

# Historical Ink: Exploring Large Language Models for Irony Detection in 19th-Century Spanish

Kevin Cohen<sup>1</sup> Laura Manrique-Gómez<sup>2</sup> Rubén Manrique<sup>1</sup>

<sup>1</sup> Systems and Computing Engineering Department, Universidad de los Andes

<sup>2</sup> History and Geography Department, Universidad de los Andes  
Bogotá D.C.

{k.cohen, l.manriqueg, rf.manrique}@uniandes.edu.co

## Abstract

This study explores the use of large language models (LLMs) to enhance datasets and improve irony detection in 19th-century Latin American newspapers. Two strategies were employed to evaluate the efficacy of BERT and GPT-4o models in capturing the subtle nuances nature of irony, through both multi-class and binary classification tasks. First, we implemented dataset enhancements focused on enriching emotional and contextual cues; however, these showed limited impact on historical language analysis. The second strategy, a semi-automated annotation process, effectively addressed class imbalance and augmented the dataset with high-quality annotations. Despite the challenges posed by the complexity of irony, this work contributes to the advancement of sentiment analysis through two key contributions: introducing a new historical Spanish dataset tagged for sentiment analysis and irony detection, and proposing a semi-automated annotation methodology where human expertise is crucial for refining LLMs results, enriched by incorporating historical and cultural contexts as core features.

## 1 Introduction

Irony is a nuanced and often subtle form of communication, especially in historical texts, where cultural context is crucial in understanding the intended meaning. Detecting irony in written language has long been a challenge for natural language processing (NLP) (González-Ibáñez et al., 2011), as it is a trope whose actual meaning differs from what is literally enunciated (Hee et al., 2018). Hence, irony detection involves identifying contradictions between what is said and the underlying meaning or context. This challenge becomes even more complex when dealing with historical texts, where linguistic expressions, cultural references,

and societal norms differ significantly from modern times.

This research focuses on irony detection in 19th-century Latin American newspapers, utilizing large language models (LLMs) and feed-forward neural networks to experiment with various strategies. The dataset, annotated by experts, was processed for multi-class and binary classification. BERT-like models were used for text encoding and transfer learning, while GPT-4o models were applied for sentiment classification with customized prompts. A BERT encoder with a feed-forward neural network was used as a comparative baseline to examine the enhancement of text analysis specific to irony detection.

A semi-automatic annotation approach was also developed to incorporate new, untagged data. By employing GPT-4o models with tailored prompts, initial classifications were machine-generated and verified by human experts, reducing annotation time and effort while maintaining high-quality standards. Integrating human expertise with machine-generated results enhanced the dataset and allowed for more effective model training. Consequently, the study contributes to research on sentiment analysis in historical texts and demonstrates LLMs' capabilities in enhancing text classification in specialized contexts.

## 2 Related Work

Philosophers and linguists have yet to reach a definitive agreement on defining certain figurative tropes including irony, sarcasm, satire, hyperbole, analogy, restatement, paradox, and parody. There have been arguments regarding subtle differences, such as the humorous intention in irony versus the explicitly offensive intention in verbal sarcasm. However, there is a broader consensus on considering irony as the overarching category (Kreuz, 2020; Colston, 2017). In NLP, irony, sarcasm, and

satire have often been used interchangeably, focusing on linguistic and sentiment features. Early work suggested satire exploits human tendencies like gullibility and confirmation bias (Rubin et al., 2016). Other approaches, achieving moderate success, explored linguistic patterns, including slang and profanity, using techniques such as Bi-normal Separation (BNS) (Burfoot and Baldwin, 2009).

Traditional rule-based methods relied on linguistic indicators like interjections and hyperbole (Riloff et al., 2013). However, deep learning models such as ELMo and BERT have revolutionized the field by incorporating embeddings that capture contextual information across sentences. Rajadesingan et al. emphasized the importance of user behavior and historical tweet data in improving sarcasm classification, demonstrating how conversational context can enhance detection accuracy (Rajadesingan et al., 2015).

Both satire and sarcasm detection have significantly benefited from integrating contextual features and deep learning, showing superior performance compared to traditional methods. Nonetheless, challenges remain due to the reliance on cultural and societal context, which complicates creating highly accurate models.

Irony detection has similarly grown in importance as researchers address the challenges of figurative language. Early approaches to irony detection primarily utilized static linguistic features and manually annotated datasets but struggled to capture irony's nuanced and dynamic nature, particularly when dealing with context-specific language, such as that found in social media and historical texts.

The introduction of deep learning models using transformer-based architectures for LLMs like BERT and GPT has significantly improved the capacity for irony detection. Huang et al. explored the application of deep learning techniques such as Recurrent Neural Networks (RNN), revealing that models with attention mechanisms outperformed others on irony detection tasks using social media data (Huang et al., 2017). These models excel by capturing both syntactic and semantic relationships within sentences, enabling comprehensive irony analysis. Similarly, Ren et al. proposed a knowledge-enhanced neural network that incorporates contextual information from external knowledge sources like Wikipedia to enhance irony detection performance (Ren et al., 2023).

Recent studies have focused on improving irony

detection by integrating emotional and contextual cues. Lin et al. introduced a newly developed method combining LLMs with emotion-centric text enhancement to improve the detection of irony (Lin et al., 2024). Their approach highlighted the significance of subtle emotional cues often overlooked in traditional models. By using GPT-4 to expand original texts with additional content, the researchers significantly enhanced irony detection in benchmark datasets. Ozturk et al. introduced a de-biasing approach for irony, satire, and sarcasm detection, utilizing generative LLMs to reduce stylistic biases produced by single-source corpus training datasets (Ozturk et al., 2024). Their findings indicate that such stylistic bias impacts model robustness and that LLMs-based enhancement can mitigate these biases. However, its effect on causal language models like Llama-3.1 remains limited. This approach aligns with recent trends in dataset enhancement and bias reduction to improve the detection of figurative language.

