ELiRF-UPV at SatiSPeech-IberLEF 2025: Multimodal Speech-text Satire Recognition in Spanish

Alejandro Joaquín Barceló Milkova[1,†], Andreu Casamayor Segarra[2,†], Vicent Ahuir[2,†] and María José Castro-Bleda[2,3,*,†]

[1]*Department of Computer Systems and Computation, Universitat Politècnica de València, Camí de Vera s/n, València, 46020, Spain*
[2]*VRAIN: Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, Camí de Vera s/n, València, 46020, Spain*
[3]*ValgrAI: Valencian Graduate School and Research Network of Artificial Intelligence, Universitat Politècnica de València, Camí de Vera s/n, València, 46020, Spain*

Abstract

This paper describes our participation in the SatiSPeech shared task at IberLEF 2025, which focuses on the automatic detection of satirical content in Spanish using both textual and acoustic modalities. The task is divided into two subtasks: satire detection from text alone, and a more challenging multimodal satire detection task combining speech and its transcription. We developed several systems leveraging pre-trained transformer-based language models and instruction-tuned large language models, applying both fine-tuning and few-shot prompting strategies. For the multimodal task, we designed an approach that integrates acoustic and textual features to capture the nuanced cues characteristic of satirical discourse. Our models were trained and evaluated using the newly introduced SatirA dataset, which includes approximately 25 hours of labeled speech and corresponding transcriptions. The results demonstrate the effectiveness of our methods, achieving first place in the text-only task and second place in the multimodal task. These findings highlight the feasibility of applying multimodal learning to complex language understanding tasks such as satire detection and underscore the value of combining linguistic and prosodic cues for improved performance.

Keywords

Natural Language Processing, Transformers-based Models, Large Language Models, Multimodal Satire Detection, Spanish Satirical Content

1. Introduction

Satire is a sophisticated and multifaceted communication form that intertwines humor, irony, and criticism to expose or mock social, political, or cultural phenomena. Unlike direct or straightforward humor, satire often hinges on subtle linguistic and contextual cues (such as tone, exaggeration, and implied meaning), making it difficult to detect, even for humans. This complexity is further magnified in multimodal scenarios, where meaning is distributed across text, speech, and prosodic elements [1]. Misinterpreting satirical content can result in confusion or unintended consequences, particularly when the subject matter is sensitive or highly critical. However, developing systems capable of understanding satire is increasingly relevant for applications such as media monitoring [2], misinformation detection, or political discourse analysis. Leveraging multimodal approaches that integrate textual, acoustic, and contextual features offers a promising path toward modeling the nuanced, context-dependent nature of satirical discourse.

Although several previous works have addressed the detection of satire in text [3, 4], fewer have explored multimodal approaches that incorporate audiovisual information along with text [5]. This

IberLEF 2025, September 2025, Zaragoza, Spain
*Corresponding author.
†These authors contributed equally.
ajbarmil@upv.edu.es (A. J. Barceló Milkova); ancase3@upv.es (A. Casamayor Segarra); vahuir@dsic.upv.es (V. Ahuir); mcastro@dsic.upv.es (M. J. Castro-Bleda)
0009-0007-4740-6461 (A. J. Barceló Milkova); 0009-0003-6000-3828 (A. Casamayor Segarra); 0000-0001-5636-651X (V. Ahuir); 0000-0003-1001-8258 (M. J. Castro-Bleda)

gap is especially evident in low-resource languages such as Spanish, where annotated datasets and multimodal benchmarks remain scarce [6].

In this work, we address the task of automatic satire detection in Spanish as a supervised classification problem. We have explored a range of models, including traditional machine learning approaches such as Support Vector Machines (SVMs) and transformer-based deep learning architectures [7]. In addition, we evaluate several large language models (LLMs) under few-shot prompting scenarios to assess their few-shot generalization capabilities. Our study is grounded on a novel multimodal dataset which contains spoken content and its corresponding transcription, labeled as satirical or non-satirical. We investigate how textual, acoustic, and contextual features can be effectively combined to improve classification performance.

The rest of the paper is structured as follows. In Sections 2 and 3, we introduce the task and describe the dataset and evaluation metrics. Section 4 presents our multimodal satire classification system. In Section 5, we report experimental results and conduct an error analysis. Finally, Section 6 concludes the paper and discusses avenues for future work.

