SatiSPeech 2025 Task - Multimodal Speech-text Satire Recognition in Spanish

Nguyen Minh Bao[1,2], Trinh Tran Tran[1,2], Nguyen Thien Bao[1,2] and Dang Van Thin[1,2]

[1]*University of Information Technology-VNUHCM, Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam*
[2]*Vietnam National University, Ho Chi Minh City, Vietnam*

Abstract

This paper presents our submission to the SatiSPeech 2025 shared task at IberLEF [1], which focuses on satire detection using unimodal (text-only) and multimodal (text + audio) approaches. For the text classification task, we used TF-IDF features, BETO [2] embeddings, and shallow models such as Logistic Regression, SVM, and XGBoost. In the multimodal setting, audio features extracted via MFCCs were processed by a CNN and fused with text features (TF-IDF + SVD and BETO) before being passed to classifiers including MLP and other shallow models. A key aspect of our system was the independent fine-tuning of the CNN, allowing it to act as a specialized expert on audio data within the final ensemble. We applied a voting ensemble to combine model predictions. Our system ranked 4th in the official evaluation phase and improved to 2nd place in the post-evaluation phase, highlighting the effectiveness of combining shallow and deep learning techniques, multimodal fusion, and targeted CNN optimization for satire detection in Spanish.

Keywords

Satire detection, Multimodal learning, Spanish language, Text classification, Audio classification,

1. Introduction

SatiSPeech at IberLEF 2025 [1] is a shared task that aims to advance the automatic detection of satire through multimodal natural language processing (NLP) techniques. Satire is a complex and subtle communicative form that intertwines humor, irony, and social commentary, often relying on context-dependent cues such as tone, prosody, exaggeration, and cultural references [3, 4]. Unlike explicit sentiment expressions, satire typically conveys meaning indirectly, which poses considerable challenges for both human annotators and machine learning systems. This shared task provides a unique opportunity for the research community to explore the boundaries of satire recognition using both text and audio modalities.

The SatiSPeech 2025 competition consists of two subtasks. Task 1: Text Satire Detection focuses on classifying Spanish-language text segments as either satirical or non-satirical. This task requires models to identify linguistic patterns indicative of satire, such as wordplay, sarcasm, or implicit social critique [5, 6]. Task 2: Multimodal Satire Detection expands this challenge by incorporating audio data, enabling participants to exploit speech characteristics such as intonation, rhythm, and vocal emphasis in combination with textual content. This multimodal approach is particularly relevant, as satire often manifests through the interplay of spoken delivery and written language [7, 8].

To support these tasks, the organizers compiled a diverse dataset of annotated audio-text pairs from various Spanish-language sources, including satirical programs and news broadcasts [9]. The data was segmented and transcribed using state-of-the-art diarization and speech recognition systems [10, 11]. Despite growing interest in satire and irony detection, most prior studies have focused on unimodal text analysis [4, 5, 6]. Recent works have shown the promise of leveraging large language models and multimodal fusion techniques for satire detection, but research in Spanish—especially with spoken content—remains limited [8].

About the dataset for each task. In term of Task 1: Text Satire Detection, it focuses on detecting satire using textual information only. Each sample consists of a manually transcribed Spanish utterance extracted from broadcast programs. Satirical texts are characterized by irony, exaggeration, or indirect critique, while non-satirical texts convey straightforward, factual messages. The input is the transcription; no audio information is used.

IberLEF 2025, September 2025, Zaragoza, Spain
23520123@gm.uit.edu.vn (N. M. Bao)

CEUR Workshop Proceedings (CEUR-WS.org)

In term of Task 2: Multimodal Satire Detection, it extends the classification challenge by incorporating both text and audio modalities. Each sample includes a speech segment along with its transcription. Participants are expected to combine linguistic and acoustic cues—such as vocal tone, rhythm, and emphasis—to infer the presence of satire. Audio segments were processed using speaker diarization [7, 10] and transcribed using Whisper [11].

The dataset includes 6,000 training samples, 384 validation samples, and 2,000 unlabeled test samples, with a nearly balanced class distribution, as shown in Table 1.

Labels	**Training set**	**Validation set**	**Test set**
Satire	2,832	178	–
Non-satire	3,168	206	–
Total	**6,000**	**384**	**2,000**

Table 1: Distribution of samples across sets

2. Methodology

2.1. Preprocessing data

For data preprocessing, we adopted distinct strategies for the text and audio modalities to best preserve the features relevant to satire detection.