Recent methodologies introduced a human-LLMs collaborative annotation framework that addresses the limitations of LLMs-generated labels by combining automated annotation with human expertise (Wang et al., 2024). Their method incorporates three major steps: LLMs predict labels and generate explanations; a verifier assesses label quality; and human annotators review and re-annotate labels flagged as low quality. This research shows the necessity for hybrid approaches that merge LLMs scalability with human precision, particularly for complex tasks such as irony detection.

Despite advancements, challenges persist in handling class imbalance, stylistic bias, and the nuanced nature of figurative language, especially in historical texts. Prior work on deep learning models and LLMs-driven enhancements often focuses on contemporary datasets, neglecting historical linguistic and cultural context variations. Moreover, research on semi-automated annotation strategies leveraging human-LLMs collaboration is limited. Our study addresses these gaps by integrating a structured semi-automated annotation process, improving dataset balance, and fine-tuning domain-specific models for 19th-century Spanish. By combining LLMs-powered augmentation with human verification, we propose a scalable method for enriching training data while maintaining historical linguistic authenticity. This work refines irony detection in historical texts and provides a broader

framework for augmenting and improving figurative language classification in underrepresented domains.

### 3 Data

The dataset used in this research initially comprised a tagged corpus of 2,734 entries. Each entry contains text samples randomly extracted from 19th-century Latin American newspapers—the LatamXIX Dataset (Manrique-Gómez et al., 2024). Tags were manually assigned by three human experts to each entry, falling into one of the following categories: "IRONÍA" (irony), "POSITIVO" (positive), "NEGATIVO" (negative), and "NEUTRO" (neutral). These tags represent the predominant sentiment of the extracted text. In our dataset, irony involves multilayered expressions of emotions, including criticism, humor, and sarcasm, as well as the use of poetic language—a distinctive feature of the era.

The primary dataset was used to conduct experiments on text enhancement and fine-tune BERT-like models for classifying irony. As mentioned previously, the dataset comprises four distinct classes. To extend the scope of experimentation, a copy of the dataset was created where the "POSITIVO," "NEGATIVO," and "NEUTRO" classes were merged into a single class labeled "NO IRONÍA" (not irony). This transformation converts the task from multi-class classification to binary classification. For better computational processing, a new column named "category\_encoded" was created in both datasets to contain the same tags encoded as numerical values, facilitating interpretation.

In the second phase of the research, the primary dataset with the original text (without enhancement) was augmented with 1,016 additional entries. These new fragments also originated from the LatamXIX Dataset and were used to implement the semi-automated methodology for annotation.

The final dataset consists of 3,750 annotated entries, resulting in a more balanced collection that corrected the initial underrepresentation of the "IRONÍA" class, preserves the historical linguistic value of the original texts, and improves the accuracy of the LLMs BERT-like models fine-tuned for the historical irony classification task<sup>1</sup>.

<sup>1</sup>The dataset is available at <https://huggingface.co/datasets/Flaglab/latam-xix-tagged-augmented> in its three versions: "primary", "enhanced", and "augmented"

## 4 Methodology

The experimentation encompasses two primary aspects: dataset enhancement and augmentation and the construction of the classification pipeline. GPT-4o was employed in conjunction with prompt engineering to enhance the text and establish a baseline for measuring improvements in the classification task. Subsequently, BERT-like models were used to classify the dataset. Detailed explanations of these components are provided in the following subsections.

### 4.1 Dataset Enhancing

A key objective of this research was to evaluate how models like GPT-4o can enhance context and text to improve sentiment analysis, specifically focusing on detecting historical irony. Several prompts were developed and tested to achieve satisfactory results. The evaluation involved a small, balanced dataset of 40 diverse entries from the original dataset. Each prompt's performance was individually analyzed to ensure that the responses were closely aligned with the manual classification of the original data. A prompt deemed effective in performing the task was then applied to the entire dataset alongside a neural network.

The final prompt used to enhance the dataset was: *"Expand this text while preserving its original meaning, placing a strong emphasis on its emotional content to enhance the identification of its overall sentiment. Respond only with the expanded text, and strive to maintain the syntax and morphology characteristic of 19th-century Latin American Spanish."*

The prompt, originally in Spanish as detailed in Appendix A, does not reference irony or specific sentiments to avoid bias that pre-classified data might introduce. This approach allows for an initial observation of how GPT-4o expands the original texts. Appendix A.1 provides an example of the GPT-4o input and output text generated from the data enhancement process.

### 4.2 Dataset Augmentation

In addition to enhancing the original dataset, a new strategy was introduced to include previously untagged data with a high potential for irony detection. This involved designing a prompt for processing 1,034 new entries, selected from sources likely to contain ironic content.

The dataset was classified using GPT-4o to iden-

tify scenarios involving "IRONÍA," "POSITIVO," "NEGATIVO," and "NEUTRO." Appendix B details the prompt used for this task, and Appendix B.1 provides an example of the GPT-4o output tag and explanation generated in the augmentation process.

The prompt design featured four key components: [1] Context: Oriented the model to analyze 19th-century Spanish texts from Latin American newspapers. [2] Irony Recognition: Guided the model to recognize contradictions in three scenarios:

- Between described reality and expression.
- Historical reality and expression.
- Expression tone as indicated by capitalization and punctuation.

[3] Exceptions: Addressed frequent misclassifications by instructing the model not to mark irony in cases of:

- Political opinions.
- Poetic language.
- Instances lacking humorous intent.