2. Task description

The SatiSPeech 2025 shared task [8] at IberLEF 2025 [9] aims to address these challenges by investigating the detection of satire in Spanish through a multimodal lens, combining textual and audio information. Satirical communication often draws on irony, double meanings, and culturally grounded references, which pose significant challenges for automatic systems [10]. The task is framed as a binary classification problem: distinguishing between satirical and non-satirical content, using both textual and audio-based cues. Central challenges include identifying which features most reliably indicate satire (e.g., linguistic structures, prosody, intonation) and overcoming the scarcity of multimodal datasets that reflect the diversity and authenticity of real-world satire. Through this initiative, the task seeks to advance the capabilities of multimodal classification models and promote novel approaches to satire recognition across languages and modalities.

Despite recent progress, multimodal satire detection remains relatively underexplored, and most existing research focuses on text or visual content. For example, [5] introduced a system that combines text and images to detect satire in news headlines, showing that multimodal models outperform unimodal counterparts. Similarly, [11] evaluated the effectiveness of large language models in identifying satirical news in Brazilian Portuguese, reporting promising results and valuable insights into the processing of satirical language characteristics. In the Spanish-language context, research has only recently begun to gain momentum. The development of the *SatiCorpus 2021* [6] marked a significant step, offering a labeled dataset of satirical and non-satirical texts evaluated using deep learning models and linguistic features.

The shared task is organized into two subtasks: Text-based Satire Detection and Multimodal Satire Detection. The first task focuses on determining whether a given text expresses satirical or non-satirical content, relying solely on linguistic and semantic cues present in the written modality. The second task extends this framework by incorporating audio data alongside text. Multimodal Satire Detection aims to leverage both spoken language and its textual transcription to assess whether an audio-text pair conveys satire, capturing prosodic, acoustic, and contextual signals that may be critical to identifying satirical intent.

3. The dataset and the evaluation metrics

To address the challenges of satire detection in a multimodal framework, a dedicated dataset was curated, combining both textual and audio information. The data were sourced from a wide selection of YouTube channels, including satirical programs such as El Intermedio, Zapeando, Homo-Zapping, and El Mundo Today, as well as non-satirical news programs like Antena 3 Noticias, El Mundo, and BBC

News. This variety ensures a rich representation of regional Spanish dialects and stylistic diversity, capturing the linguistic and cultural nuances necessary for the task.

The dataset construction process involved video extraction, followed by segmentation into manageable short audio clips using a speaker diarization tool [12, 13]. To maintain consistency, only segments of no longer than 25 seconds were retained. These clips were then transcribed using Whisper [14], allowing the generation of high-quality textual data aligned with the audio.

Annotation was carried out through a semi-supervised strategy, combining automatic classification techniques with manual validation by three expert annotators. This hybrid approach ensured both efficiency and accuracy. A subsequent manual review by task organizers was conducted to refine the labels and ensure the quality of the annotation. The final dataset features a broad spectrum of Spanish-speaking regions, promoting linguistic diversity and minimizing potential regional biases. Some examples from the dataset are shown in Table 7 in Section A.

The *SatirA* dataset consists of approximately 25 hours of labeled audio and the corresponding transcriptions. For the shared task, a curated subset of 8000 multimodal samples was selected and divided into Training (6000 samples) and Test (2000 samples) sets. Each sample contains a unique identifier, the speech signal, its transcription, and a binary label indicating whether the content is *satirical* or *non-satirical*.

Table 1
SatirA dataset sample distribution along the sets. It is also shown the percentage of satire and no-satire samples on Training.

	Total		Satire		No-satire	
Set	**Number**	**(% Total)**	**Number**	**(% Set)**	**Number**	**(% Set)**
Training	6000	(75.00%)	2832	(47.20%)	3168	(52.80%)
Test	2000	(25.00%)	-	-	-	-
Total	8000	(100%)				

Table 1 shows the descriptive statistics of the dataset used in the shared task. We can appreciate an imbalance towards `no-satire` labeled samples in the training set. Regarding the test set, we could not provide the distribution of satirical and non-satirical samples for that partition since the participants could not access the test labels.

For evaluating the system, the Macro version of Precision, Recall, and F1-score was measured. The systems were ranked by the F1-score, that is the harmonic mean of the Precision and Recall.

4. Developed Systems

4.1. Working partitions

To develop, tune, and validate our systems before submitting them to the competition, we divided the original Training set into two stratified partitions: 5100 samples for system training (Train) and 900 samples for tuning and validation of the different approaches (Dev). Depending on the task, we used only the text in the samples (Task 1: Text Satire Detection) or text+audio (Task 2: Multimodal Satire Detection) to develop the different systems. Table 2 shows the distribution of satirical and non-satirical samples within the two partitions. The partitions are the same for tasks 1 and 2.

Table 2
Distribution of samples in Train and Dev stratified partitions created from the original Training set.