2.1.1. Text Preprocessing

For the textual data, we applied tokenization to segment each utterance into linguistically meaningful units, ensuring that important elements such as punctuation and word boundaries-often crucial for conveying irony or sarcasm-were retained. Following tokenization, we standardized the resulting feature vectors using a standard scaler. This normalization step was intended to enhance the stability and convergence of downstream machine learning models by ensuring that all features contributed equally during training.

2.1.2. Audio Preprocessing

In contrast, for the audio modality, we opted to use the original, unscaled audio features as input to our models. Satirical speech often relies on subtle prosodic and acoustic cues-such as intonation, rhythm, and emphasis-that can be diminished or lost through aggressive normalization or scaling. By preserving the raw audio features, we aimed to maintain the richness of these cues, allowing our models to fully exploit the nuanced characteristics of satirical speech. This approach ensured that both textual and acoustic information were optimally prepared for the subsequent classification tasks.

2.2. Task 1 - Text Satire Detection

We explored two vectorization strategies: **TF-IDF**, which captures surface-level lexical patterns and serves as a strong baseline for linear models, and **BETO**, a pre-trained Spanish BERT model, which encodes deep contextualized features. These embeddings were then fed into different classifiers to study both linear and non-linear decision boundaries.

To improve performance and robustness, we experimented with multiple model combinations, optimized their hyperparameters using **Optuna**[12], and finally applied a **simple ensemble strategy** to aggregate predictions from the best-performing models.

2.2.1. Vector Encoding Text

We used two different strategies to convert input text into numerical representations:

- **TF-IDF (Term Frequency–Inverse Document Frequency):** This classical vectorization method assigns a weight to each word based on its frequency across documents. It helps capture surface-level lexical cues that may be indicative of satire (e.g., rare or exaggerated word usage). The TF-IDF vectors were used with simple linear classifiers to test the generalizability of shallow features in satire detection.
- **BETO Embeddings:** We used the Spanish pre-trained BETO model to extract contextualized embeddings from the input text. Specifically, the token representation from the final hidden layer was used as a fixed-size sentence vector. These embeddings aim to capture deeper semantic patterns and contextual dependencies typical of satirical language.

2.2.2. Classification Model

We employed three classification configurations:

- **TF-IDF + Logistic Regression:** This setup serves as a strong and interpretable baseline to evaluate whether surface-level lexical cues are sufficient for satire classification. By applying logistic regression to TF-IDF features, the model learns a linear decision boundary based on word frequency patterns. While this approach lacks semantic understanding, it offers insight into how much information can be captured from basic lexical statistics alone. It is also computationally efficient and robust to small training data, making it suitable for initial prototyping and benchmarking.
- **BETO + Logistic Regression:** This configuration combines the power of contextualized word embeddings with a simple linear classifier. BETO model—a BERT variant pre-trained on a large Spanish corpus—we aim to capture deeper semantic and syntactic features inherent in satirical expressions. Logistic regression on top of BETO embeddings helps to test whether these learned representations are linearly separable and whether a simple classifier can exploit the richness of deep contextual cues without introducing model complexity.
- **BETO + XGBoost:** This configuration is designed to model the non-linear and complex relationships embedded within the high-dimensional BETO features. Unlike logistic regression, which assumes a linear boundary, XGBoost leverages an ensemble of decision trees to capture intricate patterns and interactions that are often present in satirical language—such as implicit irony, multi-level semantic cues, and context-dependent humor. The use of XGBoost enables the model to adapt to subtle and non-obvious characteristics that are critical for effective satire detection, particularly in cases where linear models may underperform due to oversimplified assumptions. As such, this setup serves as the most expressive model in our pipeline.

Hyperparameters for each model were optimized using **Optuna**, a hyperparameter tuning framework based on Bayesian optimization. Each configuration was trained and validated independently.

2.2.3. Ensemble model

Finally, we applied a soft voting ensemble to combine the predictions of the three models. Specifically, we assigned a weight of 0.2 to the TF-IDF + Logistic Regression model and a weight of 0.4 to each of the two BETO-based models: BETO + Logistic Regression and BETO + XGBoost. This weighted strategy allows the ensemble to place greater emphasis on the deep contextual features learned by BETO, while still retaining the complementary lexical cues captured by TF-IDF. The ensemble improved performance and robustness on the validation set by leveraging the strengths of both shallow and deep representations, as well as combining linear and non-linear decision boundaries.

