual and multilingual transformer-based models, LLMs, and lexical-based approaches for historical TBED. Both pretrained and LLM-based models show a limited ability to interpret the historical context and emotion conventions specific to 16th-century Spanish correspondence, suggesting the lack of this kind of text in its training data. In contrast, the lexical approaches show a higher performance. The best-performing model is the Stack of TF-IDF + LR, TF-IDF + LinearSVC, and Sentence Embeddings + LR, integrating the outputs of all three models by averaging their emotion probabilities. This model achieved macro and micro F1 scores of 0.37 and 0.48, outperforming both the pretrained and fine-tuned BERT models as well as the LLMs.

Future research should address the class imbalance, through either the inclusion of more balanced training examples or more sophisticated machine learning techniques, and the limited capacity of the Stack model to capture diachronic semantic shifts in 16th-century Spanish emotion expression.

Limitations

Regarding the limitations of our study, while we focused on 16th-century Spanish corre-

[12] We excluded the nostalgia and disgust categories due to insufficient fragments for reliable evaluation.

spondence, additional resources are needed for TBED in the historical domain. The present study is only a first step toward understanding how emotions are conveyed in historical correspondence, and it demands further, deeper exploration. Such advancements require the refinement of the annotation methodology, and the enhancement of our initial lexicon (see Appendix B) that captures the distinctive textual features used to convey and accurately interpret emotion-bearing content.

A further limitation concerns the composition of the dataset itself. The TRACS is an authentic historical text corpus and is not balanced across emotion categories, with some emotions being considerably more frequent than others. This imbalance directly affects model evaluation, as classifiers tend to optimize for majority classes at the expense of underrepresented ones, making it difficult to draw reliable conclusions about model performance on rare emotion categories.

Despite achieving the best overall performance, the Stack model exhibits important limitations. Most notably, it tends to predict True Negatives more reliably than True Positives due to imbalanced data used for both training and evaluation. Furthermore, its predictions are strongly influenced by class representation in the training and evaluation data: emotions that appear more frequently are consistently classified more accurately, while underrepresented classes remain difficult to detect. These findings suggest that the model has not fully overcome the class imbalance inherent in the dataset, and that addressing this imbalance constitutes a critical direction for future work.

While this study focuses on 16th-century Spanish correspondence, the extent to which its findings generalize to other historical corpora remains an open question. Factors such as language, genre, register, temporal distance, and the availability of domain-specific resources would all influence transferability. Adapting the methodology to more recent historical corpora or to other languages would require revisiting the emotion taxonomy, retraining or fine-tuning models on new data, and reconsidering annotation guidelines. We regard cross-domain transferability as an important direction for future work.

To sum up, our experiments' outcomes highlight the importance of currently underdeveloped robustly annotated corpora, such as TRACS, for accurate domain-specific emotions detection as historical texts.

Acknowledgments

This study is part of the research project "History and emotions: Characterization and evaluation of computational linguistics and language models in emotion mining from historical sources" of Momentum CSIC Programme (ref. MMT24–IEGPS–01) within the framework of the Generación D initiative, promoted by Red.es, an entity attached to the Ministry for Digital Transformation and the Civil Service, to attract and retain talent through scholarships and training contracts funded by the Recovery, Transformation and Resilience Plan supported by the Next Generation EU funds. This work is also partially supported by funding from the Spanish National Research Council (CSIC), Special Intramural Project No. 2024ICT028. This work is also funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project ROMANET (CERV-2024-CHAR-LITI-101215052), funded by the European Union under the Citizens, Equality, Rights and Values programme, Project HEART-NLP-UJA (PID2024-156263OB-C21) and project VERITAS-H (AIA2025-163322-C64) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, Project GALENO-IA (DGP_PIDI_2024_00852) funded by the European Union under the Smart Specialization Strategy Program for Sustainability in Andalusia (S4Andalucía) 2021–2027. The authors wish to thank Sara Dueñas Romero, Adrián Moreno Muñoz and members of SINAI Research Group (University of Jaén) for their valuable support and insightful discussions throughout the development of this work. The authors are also extremely grateful to Miguel García Fernández, Cristian David López Ushakov, Judith Farré,

Sergio Bravo Sánchez, Teresa Abejón Peña, Yara Mostazo Fernández, Javier Martinez Escobar, Magdalena Caso López for their contribution and theoretical support.