	Total		Satire		No-satire	
Partition	**Number**	**(% Total)**	**Number**	**(% Part.)**	**Number**	**(% Part.)**
Train	5100	(85.00%)	2407	(47.20%)	2693	(52.80%)
Dev	900	(15.00%)	425	(47.22%)	475	(52.78%)

4.2. Task 1: Text Satire Detection

For the first downstream task, we developed some text satire detection models using different transformer-based language models and the Low-Rank Adaptation technique (LoRA) [15]. We employed LoRA since the number of training samples is scarce, and LoRA will reduce the probability of overfitting compared to regular fine-tuning. We targeted LoRA, the query and value modules, with a fixed dropout set to 0.05. We selected different base models, monolingual and multilingual, publicly available at the Hugging Face Hub [16] that were pre-trained for the Spanish language.

For this approach, we developed the following 4 systems:

- **System `T1-1`:** The first system was based on the monolingual Spanish model BETO [17] (https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased). BETO is a BERT model pre-trained on large-scale Spanish corpora.
- **System `T1-2`:** The second system was based on the monolingual Spanish model BERTIN [18] (https://huggingface.co/bertin-project/bertin-roberta-base-spanish). BERTIN is a RoBERTA model pre-trained on Spanish texts.
- **System `T1-3`:** This system was based on the multilingual model XLM-RoBERTa [19] in its `base` version (https://huggingface.co/FacebookAI/xlm-roberta-base). XLM-RoBERTa is a multilingual version of RoBERTa, pre-trained on data containing 100 languages, including Spanish.
- **System `T1-4`:** System based on the large version of the monolingual Spanish model RoBERTa-BNE [20] (https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne). RoBERTA-BNE was pre-trained with data from the web crawling performed by the National Library of Spain from 2009 to 2019.

For the fine-tuned systems, we made a hyperparameter search using the *Optuna* library [21]. Hyperparameters such as the learning rate, the batch size and the rank and scaling factor from LoRA were set by a cross-validation system dividing the training set into five folds. For each model, ten trials were performed.

The search for the learning rate consisted of a log-uniform search between 1e-5 and 1e-3, while the search for the batch size was between the specific values of 8, 16, 32 and 64. For the LoRA hyperparameters, an integer search was done for both of them: the rank searching values between 8 and 32 with steps of 8, and the scaling factor searching values between 16 and 64 with steps of 16. The hyperparameter values search results found by *Optuna* for each system are described in Table 3.

Table 3
Model training configurations with LoRA parameter for systems `T1-1`, `T1-2`, `T1-3`, `T1-4`.

System	Desc.	Learning Rate	Batch Size	LoRA Rank (r)	LoRA Scaling Factor (α)
`T1-1`	BETO	1.93×10^{-4}	32	8	16
`T1-2`	BERTIN	2.27×10^{-4}	64	8	16
`T1-3`	XLM-RoBERTa	7.57×10^{-4}	32	24	16
`T1-4`	RoBERTa-BNE	1.44×10^{-4}	32	32	48

We also developed systems based on LLMs and in-context learning prompting. We followed a 6-shot approach and tried different sets of random examples from the training test. Fig. 1 shows the prompt template and an example of how the template would look instantiated.

We explored various LLMs for satire detection in text. All models were selected based on their support for Spanish and/or their capabilities for few-shot reasoning. We developed 3 systems based on LLMs and prompting:

- **System `T1-5`:** The fifth system was based on the Qwen2.5 family of LLMs [22], specifically on the 7 billion parameters instruct version (https://huggingface.co/Qwen/Qwen2.5-7B-Instruct). Qwen2.5 is a versatile family of LLMs developed by Alibaba Cloud, designed to handle a wide range of natural language processing tasks.

6-shot Prompt for Satire Classification

Tu tarea es clasificar un texto en español como 'satire' o 'no-satire'.

Definición de sátira: La sátira es un tipo de discurso que usa el humor, la ironía, la exageración o el absurdo para criticar o ridiculizar a personas, instituciones o situaciones sociales. A menudo aparenta ser serio, pero en realidad busca provocar reflexión o burla.

Instrucciones:
- Responde SOLO con 'satire' o 'no-satire'.
- Analiza el tono, el contenido y el propósito del texto.
- Considera si hay señales de exageración, ironía o crítica disfrazada.