Finally, [4] the Task: Required explanations for classification decisions to be appended in asterisks (\*), facilitating sentiment extraction via regular expressions.

Each entry, processed by GPT-4o using this prompt, produced an output with an assigned tag (e.g., 'IRONÍA') and an explanation for the classification (e.g., The text contains contradictory statements suggesting irony). An expert reviewed these outputs to verify accuracy, significantly reducing the time and effort compared to manual annotation from scratch.

The human verification process identified 18 entries with low-quality OCR transcription. These 'unreadable' entries were excluded from the final dataset augmentation. The verified new sample resulted in 1,016 entries added to the primary dataset. The augmented dataset was then used to fine-tune the model, detailed in the next section, to enhance its performance in irony detection tasks. Figure 1 illustrates this semi-automatic annotation methodology, and Table 1 summarizes the datasets mentioned:

Dataset	Num. Entries	Augmented	Enhanced	Experiment
PRIMARY	2734	NO	NO	Baseline
ENHANCED	2734	NO	YES	Prompt Based Enhancement
AUGMENTED	3750	YES	NO	Semi-Automatic Annotation

Table 1: Datasets Summary. *PRIMARY*: Primary Human Annotated Dataset. *ENHANCED*: Enhanced Dataset Automatic Annotated. *AUGMENTED*: Augmented Semi-automatic Annotated Dataset.

### 4.3 BERT-based Classification Pipeline

The architecture designed for irony classification consists of an LLMs BERT encoder with a feedforward neural network head. The model comprises three layers with the following features:

- An input layer of size 768, selected to match the standard dimensions of the contextual vector representations produced by BERT-like models.
- A first hidden layer employing a ReLU activation function, with a weight matrix of dimensions 768 x 50.
- A fully connected layer that maps the hidden layer to the output layer, with a weight matrix of dimensions 50 x output\_dim, where output\_dim can be either four or two, depending on whether the classification is multi-class or binary.

As shown in Figure 2, the architecture uses the contextual embedding generated by a BERT family model for text representation. At the bottom of each layer, the corresponding dimensions and activation functions are indicated. In the model's output layer, the dimensions vary based on the type of classification (binary or multi-class). The activation function is a sigmoid for binary classification, and the number of nodes is adjusted to two. The following BERT-like models were evaluated:

- bert-base-uncased
- bert-base-multilingual-uncased
- dccuchile/bert-base-spanish-wwm-uncased
- dccuchile/bert-base-spanish-wwm-cased
- beto-cased-finetuned-xix-latam.

The selection included standard BERT models, both base and multilingual versions, as well as models tailored for contemporary Spanish, and a version trained on 19th-century Spanish texts (Montes et al., 2024).



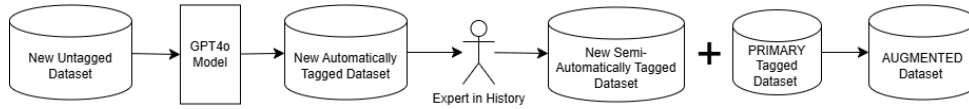


Figure 1: Semi-Automatic Annotation Methodology.

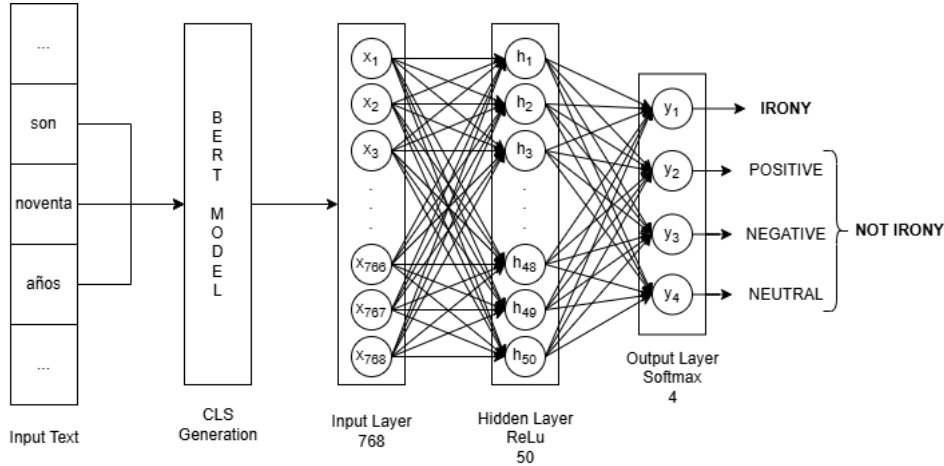


Figure 2: Architecture of the BERT-based Classification Pipeline.

The training process was configured to run for a maximum of 1500 epochs. However, in practice, training was automatically halted when the validation loss began to diverge significantly from the training loss. This mechanism was employed to prevent overfitting.

#### 4.4 Experiments

The experiments were evaluated in three main phases, each designed to assess different strategies for irony detection. The first phase establishes **baselines** using the *PRIMARY* dataset (see Table 1). This phase consists of two separate evaluations: first, GPT-4o is used to directly classify the samples based on a prompt that provides context, defines irony, outlines exceptions, and specifies the task, as detailed in Appendix B. The initial prompt-based tagging made by GPT-4o was followed by independent processing with the BERT-based Classification Pipeline (see Figure 2). These baselines served as reference points to compare the impact of dataset enhancements and augmentations.

In the second phase, we tested the hypothesis that GPT-4o can **enhance** the original text to facilitate classification. In this experiment, GPT-4o enriched the emotional and contextual features of the historical texts, generating the *ENHANCED* dataset (see Table 1). This dataset was then processed using the BERT-based Classification Pipeline to assess whether the additional contextual and emo-

tional cues improved classification performance.