References

Acheampong, F. A., H. Nunoo-Mensah, and W. Chen. 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829.

Acheampong, F. A., C. Wenyu, and H. Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

Adamu, H., M. A. Azmi Murad, and N. A. Nasharuddin. 2026. A hybrid stacked ensemble learning framework for multilabel text emotion detection. *Scientific Reports*, 16(1):7714.

Al Maruf, A., F. Khanam, M. M. Haque, Z. M. Jiyad, M. F. Mridha, and Z. Aung. 2024. Challenges and opportunities of text-based emotion detection: A survey. *IEEE access*, 12:18416–18450.

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

Bagdon, C., P. Karmalkar, H. Gurulingappa, and R. Klinger. 2024. "you are an expert annotator": Automatic best–worst-scaling annotations for emotion intensity modeling. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7924–7936, Mexico City, Mexico, June. Association for Computational Linguistics.

Bostan, L.-A.-M. and R. Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In E. M. Bender, L. Derczynski, and P. Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Buechel, S. and U. Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In M. Lapata, P. Blunsom, and A. Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain, April. Association for Computational Linguistics.

Calzolari, N., F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis. 2022. Proceedings of the thirteenth language resources and evaluation conference. Marseille, France, June. European Language Resources Association.

Canales, L. and P. Martínez-Barco. 2014. Emotion detection from text: A survey. In M. Hernandez and J. de Jesus Aguiar Pontes, editors, *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, pages 37–43, Quito, Ecuador, October. Association for Computational Linguistics.

CATMA. 2020. Catma.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Chatzimparmpas, A., R. M. Martins, K. Kucher, and A. Kerren. 2021. Stackgenvis: Alignment of data, algorithms, and models for stacking ensemble learning using performance metrics. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1547–1557, February.

CODEA. 2015. Codea+ 2022 (corpus de documentos españoles anteriores a 1900).

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages

8440–8451, Online, July. Association for Computational Linguistics.

CORDE. 2007. Corpus diacrónico del español.

DARIAH. 2012. Digital research infrastructure for the arts and humanities.

del Arco, F. M. P., C. Strapparava, L. A. U. Lopez, and M. T. Martín-Valdivia. 2020. Emoevent: A multilingual emotion corpus based on different events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1492–1498.

Ehrlicher, H., R. Klinger, J. Lehmann, and S. Pado. 2019. Measuring historical emotions and their evolution: An interdisciplinary endeavour to investigate the 'emotions of encounter'. *Liinc em Revista*, 15(1).

Ekman, P. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Ekman, P. and W. V. Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Eustace, N., E. Lean, J. Livingston, J. Plamper, W. M. Reddy, and B. H. Rosenwein. 2012. Ahr conversation: The historical study of emotions. *The American Historical Review*, 117(5):1487–1531, 12.

Feldman, D. B. and H. Jazaieri. 2024. Feeling hopeful: development and validation of the trait emotion hope scale. *Frontiers in Psychology*, 15:1322807.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Garouani, M., A. Barhrhouj, and O. Teste. 2025. Xstacking : An effective and inherently explainable framework for stacked ensemble learning. *Information Fusion*, 124:103358.

Gilbert, A. 2019. Beyond nostalgia: Other historical emotions. *History and Anthropology*, 30:293 – 312.

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

Greschner, L. and R. Klinger. 2025. Fearful falcons and angry llamas: Emotion category annotations of arguments by humans and LLMs. In M. Hämäläinen, E. Öhman, Y. Bizzoni, S. Miyagawa, and K. Alnajjar, editors, *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 628–646, Albuquerque, USA, May. Association for Computational Linguistics.

Greschner, L., S. Weber, and R. Klinger. 2025. Trust me, i can convince you: The contextualized argument appraisal framework. *ArXiv*, abs/2509.17844.

Haralabopoulos, G., I. Anagnostopoulos, and D. McAuley. 2020. Ensemble deep learning for multilabel binary classification of user-generated content. *Algorithms*, 13(4).

Hinrichs, E. and S. Krauwer. 2014. The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1525–1531, May.

Hsu, C.-C., S.-Y. Chen, C.-C. Kuo, T.-H. Huang, and L.-W. Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Huang, D., Q. Li, C. Yan, Z. Cheng, Z. Han, Y. Huang, X. Li, B. Li, X. Wang, Z. Lian, et al. 2025. Emotion-qwen: A unified framework for emotion and vision understanding. *arXiv preprint arXiv:2505.06685*.