Ejemplos:

- Texto: "¡Ey! ¡Ey! ¡Ey! Y todos hemos escuchado la frase, el que se mueve no sale en la foto..."
  Clasificación: satire
- Texto: "La Organización Mundial de la Salud advirtió que en las próximas dos décadas..."
  Clasificación: no-satire
- Texto: "Científicos de Texas logran disparar al coronavirus y agujerearle el sombrero..."
  Clasificación: satire
- Texto: "Estampa su firma para que el Parlamento se constituya el 26 de octubre..."
  Clasificación: no-satire
- Texto: "Sí, yo solo quiero decir que para la semana que viene..."
  Clasificación: satire
- Texto: "Estampa su firma para que el Parlamento se constituya el 26 de octubre..."
  Clasificación: no-satire

Ahora clasifica este texto:
Texto: "En Francia, se juega una nueva carta en favor de su política de austeridad..."
Clasificación:

Figure 1: An example of a 6-shot prompt employed for text-based satire detection. The text in black is part of the prompt template, and the text in blue will vary in each prediction petition.

- **System `T1-6`:** In this system, we employed the 8 billion parameters instruct version of the Llama-3.1 LLMs [23] (https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct). Llama 3.1 was developed by Meta and represented a significant leap in capability, scale, and multilingual support compared to previous Llama versions.
- **System `T1-7`:** In the seventh system we employed the LLM 7 billion parameters instruct version of Mistral [24] (https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3). The Mistral model was developed by the Mistral AI team, and it offers a balance between performance and efficiency, making it suitable for a wide range of applications, including those with limited computational resources.

4.3. Task 2: Multimodal Satire Detection

A multimodal approach combining both audio and transcription analysis was pursued. The goal was to obtain, for each sample, an embedding representative of its audio and a different embedding representative of its text. Then, both embeddings would be concatenated and used as the input of a classifier that performed the classification task.

To obtain the audio embeddings, we employed three different models: (1) OpenAI's *Whisper* [25] in its base version (https://openai.com/index/whisper/), (2) the base version of Meta's *Wav2Vec2* [26] models (https://huggingface.co/facebook/wav2vec2-base), and (3) Meta's *Hubert* [27] base version model (https://huggingface.co/facebook/hubert-base-ls960). Although *Wav2Vec2* and *Hubert* directly

output the desired embeddings, *Whisper* does not-as we had to extract the embeddings from the output of its encoder. For the text embeddings, we used the mean pooling over the last hidden state of the XLM-RoBERTa to extract the contextual embeddings of the text; we employed the base model version.

Then, different types of classifiers such as Multi-Layer Perceptron (MLP), Support Vector Classifier (SVC), Linear SVC (lSVC), Gradient Boosting classifier (GB), and K-Neighbors classifier (K-Ns) were trained and evaluated; all classifiers were trained using the *Scikit-learn* library for machine learning in Python [28]. A grid search using a 5-fold cross-validation was performed in order to find optimal values for the parameters of each of these classifiers.

For task 2, we developed several multimodal classification systems by varying both the audio embedding extraction methods and the classification techniques. These systems were trained and fine-tuned using the Train and Dev partitions. In total, we implemented seven distinct multimodal classifiers:

- **System** `T2-1`**:** This system extracts the audio embedding from the Whisper encoder module of the model, and the text embedding is extracted using XML-RoBERTa. These embeddings are concatenated and passed to an MLP classifier.
- **System** `T2-2`**:** For audio embedding system employs Wav2Vec2 and XML-RoBERTa for the text. These embeddings are concatenated and passed to an MLP classifier.
- **System** `T2-3`**:** For audio embedding system employs the Hubert model and the XML-RoBERTa model for the text. These embeddings are concatenated and passed to an MLP classifier.
- **System** `T2-4`**:** Same as `T2-2` (Wave2Vect2 and XML-RoBERTa), but it uses an SVC for the classification task.
- **System** `T2-5`**:** Same as `T2-2`, but it uses a Linear SVC for the classification task.
- **System** `T2-6`**:** Same as `T2-2`, but it uses Gradient Boosting for the classification task.
- **System** `T2-7`**:** Same as `T2-2`, but it uses K-Neighbors for the classification task.

We also explored the possibility of increasing the amount of data, changing the way the audio embedding was extracted, and applied noise on the audio side. We obtained 3 more systems:

- **System** `T2-8`**:** The system is identical to the `T2-4` (Wav2Vec2 and XML-RoBERTa with SVC for classification). However, it was trained with the 6000 samples available in the Training set provided by the organizers of the shared task.
- **System** `T2-9`**:** This system uses the same components as `T2-4`, however it uses attention pooling for audio embedding extraction. This system was also trained with the Training set of the shared task.
- **System** `T2-10`**:** This system is identical to the `T2-9` system. In addition to training it with the 6000 available samples, we introduce a new synthetic sample per original sample by randomly introducing volume variation, white noise, frequency masking, and time masking in the audio to improve the robustness of the system. We trained the system with 12 000 samples.