The final phase involved **augmenting** the dataset through a semi-automated annotation process. GPT-4o classified entries in the new samples, providing labels and detailed justifications. Human experts reviewed these automatic annotations to ensure accuracy and preserve the historical value of the dataset. The newly verified entries were integrated into the *PRIMARY* dataset, generating the *AUGMENTED* dataset (see Table 1), which was subsequently processed using the BERT-based Classification Pipeline. In this final phase, only the top three performing BERT-like models from the earlier experiments were used for classification.

After each phase, the neural network results were evaluated using precision, recall, accuracy, and F1 score metrics. Testing was conducted separately for both binary and multi-class classifications.

## 5 Results

### 5.1 Baselines

The results obtained using GPT-4o with the *PRIMARY* dataset are presented in Table 2.

Model	Category	Precision	Recall	F1 Score	Accuracy
Base GPT 4o -Prompt	IRONY	0.24	0.80	0.37	0.39
	NEGATIVE	0.55	0.33	0.41	
	NEUTRAL	0.44	0.76	0.55	
	POSITIVE	0.86	0.01	0.03	
	W. AVG	0.60	0.39	0.31	

Table 2: Results for the multi-class classification task using GPT-4o with the prompt specified in Appendix B.

The results indicate that the 'IRONY' class achieved a precision of only 0.24, with an overall accuracy of 0.39. Thus, relying solely on GPT-4o models and prompting is not a viable approach for classifying historical Spanish texts.

Model	Category	Precision	Recall	F1 Score	Accuracy
Base GPT 4o -Prompt	IRONY	0.24	0.80	0.37	0.72
	NOT IRONY	0.97	0.71	0.82	
	AVG	0.60	0.75	0.59	

Table 3: Results for binary classification task using GPT-4o with the prompt specified in Appendix B.

The 'NOT IRONY' class showed improvement in the binary classification scenario (see Table 3). Although the model performed better when detecting non-ironic situations.

Next, we discuss the results obtained using the complete BERT-based Classification Pipeline described in the methodology, which includes contextual embeddings from BERT-family encoder models. The best-performing encoder results are presented in Table 4, with a comprehensive list of all tested encoders in Appendix C.1.

Model	Category	Precision	Recall	F1 Score	Accuracy
dccuchile/bert-base-spanish-wwm-cased	IRONY	0.61	0.47	0.53	0.66
	NEGATIVE	0.60	0.62	0.61	
	NEUTRAL	0.72	0.66	0.69	
	POSITIVE	0.66	0.75	0.70	
	W. AVG	0.66	0.66	0.65	

Table 4: Results of the BERT-based Classification Pipeline. The table presents only the best-performing encoder model for the multi-class task.

Table 4 shows an accuracy of 0.66, effectively detecting the 'IRONY' class and other sentiment categories. Although not extraordinary, these results show notable improvement over the GPT-4o classification outcomes. The results in the binary classification scenario, as shown in Table 5, corroborate this trend. The model achieved near-perfect classification for the 'NOT IRONY' category. However, these results should be interpreted cautiously due to the potential bias from the underrepresentation of the "IRONY" class in the PRIMARY dataset. Additionally, significant room for improvement in classifying the 'IRONY' category remains. These results, obtained using the neural network with the "dccuchile/bert-base-spanish-wwm-case" encoder, serve as the benchmark for subsequent experiments.

Model	Category	Precision	Recall	F1 Score	Accuracy
dccuchile/bert-base-spanish-wwm-cased	IRONY	0.80	0.34	0.48	0.91
	NOT IRONY	0.92	0.99	0.95	
	AVG	0.86	0.66	0.72	

Table 5: Results of the BERT-based Classification Pipeline. The table presents only the best-performing encoder model for the binary task.

## 5.2 Enhancement

This section presents the results obtained using the ENHANCED dataset and the BERT-based classification pipeline. Table 6 reports the results for the multiclass scenario, while Table 7 shows the binary classification results for the best-performing encoder. Appendix C.2 shows a complete set of tabulated results.

Model	Category	Precision	Recall	F1 Score	Accuracy
beto-cased-finetuned-xix-latam	IRONY	<b>0.65</b>	<b>0.51</b>	<b>0.57</b>	0.60
	NEGATIVE	0.54	<b>0.62</b>	<b>0.58</b>	
	NEUTRAL	0.69	0.53	0.60	
	POSITIVE	0.58	0.69	0.63	
	W. AVG	0.61	0.60	0.60	

Table 6: Results of the BERT-based Classification Pipeline on the ENHANCED dataset. The table presents only the best-performing encoder model for the multi-class task.

In the multiclass scenario, while the 'IRONY' class slightly improved over the baseline, the overall performance remained similar. The 'IRONY' classification exhibited greater reliability in binary classification, without a clear improvement over the baseline. These results may be influenced by the unique characteristics of historical texts and the challenges the GPT-4o model faces in capturing deeper emotional and contextual cues.

Model	Category	Precision	Recall	F1 Score	Accuracy
dccuchile/bert-base-spanish-wwm-uncased	IRONY	0.65	<b>0.32</b>	<b>0.43</b>	0.90
	NOT IRONY	0.92	0.98	0.95	
	AVG	0.78	<b>0.65</b>	<b>0.69</b>	

Table 7: Results of the BERT-based Classification Pipeline on the ENHANCED dataset. The table presents only the best-performing encoder model for the binary task.

## 5.3 Augmentation

The following tables present the results obtained when experimenting with the AUGMENTED dataset, using semi-automatically annotated data. As mentioned previously, only the top three encoder models, according to 'IRONY' class metrics, were considered in this experiment.