Hui, B., J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Lu, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2409.12186*.

Kamath, G. T. A., J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ram'e, M. Rivière,

L. Rouillard, T. Mesnard, G. Cideron, J.-B. Grill, S. Ramos, E. Yvinec, M. Casbon, E. Pot, I. Penchev, G. Liu, F. Visin, K. Kenealy, L. Beyer, X. Zhai, A. Tsitsulin, R. I. Busa-Fekete, A. Feng, N. Sachdeva, B. Coleman, Y. Gao, B. Mustafa, I. Barr, E. Parisotto, D. Tian, M. Eyal, C. Cherry, J.-T. Peter, D. Sinopalnikov, S. Bhupatiraju, R. Agarwal, M. Kazemi, D. Malkin, R. Kumar, D. Vilar, I. Brusilovsky, J. Luo, A. Steiner, A. Friesen, A. Sharma, A. Sharma, A. M. Gilady, A. Goedeckemeyer, A. Saade, A. Kolesnikov, A. Bendebury, A. Abdagic, A. Vadi, A. Gyorgy, A. S. Pinto, A. Das, A. Bapna, A. Miech, A. Yang, A. Paterson, A. Shenoy, A. Chakrabarti, B. Piot, B. Wu, B. Shahriari, B. Petrini, C. Chen, C. L. Lan, C. A. Choquette-Choo, C. Carey, C. Brick, D. Deutsch, D. Eisenbud, D. Cattle, D. Z. Cheng, D. Paparas, D. S. Sreepathihalli, D. Reid, D. Tran, D. Zelle, E. Noland, E. Huizenga, E. Kharitonov, F. Liu, G. Amirkhanyan, G. Cameron, H. Hashemi, H. Klimczak-Pluci'nska, H. Singh, H. Mehta, H. T. Lehri, H. Hazimeh, I. Ballantyne, I. Szpektor, I. Nardini, J. Pouget-Abadie, J. Chan, J. Stanton, J. M. Wieting, J. Lai, J. Orbay, J. Fernandez, J. Newlan, J. Ji, J. Singh, K. Black, K. Yu, K. Hui, K. Vodrahalli, K. Greff, L. Qiu, M. Valentine, M. Coelho, M. Ritter, M. Hoffman, M. Watson, M. Chaturvedi, M. Moynihan, M. Ma, N. Babar, N. Noy, N. Byrd, N. Roy, N. Momchev, N. Chauhan, O. Bunyan, P. Botarda, P. Caron, P. K. Rubenstein, P. Culliton, P. Schmid, P. G. Sessa, P. mei Xu, P. Stańczyk, P. D. Tafti, R. Shivanna, R. Wu, R. Pan, R. A. Rokni, R. Willoughby, R. Vallu, R. Mullins, S. Jerome, S. Smoot, S. Girgin, S. Iqbal, S. Reddy, S. Sheth, S. Põder, S. Bhatnagar, S. R. Panyam, S. Eiger, S. Zhang, T. Liu, T. Yacovone, T. Liechty, U. Kalra, U. Evci, V. Misra, V. Roseberry, V. Feinberg, V. Kolesnikov, W. Han, W. Kwon, X. Chen, Y. Chow, Y. Zhu, Z. Wei, Z. Egyed, V. Cotruta, M. Giang, P. Kirk, A. Rao, J. Lo, E. Moreira, L. G. Martins, O. Sanseviero, L. Gonzalez, Z. Gleicher, T. Warkentin, V. S. Mirrokni, E. Senter, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, Y. Matias, D. Sculley, S. Petrov, N. Fiedel, N. Shazeer, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, J.-B. Alayrac, R. Anil, D. Lepikhin, S. Borgeaud, O. Bachem, A. Joulin, A. Andreev, C. Hardin, R. Dadashi, and L. Hussenot. 2025. Gemma 3 technical report. *ArXiv*, abs/2503.19786.

Kusal, S., S. Patil, A. Peerbhai, K. Kotecha, G. Selvachandran, and A. Abraham. 2026. Meta-learning ensemble for emotion detection in conversational text. *Neural Computing and Applications*, 38(4):54.