Although there was no hyperparameter optimization for the MLP-based systems, all other systems were optimized by performing a grid search on its parameters using a 5-fold cross-validation system. For SVCs, its penalty parameter (C) is chosen between $[0.1, 1, 10]$, and its gamma parameter is chosen between 'scale' or 'auto'. For linear SVCs, its penalty parameter is also chosen between the same values, its tolerance is decided between $[0.1, 0.01, 0.001]$, and its loss function is decided between 'hinge' and 'squared_hinge'. For gradient boosting classifiers, the number of estimators is decided between 100 or 300, the learning rate is chosen between 0.1 and 0.05 and its maximum depth is chosen between 3 and 5. For K-Neighbors classifiers, the number of neighbors is decided between $[3, 5, 7]$, with weights that can be uniform or up to distance, and its metric can be either Euclidean distance or Manhattan distance. The hyperparameter values search results found by the grid search for each system are described in Table 4.

Table 4
Classifier hyperparameter configuration for systems T2-4, T2-5, T2-6, T2-7.

System	Desc.	Key Parameters	Value	Additional Info
T2-4	Wav2Vec2 + XML-RoBERTa → SVC	Penalty parameter (C)	10	Gamma: scale
T2-5	Wav2Vec2 + XML-RoBERTa → lSVC	Penalty parameter (C)	0.1	Tolerance: 0.1 Loss: squared_hinge
T2-6	Wav2Vec2 + XML-RoBERTa → GB	Estimators	300	Learning rate: 0.1 Max depth: 5
T2-7	Wav2Vec2 + XML-RoBERTa → K-Ns	Neighbors	3	Weights: uniform Metric: manhattan

5. Experimental results and discussion

This section presents the results of our systems on both the development (Dev) partition (defined in Section 4.1) and the official Test set provided for the shared task.

5.1. Task 1: Text Satire Detection

Table 5 shows the results for the first task. T1-B1 is the baseline provided by the organizers, which is based on using a *TF-IDF* vectorizer to transform samples into vectors, applying a Min-Max to them, and then use them to train a linear Support Vector Classifier [8]. Systems T1-1 to T1-4 are based on fine-tuning using the LoRA technique, while T1-5 to T1-7 are prompt-based models.

Table 5
Task 1: Performance results for the Dev partition and the Test set of the shared task. The super index on our best system in Test indicates the place obtained in the contest.

System	Descript.	Dev	Test
T1-B1	Shared task baseline	—-	0.7937
T1-1	BETO	0.9463	0.8424
T1-2	BERTIN	0.9508	0.7994
T1-3	XLM-RoBERTa	0.9497	**0.8564**[1]
T1-5	Qwen2.5-7B	0.8078	0.8542
T1-6	Llama-3.1-8B	0.6800	0.5824
T1-7	Mistral-7B	**0.8875**	0.8099

Compared to the baseline system T1-B1, all of our systems, except T1-6, outperformed the base performance, indicating that modern transformer models are effective for satire detection.

Analyzing the results of the models obtained with LoRA fine-tuning, we observe that these systems have obtained the best performance on Dev compared to the prompting-based systems. However, their performance dropped significantly on the Test set, compared to the prompt-based systems. This suggests that LoRA-based models may generalize less effectively than larger, prompt-based models. Despite the drop, the best overall result was achieved by a LoRA system (T1-3), closely followed by a prompt-based system (T1-5), which interestingly performed better on the Test set than on Dev. These results led us to achieve first place in the text-based satire detection task.

Notably, although this task is conducted in Spanish texts, the monolingual models T1-1, T1-2, and T1-4 presented a lower performance on the Test set than some multilingual models (T1-3 and T1-5). This may suggest that the Test samples included foreign expressions or multilingual cues that were better handled by multilingual models.

Fig. 2 visually summarizes the results from Table 5, with systems ordered by Test performance. The performance gap between Dev and Test is more pronounced in monolingual-based systems than in multilingual ones. As stated before, the multilingual model T1-5 even improve from Dev to Test.