Model	Category	Precision	Recall	F1 Score	Accuracy
dccuchile/bert-base-spanish-wwm-uncased	IRONY	<b>0.65</b>	<b>0.72</b>	<b>0.68</b>	0.63
	NEGATIVE	0.53	<b>0.59</b>	0.56	
	NEUTRAL	0.68	<b>0.63</b>	0.65	
	POSITIVE	<b>0.66</b>	0.61	0.64	
dccuchile/bert-base-spanish-wwm-cased	W. AVG	<b>0.64</b>	0.63	0.63	0.61
	IRONY	<b>0.65</b>	<b>0.67</b>	<b>0.66</b>	
	NEGATIVE	0.51	0.56	0.53	
	NEUTRAL	0.67	0.61	0.64	
beto-cased-finetuned-xix-latam	POSITIVE	0.62	0.61	0.62	0.61
	W. AVG	0.59	0.59	0.59	
	IRONY	<b>0.70</b>	<b>0.66</b>	<b>0.68</b>	
	NEGATIVE	0.50	0.55	0.52	
	NEUTRAL	0.64	<b>0.64</b>	<b>0.64</b>	
	POSITIVE	<b>0.62</b>	0.60	0.61	
	W. AVG	0.62	0.61	0.62	

Table 8: Results of the BERT-based Classification Pipeline on the AUGMENTED dataset for the multi-class task.

Unlike in previous experiments, the results presented in Table 8 show more promising improvements for the IRONY class. While the precision metric was similar to that from Table 7, recall was significantly improved. This indicates that with the training data, the model can detect irony in more cases without compromising precision.

Model	Category	Precision	Recall	F1 Score	Accuracy
dccuchile/bert-base-spanish-wwm-uncased	IRONY	0.65	<b>0.62</b>	<b>0.64</b>	0.85
	NOT IRONY	0.90	0.91	0.91	
	AVG	0.78	<b>0.77</b>	<b>0.77</b>	
dccuchile/bert-base-spanish-wwm-cased	IRONY	0.72	<b>0.70</b>	<b>0.71</b>	0.88
	NOT IRONY	<b>0.92</b>	0.93	0.93	
	AVG	0.82	<b>0.81</b>	<b>0.82</b>	
beto-cased-finetuned-xix-latam	IRONY	0.68	<b>0.67</b>	<b>0.67</b>	0.87
	NOT IRONY	<b>0.91</b>	0.92	0.91	
	AVG	0.80	<b>0.79</b>	<b>0.79</b>	

Table 9: Results of the BERT-based Classification Pipeline on the AUGMENTED dataset for the binary task.

In the binary classification approach, the results in Table 9 showed substantial improvement in the typical precision increase compared to multi-class experiments and in addressing the usual problem of low recall. The model dccuchile/bert-base-spanish-wwm-cased demonstrates an F1 score for the 'IRONY' class above any other obtained in the experiments.

## 6 Discussion

As expected, the baseline using GPT-4o exclusively for classification demonstrated the weakest performance among the models. Its accuracy and recall metrics were significantly lower than those based on our BERT-based pipeline and data augmentation techniques, with accuracy values of 0.39 for multi-class and 0.72 for binary classification. This was particularly evident in multi-class classification, where irony detection proved especially challenging. These results highlight the limitations of

relying solely on GPT-4o for nuanced text analysis tasks.

The binary classification approach delivered superior overall performance, with accuracy values approaching one hundred percent. Simplifying the task to binary classification (ironic vs. non-ironic) improved generalization and precision across most configurations. However, the setup consistently struggled with recall for the irony class, which ranged from 0.09 to a maximum of 0.34 in baseline configurations. This indicates that while the model accurately predicted non-ironic cases, it frequently failed to recognize ironic instances.

The results of the semi-automated annotation process emphasize the valuable role of human verification in supporting automated annotation methods. The evaluation through human inspection of GPT-4o suggestions for tagging irony revealed some differences between machine-generated tags and human evaluations, as shown in Table 10. While GPT-4o marked 73.6% of entries as ironic, human evaluators assigned this tag to only 53.1% of entries. Additionally, there was a notable difference in the tagging of negative, positive, and neutral sentiments, with humans detecting more negative sentiments and some positive sentiments that GPT-4o overlooked. GPT-4o struggled with some entries with low-quality OCR transcriptions, misleading the model to produce hallucinations. These entries were introduced in the 'unreadable' category—1.7% and were excluded from the final dataset.

Tag	GPT-4o Tag	Human Tag
<b>Irony</b>	73.6%	53.1%
<b>Negative</b>	3.9%	13.1%
<b>Positive</b>	0%	2.9%
<b>Neutral</b>	22.6%	29.1%
<b>Unreadable</b>	-	1.7%

Table 10: Comparison of GPT-4o and Human Tags.

The disparity in positive sentiment detection highlights GPT-4o's limitation in contextualizing historical nuances. For instance, the model struggled to discern an author's potentially positive intent mostly because, during that period, poetic language was often employed to praise individuals or concepts, which was not indicative of irony. However, in modern contexts, such excessive praise, particularly concerning politics, might typically be interpreted as ironic, highlighting the differing interpretations across eras. These findings further illustrate GPT-4o's cultural and historical biases,

suggesting that while these models provide substantial assistance, human expertise remains essential for accurate sentiment analysis in complex historical datasets.

The process facilitated the generation of classifications by processing a set of previously untagged entries with a higher likelihood of irony using a tailored GPT-4o prompt. The GPT-4o automatic classifications and the detailed justifications were reviewed by human experts for accuracy. Confirmed annotations were integrated into the PRIMARY dataset, enriching and balancing it. This effectively addressed class imbalance as shown in Table 11, a problematic limitation in earlier experiments, while expanding the dataset with additional examples of ironic content.