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

Li, Y., H. Su, X. Shen, W. Li, Z. Cao, and S. Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In G. Kondrak and T. Watanabe, editors, *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Linardatou, A. and P. Platanou. 2025. Angeliki linardatou at SemEval-2025 task 11: Multi-label emotion detection. In S. Rosenthal, A. Rosá, D. Ghosh, and M. Zampieri, editors, *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 502–507, Vienna, Austria, July. Association for Computational Linguistics.

Martínez-Cámara, E., M. T. Martín-Valdivia, J. M. Perea-Ortega, and L. A. Ureña-López. 2011. Técnicas de clasificación de opiniones aplicadas a un corpus en español. *Procesamiento del Lenguaje Natural*, 47:163–170.

Mohammad, S. and P. Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In D. Inkpen and C. Strapparava, editors, *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA, June. Association for Computational Linguistics.

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

Muhammad, S. H., N. Ousidhoum, I. Abdulmumin, J. P. Wahle, T. Ruas, M. Beloucif, C. de Kock, N. Surange, D. Teodorescu, I. S. Ahmad, D. I. Adelani, A. F. Aji, F. D. M. A. Ali, I. Alimova, V. Araujo, N. Babakov, N. Baes, A.-M. Bucur, A. Bukula, G. Cao, R. Tufiño, R. Chevi, C. I. Chukwuneke, A. Ciobotaru, D. Dementieva, M. S. Gadanya, R. Geislinger, B. Gipp, O. Hourrane, O. Ignat, F. I. Lawan, R. Mabuya, R. Mahendra, V. Marivate, A. Panchenko, A. Piper, C. H. P. Ferreira, V. Protasov, S. Rutunda, M. Shrivastava, A. C. Udrea, L. D. A. Wanzare, S. Wu, F. V. Wunderlich, H. M. Zhafran, T. Zhang, Y. Zhou, and S. M. Mohammad. 2025a. BRIGHTER: BRIdging the gap in human-annotated textual emotion recognition datasets for 28 languages, July.

Muhammad, S. H., N. Ousidhoum, I. Abdulmumin, S. M. Yimam, J. P. Wahle, T. Lima Ruas, M. Beloucif, C. De Kock, T. D. Belay, I. S. Ahmad, N. Surange, D. Teodorescu, D. I. Adelani, A. F. Aji, F. D. M. Ali, V. Araujo, A. A. Ayele, O. Ignat, A. Panchenko, Y. Zhou, and S. Mohammad. 2025b. SemEval-2025 task 11: Bridging the gap in text-based emotion detection. In S. Rosenthal, A. Rosá, D. Ghosh, and M. Zampieri, editors, *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2558–2569, Vienna, Austria, July. Association for Computational Linguistics.

Ortega-Sánchez, D., J. Pagès Blanch, and C. Pérez-González. 2020. Emotions and construction of national identities in historical education. *Education Sciences*, 10(11):322.

Pérez, J. M., D. A. Furman, L. Alonso Alemany, and F. M. Luque. 2022. RoBERTuito: a pre-trained language model for social media text in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7235–7243, Marseille, France, June. European Language Resources Association.

Piperidis, S., N. Bel, H. van den Heuvel, N. Ide, S. Krek, and A. Toral, editors. 2026. *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain, May. European Language Resources Association (ELRA).

Poria, S., D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July. Association for Computational Linguistics.

Preoţiuc-Pietro, D., H. A. Schwartz, G. Park, J. Eichstaedt, M. Kern, L. Ungar, and E. Shulman. 2016. Modelling valence and arousal in facebook posts. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 9–15.

Reimers, N. and I. Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Sarymsakova, A., M. García-Fernández, C. D. López Ushakov, S. Bravo Sánchez, T. Abejón Peña, Y. Mostazo Fernández, J. Martínez Escobar, M. Caso-López, J. Farré Vidal, and P. Martín-Rodilla. 2026. TRACS: Textual records of affective correspondence in spanish.

Schäfer, J., J. Wagner, and R. Klinger. 2026. Appraisal trajectories in narratives reveal distinct patterns of emotion evocation. In J. Barnes, V. Barriere, O. De Clercq, R. Klinger, C. Nouri, D. Nozza, and P. Singh, editors, *The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA 2026)*, pages 73–82, Rabat, Morocco, March. Association for Computational Linguistics.