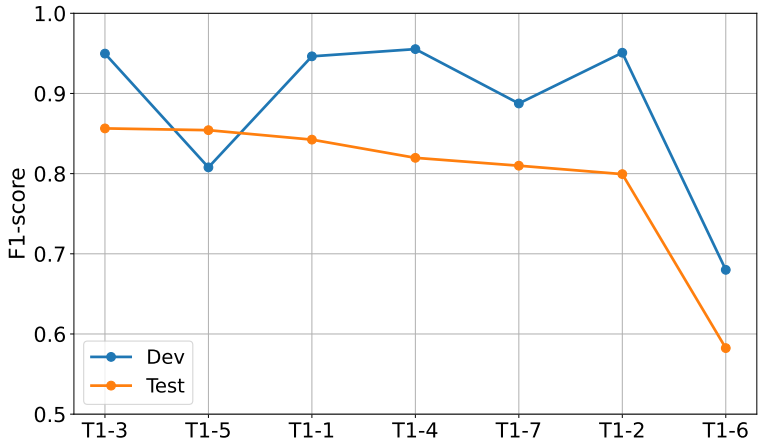

Figure 2: Evolution of the F1-Score of our systems, sorted by Test score (best to worst), for Dev and Test for Task 1.

Table 6
Task 2: Performance results for the Dev partition and the Test set of the shared task. The super index on our best system in Test indicates the place obtained in the contest.

System	Desc.	Dev	Test
T2-B1	Shared task baseline	–	0.7992
T2-1	Whisper + XML-RoBERTa → MLP	0.9678	0.8532
T2-2	Wav2Vec2 + XML-RoBERTa → MLP	0.9733	0.8550
T2-3	Hubert + XML-RoBERTa → MLP	0.9656	0.8543
T2-4	Wav2Vec2 + XML-RoBERTa → SVC	**0.9755**	0.8639
T2-5	Wav2Vec2 + XML-RoBERTa → lSVC	0.9621	0.8390
T2-6	Wav2Vec2 + XML-RoBERTa → GB	0.9656	0.8392
T2-7	Wav2Vec2 + XML-RoBERTa → K-Ns	0.9542	0.8495
T2-8	Wav2Vec2 + XML-RoBERTa → SVC (6K)	–	**0.8644**[2]
T2-9	Wav2Vec2 + XML-RoBERTa → SVC (6K+AAP)	–	0.8639
T2-10	Wav2Vec2 + XML-RoBERTa → SVC (6K+AAP+Aud.DA)	–	0.8488

5.2. Task 2: Multimodal Satire Detection

Table 6 shows the results for the second task. T2-B1 is the baseline provided by the organizers. The baseline relies on *MFCC* features extracted from the audio, concatenated with the textual features from Task 1, and using a Support Vector Classifier (SVC) for classification [8]. Systems T2-1 to T2-7 were trained using our internal Train partition (as defined in Section 4.1), while T2-8 to T2-10 were trained on the official Training set provided by the shared task organizers. Consequently, no Dev results are available for TT2-8 to TT2-10.

All our developed systems outperform the baseline system on the Test set. Among the systems trained with our internal Dev data, the combination of the Wav2Vec2 with XML-RoBERTa (system T2-2) performed better the other two audio models: Whisper (T2-1) and HuBERT (T2-3). This suggests that Wav2Vec2 provides more robust acoustic representations in this context.

When comparing classification strategies, T2-4 (SVC) outperformed T2-2 (MLP), even with the same audio and text embeddings, indicating that SVC may offer better generalization or be more stable for smaller training sets. Other classical classifiers, such as linear SVC (T2-5), Gradient Boosting (T2-6), and K-Neighbors (T2-7), showed lower performance than SVC and MLP, especially on the Test set, suggesting limited suitability for this multimodal setup.

The highest overall performance was obtained with system T2-8, trained on the full 6000 samples of the official Training set. This confirms that increasing the training data size leads to significant performance improvements, especially for high-capacity models and multimodal architectures.

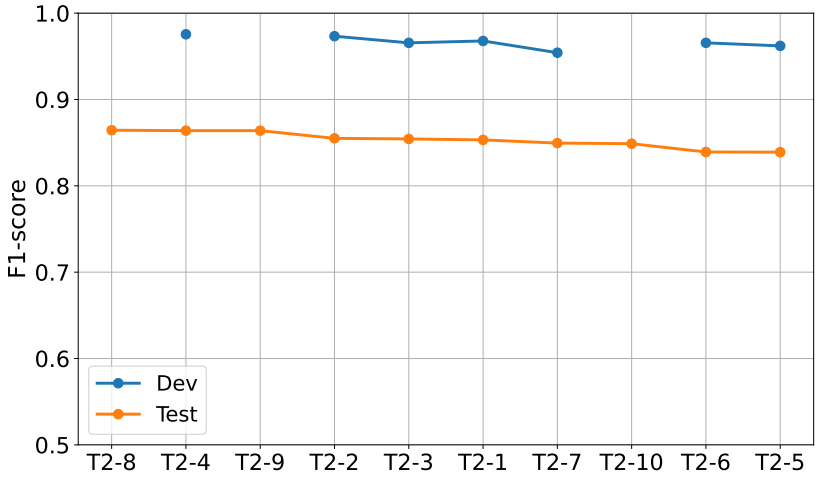

Figure 3: Evolution of the F1-Score of our systems, sorted by Test score (best to worst), for Dev and Test for Task 2.