Category	Primary	Augmented
<b>Irony</b>	10.68%	22.40%
<b>Negative</b>	25.64%	22.32%
<b>Neutral</b>	28.89%	29.12%
<b>Positive</b>	34.78%	26.16%

Table 11: *Classes Distribution in the PRIMARY and the AUGMENTED datasets.*

For multi-class classification, the *AUGMENTED* dataset led to higher recall values for irony detection, suggesting that the semi-automated annotation strategy contributed positively. However, the performance of non-ironic classes showed slight reductions, highlighting the trade-offs involved in emphasizing irony-related signals. In binary classification, combining the semi-automated annotation process with the dccuchile/bert-base-spanish-wwm-cased model yielded strong results, with precision and recall values of 0.70 and 0.93, respectively. Notably, recall improved significantly, increasing by up to 0.50 compared to previous experiments. These results indicate that the semi-automated process helped expand the dataset and mitigate imbalance, enhancing irony detection<sup>2</sup>.

Interestingly, the *ENHANCED* dataset, designed for emotional and contextual enrichment, while effectively sharpening certain sentiment cues, did not result in significant gains in irony detection over the original dataset. This finding indicates that enhancing emotional intensity alone does not necessarily capture the subtleties of historical irony. For example, GPT-4o struggled to identify irony

<sup>2</sup>The model is available at <https://huggingface.co/Flaglab/latam-xix-irony>. Detailed processing steps can be found at <https://github.com/historicalink/irony-detection>



Figure 3: An example of a cartoon depicting the ironic situation lived during the time of *porfiriato* in Mexico. El Hijo del Ahuizote, Mexico. Dec 7th, 1890. From the cartoons exhibition at Museo del Estanquillo in 2024.

when applying enhancement techniques to a sentence expressing political contradictions during the *Porfiriato* in Mexico<sup>3</sup>, primarily due to a lack of historical context. As depicted in Figure 3, Porfirio Díaz, who initially opposed re-elections citing constitutional violations, ironically remained in power for 35 years. Although enhancing has been successful in previous works (Lin et al., 2024), it did not effectively capture the specific cultural and historical features present in Latin-American historical texts. This highlights the cultural bias embedded in commercial models such as GPT-4o, emphasizing the need for more tailored approaches when working with historical and culturally nuanced materials.

Overall, the semi-automated annotation process, especially in binary classification, achieved the best performance for irony detection. This approach’s ability to expand the dataset and address class im-

<sup>3</sup>The translated sentence read: ‘Come, we said with enthusiasm, he is the one who will put us on the horns of the moon, with his respect for the law, his pure patriotism, and his famed honesty. You will see what Mr. Porfirio can do with the Tuxtepec Plan. We were already quite satisfied with our man in power, and everything was set and of good quality.’ The original text in Spanish is: ‘Ven, dijimos con entusiasmo, es el que nos va a poner en los cuernos de la luna, con su respeto a la ley, y su puro patriotismo y su mentada honradez. Verán lo que es D. Porfirio con el plan de Tuxtepecl. Estábamos ya muy anchos con nuestro hombre en el poder y ya con toda la cosa muy lista y de buen jaéz’



balance marks the best approach we found for this task.

Despite these advances, irony detection continues to present significant challenges due to its inherent complexity. Historical texts bring additional layers of difficulty with their unique linguistic and cultural references. Future research should focus on refining domain-specific prompts, evaluating alternative models, and developing increasingly automated architectures, including agent-based workflows that systematically incorporate historical context to enhance irony detection accuracy. Continued efforts to expand and enrich historical datasets will contribute to more reliable and generalized methods for irony detection and sentiment analysis across diverse cultural and historical contexts, ultimately enriching both humanities scholarship and the capabilities of large LLMs.

## 7 Acknowledgements

We thank the two anonymous NAACL 2025 NLP4DH conference reviewers for their helpful feedback and suggestions.

## 8 Limitations

Although five different models were employed in the classification experiments, a notable limitation remains the reliance on GPT-4o for the initial dataset enhancement and augmentation steps. While this reliance does not compromise the significance and robustness of our findings, it is costly and restricts scalability. Future research could benefit from exploring alternative models and systematically comparing their effectiveness in semi-automated data augmentation tasks, aiming to identify options that are both accessible and cost-effective for broader implementation.

## References

- Clint Burfoot and Timothy Baldwin. 2009. [Automatic satire detection: Are you having a laugh?](#) In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 161–164, Suntec, Singapore. Association for Computational Linguistics.
- Herbert L. Colston. 2017. *Irony and Sarcasm*. Routledge.
- Roberto I. González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. [Identifying sarcasm in twitter: A closer look](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Cynthia Van Hee, Els Lefever, and Veronique Hoste. 2018. [Semeval-2018 task 3: Irony detection in english tweets](#). In *International Workshop on Semantic Evaluation*.
- Yu-Hsiang Huang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2017. Irony detection with attentive recurrent neural networks. In *Advances in Information Retrieval*, pages 534–540, Cham. Springer International Publishing.
- Roger Kreuz. 2020. *Irony and Sarcasm*. The MIT Press.
- Yucheng Lin, Yuhan Xia, and Yunfei Long. 2024. [Augmenting emotion features in irony detection with large language modeling](#). *Preprint*, arXiv:2404.12291.
- Laura Manrique-Gómez, Tony Montes, Arturo Rodríguez-Herrera, and Rubén Manrique. 2024. [Historical ink: 19th century latin american spanish newspaper corpus with llm ocr correction](#).
- Tony Montes, Laura Manrique-Gómez, and Rubén Manrique. 2024. [Historical ink: Semantic shift detection for 19th century spanish](#). *Preprint*, arXiv:2407.12852.
- Asli Umay Ozturk, Recep Firat Cekineli, and Pinar Karagoz. 2024. [Make satire boring again: Reducing stylistic bias of satirical corpus by utilizing generative llms](#). *Preprint*, arXiv:2412.09247.
- Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. [Sarcasm detection on twitter: A behavioral modeling approach](#). In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM '15*, page 97–106, New York, NY, USA. Association for Computing Machinery.
- Yafeng Ren, Zilin Wang, Qiong Peng, and Donghong Ji. 2023. [A knowledge-augmented neural network model for sarcasm detection](#). *Information Processing Management*, 60(6):103521.
- Ellen Riloff, Ashequl Qadir, Praful Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. [Sarcasm as contrast between a positive sentiment and negative situation](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. [Fake news or truth? using satirical cues to detect potentially misleading news](#). In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, San Diego, California. Association for Computational Linguistics.
- Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. [Human-llm collaborative annotation through effective verification of llm labels](#). In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24*, New York, NY, USA. Association for Computing Machinery.