Schmidt, T., K. Dennerlein, and C. Wolff. 2021. Towards a Corpus of Historical German Plays with Emotion Annotations. In D. Gromann, G. Sérasset, T. De-

clerck, J. P. McCrae, J. Gracia, J. Bosque-Gil, F. Bobillo, and B. Heinisch, editors, *3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *Open Access Series in Informatics (OASIcs)*, pages 9:1–9:11, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Soligo, A., V. Mikulik, and W. Saunders. 2026. Gemma needs help: Investigating and mitigating emotional instability in llms.

Troiano, E., L. Oberländer, and R. Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72, March.

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

van Tilburg, W. A. 2023. Locating nostalgia among the emotions: A bridge from loss to love. *Current Opinion in Psychology*, 49:101543.

Vera, D., O. Araque, and C. Iglesias. 2021. Gsi-upm at iberlef2021: Emotion analysis of spanish tweets by fine-tuning the xlm-roberta language model. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings, CEUR-WS, Málaga, Spain.*

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

Weber, S., L. Greschner, and R. Klinger. 2026. Less is more? the role of demographic author information in emotion classification of ambiguous text.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. 2020. Transformers: State-of-the-art natural language processing. In Q. Liu and D. Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Yacoubi, I., R. Ferjaoui, W. E. Djeddi, and A. B. Khalifa. 2025. Advancing emotion recognition through llama3 and lora fine-tuning. In *2025 IEEE 22nd International Multi-Conference on Systems, Signals and Devices (SSD)*, pages 348–353.

Zhang, J. and L. Zhong. 2025. Decoding emotion in the deep: A systematic study of how llms represent, retain, and express emotion. *arXiv preprint arXiv:2510.04064*.

Zhao, J., K. Liu, and L. Xu. 2016. *Sentiment analysis: Mining opinions, sentiments, and emotions*. MIT Press.

A Appendix 1: Demographic Information on Expert Annotators.

The profiles of the ten experts, as well as their backgrounds, are provided in Figure 6.

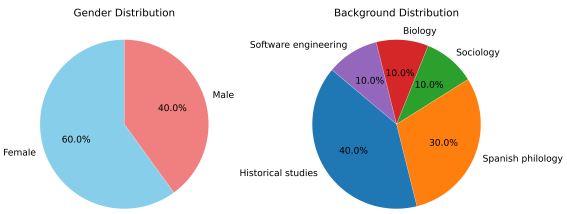

Figure 6: Distribution of annotators by gender and academic background.

B Appendix 2: Initial Emotions Lexicon and Instances from TRACS.

To address annotation disagreements, we followed Mohammad and Turney (2010) and adopted the NRC lexicon format, additionally developing a small set of supplementary lexicons in the same structure. Among these, clause-based lexicons proved especially useful for resolving ambiguous cases for TRACS dataset, as many expressions in 16th-century Spanish have undergone diachronic semantic shifts that diverge from their modern interpretations. Drawing on the historical and philological expertise of our annotation team, we constructed this lexicon (see Table 7) to

English	Spanish	A	F	J	Sad	Sur	H
to be astonished by	admirar	0	0	0	0	1	0
to get by	entretenerse	0	0	0	1	0	0
against all odds	con todo porfío	0	0	0	0	0	1
because of my sins	por mis pecados	0	0	0	0	0	0
to be widow	estar viuda	0	0	0	0	0	0
to be held captive	estar cautivo	0	0	0	0	0	0
kiss someone's hands/feet	besar las manos/los pies	0	0	0	0	0	0

Table 7: Clause-based lexicon entries from TRACS. A – anger, F – fear, J – joy, Sad – sadness, Sur – surprise, H – hope.

disambiguate emotion categories and ensure more consistent judgments when annotators disagreed. Also, the examples from TRACS corpus can be found in Table 8.

C Appendix 3: Prompting structure.

Table 9 details the prompt settings used in the LLM evaluation described in Section 4.2.

D Appendix 4: P.S. Training Data Label Distribution

Figure 7 shows the label distribution across the 1.100 instances drawn from the Post Scriptum (P.S.) corpus we used in our experiments in Section 4.