Regarding on the last two systems (`T2-9` and `T2-10`), both of which build upon `T2-8`, we experimented with audio attention pooling (AAP) and data augmentation. However, applying AAP in `T2-9` slightly decreased performance, and combining AAP with audio data augmentation in `T2-10` led to a further drop. We hypothesize that attention pooling may reduce generalization due to the limited size of training data, and that the added noise from augmentation did not introduce sufficient variability to be beneficial. Further investigation is needed to better understand these results and refine the use of AAP in future iterations.

As can be observed in Table 6, with the system `T2-8` we achieved the second position in the competition. This system achieved a relative performance of 97.85% in relation to the best result achieved in Task 2 (0.8834 of F1-score), demonstrating the competitiveness of our approach.

Fig. 3 visually summarizes the results of Table 6, with systems sorted by Test performance. It can be noticed in the systems that were evaluated on Dev and Test that there is a noticeable performance reduction from Dev to Test. We can remark `T2-6` and `T2-5` systems, which were promising with the development results, but on Test they did not generalize properly. Therefore, Dev results should always be taken with caution when we are making decisions during the system development.

6. Conclusions and Future work

This study proposes various possible solutions to both tasks of the SatiSPeech 2025 contest [8], analyzing samples containing both text and audio in order to classify them into a "satire" or a "no-satire" category. Task 1 required to only analyze the transcriptions of the samples, while Task 2 allowed to analyze both the audio and the transcription of each sample.

For Task 1, we adjust for the downstream task various pre-trained language models, both Spanish-specialized and multilingual, using LoRA due to the relatively low amount of samples. We also explored LLMs and prompting techniques to study whether these models could reach the performance of the fine-tuned ones. For Task 2, we used both audio and text deep learning models in order to extract representative embeddings. These embeddings were then used to train various types of classifiers, such as MultiLayer Perceptrons or Support Vector Classifiers.

These methodologies have helped us achieve first place in Task 1 by fine-tuning a multilingual classification model and second place in Task 2 by using a support vector machine to classify samples, which was a very favorable outcome for our approaches to the proposed problem.

For future research, it would be worthwhile to try different forms of data augmentation in order to reach better results with larger models. The close results between prompting and fine-tuning also

make a compelling case for expanding the hyperparameter search for the fine-tuned systems in Task 1, especially considering the differences between fine-tuned models observed on the test set. Additionally, tuning the prompts using alternative samples could further improve performance. Other future line of research will be to test ensemble methods, that is, combining the strengths of fine-tuned and prompting-based models, or integrating multiple classifiers in Task 2. Finally, given the superior performance of multilingual models, exploring how these systems handle cross-lingual or code-switched satire could provide useful insights for broader applications.

7. Ethics Statement

We have not used additional data to those provided by the competition. The pretrained models used are obtained from HuggingFace models hub, under the Apache License 2.0, except for OpenAI's *Whisper*, which was obtained using its own official python package.

Acknowledgments

This work is partially supported by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe" under grant PID2021-126061OB-C41. Partially supported by the Vicerrectorado de Investigación de la Universitat Politècnica de València PAID-01-23. It is also partially supported by the Spanish Ministerio de Universidades under the grant FPU21/05288 for university teacher training.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] T. Jiang, H. Li, Y. Hou, Cultural differences in humor perception, usage, and implications, Frontiers in Psychology 10 (2019).

[2] V. Rubin, N. Conroy, Y. Chen, S. Cornwell, Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News, in: Proceedings of the Second Workshop on Computational Approaches to Deception Detection, Association for Computational Linguistics, San Diego, California, 2016, pp. 7–17. URL: https://aclanthology.org/W16-0802/. doi:10.18653/v1/W16-0802.

[3] C. Burfoot, T. Baldwin, Automatic Satire Detection: Are You Having a Laugh?, in: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Association for Computational Linguistics, Suntec, Singapore, 2009, pp. 161–164. URL: https://aclanthology.org/P09-2041/.

[4] A. Reyes, P. Rosso, T. Veale, A multidimensional approach for detecting irony in Twitter, Lang Resources & Evaluation 47 (2013) 239–268. URL: https://doi.org/10.1007/s10579-012-9196-x.