## Appendix A Prompt used to enhance the sentiment and context of the dataset

"Expande este texto de manera de que mantenga su significado original, se debe hacer mucho énfasis en la carga emocional del texto, de manera que la versión final obtenida permita una mejor identificación del sentimiento general del mismo. Únicamente responde con el texto expandido y esfuérate por conservar la sintaxis y morfología del español latinoamericano del siglo 19"

### A.1 Input and Output Examples

**Input text:** *"habla de noventa y nueve años, y noventa y nueve mas de transcurso para que el gobierno pueda entrar en posesion de el, es decir, dos siglos, en cuyo tiempo, probablemente, cuanto a la tarifa de trasportes, q discrecion del sr. cisneros establecerla durante el tiempo de la construccion del camino, y diez años mas y mientras que antes se obligaba a cobrar solamente la tercera parte de la tarifa por las tropas del gobierno y los efectos del servicio publico, posteriormente exige la mitad de ella. una tarifa arbitraria y no establecida con previo acuerdo para ejercer toda la extorsion que se quiera. aun faltan por analizar algunos otros puntos importantes pero por temor de cansa vuestra atencion, y por haber sido ya, repetiperiodica, prescindimos de ocuparnos de ellos. es 6 todas luces de vital importancia la continuacion del ferrocarril de girardot mas, en nuestro humilde concepto, es preferi. ble a un contrato sumamente gravoso, que la verifique el gobierno por administracion pues que a la vista esta la mejora de la parte de la direccion, y, porque de ello es tambien miguel guerrero y la acertada direccion del inteligente ingeniero d."*

**Output text:** *"Habla de noventa y nueve años, y noventa y nueve más de transcurso para que el gobierno pueda entrar en posesión de él, es decir, dos siglos completos, en cuyo tiempo, probablemente, en cuanto a la tarifa de transportes, queda a discreción del señor Cisneros establecerla durante el tiempo de la construcción del camino, y diez años más. Y mientras que antes se obligaba a cobrar solamente la tercera parte de la tarifa por las tropas del gobierno y los efectos del servicio público, posteriormente se exige la mitad de ella. Una tarifa arbitraria y no establecida con previo acuerdo para ejercer toda la extorsión que se quiera. Aún faltan por analizar algunos otros puntos importantes,*

*pero por temor de cansar vuestra atención, y por haber sido ya repetidos periódicamente, prescindimos de ocuparnos de ellos. Es a todas luces de vital importancia la continuación del ferrocarril de Girardot, mas, en nuestro humilde concepto, es preferible a un contrato sumamente gravoso, que la verifique el gobierno por administración, pues que a la vista está; la mejora de la parte de la dirección, y, porque de ello es también Miguel Guerrero y la acertada dirección del inteligente ingeniero don."*

## Appendix B Prompt used for the semi-automatic labeling process (Translated from Spanish <sup>4</sup>)

"A text in Spanish from 19th-century Latin American press will be received. This text may or may not contain some form of irony, meaning it fulfills one of the following situations: - It presents a contradiction between the reality described in the context and what is said. - It presents a contradiction between the historical reality of 19th-century Latin America and what is said. - It presents a contradiction between what is said and the tone in which it is said (based on the use of capitalization and punctuation).

This text may contain a critique of a contradictory political or social situation that occurred, but it is not necessarily ironic; it could be a negative political opinion. The text may also contain contradictory comparisons or hyperboles, but it is not necessarily ironic; it could be an expression with poetic language. For it to be an ironic contradiction, there must be an intent of humor or mockery in the text, not merely an intent of political critique or contradiction or an intent of figurative or poetic description.

The task is to identify whether there is irony present in any contradiction in the text and explain why it is contradictory and what the author's intention is. If no irony is detected, you must explain why it is not irony and indicate whether the text has a positive, negative, or neutral sentiment. The response must begin with one of these 4 words based on the inference: "IRONY," "POSITIVE," "NEGATIVE," "NEUTRAL," written in single quotation marks ("). Next, the explanation of what the contradiction is (if ironic) or why it is not irony must

<sup>4</sup>Prompts in Spanish can be found at <https://github.com/historicalink/ironydetection>.

be added between asterisks (\*). You must not include anything beyond what is requested. The final response must not exceed 500 words, including the description."