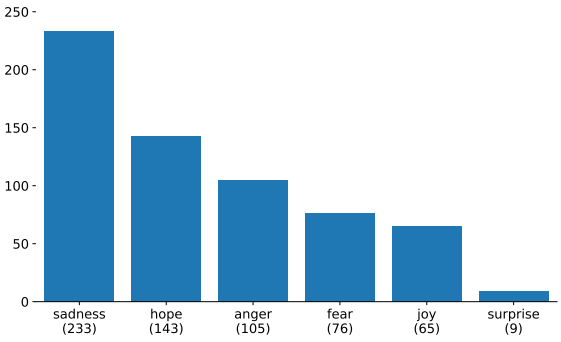

Figure 7: Label distribution for 1.100 instances from the Post Scriptum (P.S.) corpus.

Text	A	D	F	J	Sad	Sur	N	H
Y mi hermana lo hizo conmigo que Dios se lo perdone porque todo lo que tenía lo dejó a Alonso de Vera y lo hizo heredero y a mí no me dejó sino mucha dolencia porque ella estuvo enferma cuatro años en una cama y me dejó tan molida que nunca seré para ser mujer.	1	0	0	0	1	1	0	0
Por lo cual os ruego señora prima que con el portador o con otro que más presto venga me escribais largamente de toda vuestra vida o de lo que esperais de hacer de vos si tenéis voluntad de venir a esta tierra aunque no hay cosa que más desee yo y Gonzalo Martín mi marido que veros en esta tierra.	0	0	0	1	0	0	0	1

Table 8: Instances extracted from the TRACS corpus. A – anger, D – disgust, F – fear, J – joy, Sad – sadness, Sur – surprise, N – nostalgia, H – hope.

	Zero-shot	Few-shot
Role	Eres un experto en análisis de emociones en textos históricos en español.	
Task	Tu tarea es identificar qué emociones están presentes en fragmentos de cartas del siglo XVI escritas en español.	
Output format	Entrada: Muchos días ha que no recibo carta vuestra y mucho me aflijo por ello. Salida: 0,0,0,1,0,0,1	
Task description	Las únicas emociones posibles son exactamente estas siete: *anger* (ira), *disgust* (asco), *fear* (miedo), *joy* (alegría), *sadness* (tristeza), *surprise* (sorpresa), *hope* (esperanza). Para cada fragmento, devuelve una línea con 7 valores binarios separados por comas, en este orden fijo: anger, disgust, fear, joy, sadness, surprise, hope.	
Task guidelines	1) Devuelve exactamente UNA línea por fragmento, con exactamente 7 valores separados por comas (0 o 1). 2) Usa 1 si la emoción está presente en el fragmento, 0 si no lo está. 3) Un fragmento puede contener varias emociones (varios 1) o ninguna (todos 0). 4) No escribas explicaciones, encabezados, líneas en blanco ni ningún otro texto. 5) No repitas el texto de entrada ni añadas comentarios. 6) No incluyas los ejemplos en tu respuesta. 7) Mantén exactamente el mismo orden de los fragmentos que en la entrada.	
Example	N/A	A continuación se muestran tres fragmentos reales de las cartas con sus anotaciones correctas, realizadas por expertos. Úsalos como referencia para entender cómo deben clasificarse estos textos: Entrada: (1) Y mi hermana lo hizo conmigo que Dios se lo perdone. . . [1] (2) Mucho me huelgo yo de que estés ya tan mejorado. . . (3) Es verdad también que entre ellas hay muchas viciosas. . . Salida: 1,0,0,1,1,0,0 0,0,1,1,0,1,0 0,0,0,0,1,0,0
Text input	Ahora clasifica los siguientes fragmentos: {text1}, {text2}, . . .	

Table 9: Full prompt structure for zero-shot and few-shot emotion detection in historical Spanish correspondence.

[1] Full text of the three few-shot examples: (1) Y mi hermana lo hizo conmigo que Dios se lo perdone porque todo lo que tenía lo dejó a Alonso de Vera y lo hizo heredero y a mí no me dejó sino mucha dolencia porque ella estuvo enferma cuatro años en una cama y me dejó tan molida que nunca seré para ser mujer. (2) Mucho me huelgo yo de que estés ya tan mejorado que he estado sin saber de ti y lo que te pido es que te cuides mucho que ya sabes eres delicado y que cualquier cosa te hace mal. (3) Es verdad también que entre ellas hay muchas viciosas como entre todas las católicas para eso se les predica mucho que no es decible qué cuidado les cuesta a los padres de la compañía que como tú dices parece les ha infundido Nuestro Señor particular gracia.