[5] L. Li, O. Levi, P. Hosseini, D. Broniatowski, A multi-modal method for satire detection using textual and visual cues, in: Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, International Committee on Computational Linguistics (ICCL), Barcelona, Spain (Online), 2020, pp. 33–38.

[6] J. A. García-Díaz, R. Valencia-García, Compilation and evaluation of the Spanish SATICorpus 2021 for satire identification using linguistic features and transformers, Complex & Intelligent Systems 8 (2022) 1723–1736.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6000–6010.

[8] R. Pan, J. A. García-Díaz, T. Bernal-Beltrán, F. García-Sánchez, R. Valencia-García, Overview of SatiSPeech at IberLEF 2025: Multimodal Audio-Text Satire Classification in Spanish, Procesamiento del Lenguaje Natural 75 (2025).

[9] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.

[10] M. d. P. Salas-Zárate, G. Alor-Hernández, J. L. Sánchez-Cervantes, M. A. Paredes-Valverde, J. L. García-Alcaraz, R. Valencia-García, Review of English literature on figurative language applied to social networks, Knowledge and Information Systems 62 (2020) 2105–2137.

[11] G. Wick-Pedro, C. F. da Silva, M. L. Inácio, O. A. Vale, H. de Medeiros Caseli, Using large language models for identifying satirical news in brazilian portuguese, in: Proceedings of the 16th International Conference on Computational Processing of Portuguese, 2024, pp. 156–167.

[12] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, M.-P. Gill, pyannote.audio: neural building blocks for speaker diarization, in: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 7124–7128.

[13] H. Bredin, A. Laurent, End-to-end speaker segmentation for overlap-aware resegmentation, in: Interspeech 2021, 2021, pp. 3111–3115.

[14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: International Conference on Machine Learning, PMLR, 2023, pp. 28492–28518.

[15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[16] T. Wolf, 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. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

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

[18] J. D. la Rosa y Eduardo G. Ponferrada y Manu Romero y Paulo Villegas y Pablo González de Prado Salas y María Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403.

[19] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

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

[21] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA, 2019, p. 2623–2631. URL: https://doi.org/10.1145/3292500.3330701. doi:10.1145/3292500.3330701.

[22] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, Z. Qiu, Qwen2.5 technical report, 2025. URL:

https://arxiv.org/abs/2412.15115. arXiv:2412.15115.

[23] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[24] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, 2022. URL: https://arxiv.org/abs/2212.04356. arXiv:2212.04356.

[26] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL: https://arxiv.org/abs/2006.11477. arXiv:2006.11477.

[27] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021. URL: https://arxiv.org/abs/2106.07447. arXiv:2106.07447.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

A. Examples from the SatirA dataset 2025

Table 7 shows some examples extracted from the dataset provided in the SatiSPeech25 shared task.

Table 7
Some examples of the *SatirA* dataset 2025.

Transcription	Label
Por las dos cosas se te saltan las lágrimas. Tú dirás. ¿Para qué te quieres ir, hombre? Si aquí estamos genial. Lo que dices es que esta mañana casi me pica un alacrán. ¿Un alacrán? ¡Me río yo! Si te pica un alacrán, puedes morir después de agonizar durante tres días entre dolores insoportables. Pero conozco bichos peores.	satire
Tras un estudio con 18 adultos, científicos de la Universidad McGill en Canadá descubrieron que un par de juegos en Tetris puede resultar útil a la hora de tratar la ambliopía. Según los investigadores, esto demostró ser más eficaz que el tradicional método de cubrir el ojo sano con un parche.	no-satire
En mis 40 años de política no he dejado una promesa sin cumplir. Dije que cenábamos anchoucas y cenamos anchoucas. ¡No hombre, no hombre, no! Pero Miguel Ángel, por el amor de Dios, ¿qué haces? Si hasta tenía el palo ya para hacer el espeto. ¡Déjate de doradas! No me explico que no pique ni una anchoa.	satire
No, me parece estupendo, me parece estupendo. Yo creo que cuanto más empleos creemos, mejor. Lo que me parece es... Señor Rodríguez, usted hace las preguntas de una manera curiosísima, que es, yo le escucho, entonces yo quiero responder y usted pasa de mi respuesta. Pero si yo le pregunto en siete segundos. Miguel Ángel, no le interrumpas. Usted tarda cinco minutos y medio en contestar algo que todavía no ha entendido lo del papel. Pero es que no le dejas, es cierto que no le dejas hablar. Lo del papel espero leerlo en Twitter, porque aquí no lo he entendido.	no-satire

B. Code developed

All code programmed in order to create and test models up until the deadline of the contest can be found in https://github.com/ajbarmil/ELiRF-satispeech-2025