## B.1 Input and Output Examples

**Input Text:** The same as in the appendix A.

**Output text:** "'NEGATIVE' \*The text provides a negative critique of the government's management and the arbitrary transportation fee imposed, but there is no comedic or mocking intention indicating irony. The contradiction mentioned constitutes a direct and serious criticism of the political and administrative situation, without elements of irony.\*"

## Appendix C Tables

### C.1 Baselines Results

Model	Category	Precision	Recall	F1 Score	Accuracy
bert-base-uncased	IRONY	0.28	0.11	0.15	0.54
	NEGATIVE	0.47	0.60	0.53	
	NEUTRAL	0.60	0.54	0.57	
	POSITIVE	0.58	0.63	0.60	
	W. AVG	0.52	0.54	0.52	
bert-base-multilingual-uncased	IRONY	0.52	0.34	0.41	0.59
	NEGATIVE	0.49	0.52	0.50	
	NEUTRAL	0.68	0.60	0.64	
	POSITIVE	0.61	0.70	0.65	
	W. AVG	0.59	0.59	0.58	
dccuchile/bert-base-spanish-wwm-uncased	IRONY	0.54	0.43	0.48	0.64
	NEGATIVE	0.56	0.59	0.58	
	NEUTRAL	0.74	0.63	0.68	
	POSITIVE	0.65	0.76	0.70	
	W. AVG	0.64	0.64	0.64	
dccuchile/bert-base-spanish-wwm-cased	IRONY	0.61	0.47	0.53	0.66
	NEGATIVE	0.60	0.62	0.61	
	NEUTRAL	0.72	0.66	0.69	
	POSITIVE	0.66	0.75	0.70	
	W. AVG	0.66	0.66	0.65	
beto-cased-finetuned-xix-latam	IRONY	0.59	0.47	0.52	0.63
	NEGATIVE	0.56	0.60	0.58	
	NEUTRAL	0.71	0.59	0.64	
	POSITIVE	0.62	0.74	0.68	
	W. AVG	0.63	0.63	0.63	

Table 12: Baseline Results for multi-class classification

Model	Category	Precision	Recall	F1 Score	Accuracy
bert-base-uncased	IRONY	0.80	0.09	0.15	0.89
	NOT IRONY	0.89	1.00	0.94	
	AVG	0.85	0.54	0.55	
bert-base-multilingual-uncased	IRONY	0.75	0.32	0.45	0.91
	NOT IRONY	0.92	0.99	0.95	
	AVG	0.83	0.65	0.70	
dccuchile/bert-base-spanish-wwm-uncased	IRONY	0.79	0.23	0.36	0.91
	NOT IRONY	0.91	0.99	0.95	
	GENERAL	0.85	0.61	0.65	
dccuchile/bert-base-spanish-wwm-cased	IRONY	0.80	0.34	0.48	0.91
	NOT IRONY	0.92	0.99	0.95	
	AVG	0.86	0.66	0.72	
beto-cased-finetuned-xix-latam	IRONY	0.71	0.21	0.33	0.90
	NOT IRONY	0.91	0.99	0.95	
	AVG	0.81	0.60	0.64	

Table 13: Baseline Results for binary classification

## C.2 Data Enhancement Results

Model	Category	Precision	Recall	F1 Score	Accuracy
bert-base-uncased	IRONY	0.65	0.23	0.34	0.53
	NEGATIVE	0.47	0.64	0.54	
	NEUTRAL	0.62	0.50	0.55	
	POSITIVE	0.52	0.58	0.55	
	W. AVG	0.55	0.53	0.52	
bert-base-multilingual-uncased	IRONY	0.67	0.21	0.32	0.57
	NEGATIVE	0.50	0.60	0.55	
	NEUTRAL	0.65	0.59	0.62	
	POSITIVE	0.55	0.65	0.60	
	W. AVG	0.58	0.57	0.56	
dccuchile/bert-base-spanish-wwm-uncased	IRONY	0.54	0.40	0.46	0.63
	NEGATIVE	0.59	0.64	0.61	
	NEUTRAL	0.72	0.59	0.65	
	POSITIVE	0.61	0.73	0.66	
	W. AVG	0.63	0.63	0.63	
dccuchile/bert-base-spanish-wwm-cased	IRONY	0.53	0.45	0.48	0.64
	NEGATIVE	0.62	0.66	0.64	
	NEUTRAL	0.72	0.55	0.63	
	POSITIVE	0.62	0.76	0.68	
	W. AVG	0.62	0.61	0.61	
beto-cased-finetuned-xix-latam	IRONY	0.65	0.51	0.57	0.60
	NEGATIVE	0.54	0.62	0.58	
	NEUTRAL	0.69	0.53	0.60	
	POSITIVE	0.58	0.69	0.63	
	W. AVG	0.61	0.60	0.60	

Table 14: Enhancement Results. Multi-class classification tasks

Model	Category	Precision	Recall	F1 Score	Accuracy
bert-base-uncased	IRONY	0.70	0.15	0.25	0.90
	NOT IRONY	0.90	0.99	0.94	
	AVG	0.80	0.57	0.59	
bert-base-multilingual-uncased	IRONY	0.62	0.21	0.32	0.90
	NOT IRONY	0.91	0.98	0.94	
	AVG	0.77	0.60	0.63	
dccuchile/bert-base-spanish-wwm-uncased	IRONY	0.65	0.32	0.43	0.90
	NOT IRONY	0.92	0.98	0.95	
	AVG	0.78	0.65	0.69	
dccuchile/bert-base-spanish-wwm-cased	IRONY	0.58	0.15	0.24	0.89
	NOT IRONY	0.90	0.99	0.94	
	AVG	0.74	0.57	0.59	
beto-cased-finetuned-xix-latam	IRONY	0.50	0.17	0.25	0.89
	NOT IRONY	0.90	0.98	0.94	
	AVG	0.70	0.57	0.60	

Table 15: Enhancement Results. Binary classification tasks