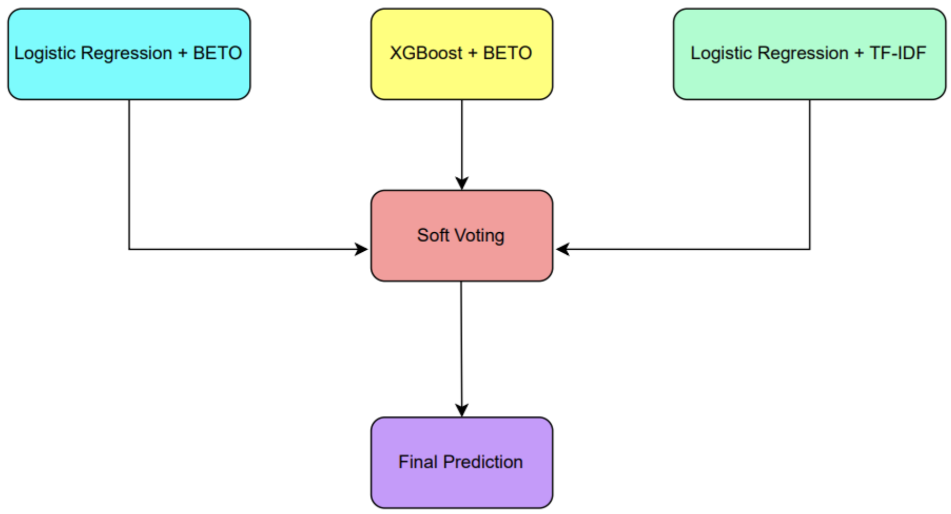

Figure 2.2.3: Soft Voting Technique for Task 1

2.3. Task 2 - Multimodal Satire Detection

For Task 2, we follow a multimodal pipeline that processes audio and text in parallel. Audio features are extracted and embedded through a CNN-based module, while textual features are obtained via embedding techniques. The resulting representations are concatenated into a joint feature vector and passed through a multimodal classification block. Final predictions are produced using a Hard-voting ensemble to enhance overall performance.

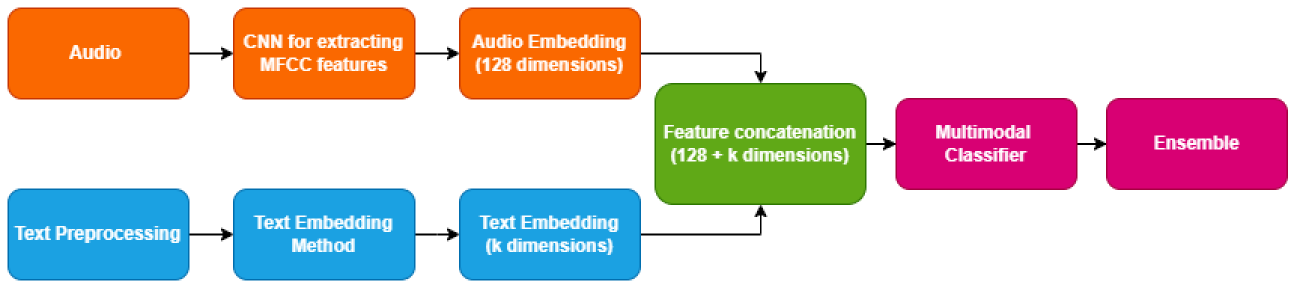

Figure 2.3: Diagram of Multimodal Satire Detection

2.3.1. Vector Encoding Text

For textual representation, we employed two encoding strategies to transform the input into fixed-size vectors suitable for multimodal integration:

- **TF-IDF + SVD:** This method uses Term Frequency-Inverse Document Frequency (TF-IDF) to capture the importance of words across the corpus. As TF-IDF typically yields high-dimensional sparse vectors, we apply Singular Value Decomposition (SVD) to reduce the dimensionality to 300. This step not only improves computational efficiency but also ensures compatibility with downstream neural components such as the MLP, which perform better with dense, fixed-length inputs.
- **BETO Embeddings:** We utilize BETO, a BERT-based pretrained language model for Spanish. Sentence-level embeddings are obtained by mean pooling over token representations from the final hidden layer, resulting in 768-dimensional dense vectors that encode rich contextual and semantic information.

These vector representations are later concatenated with the audio embeddings in the multimodal pipeline for final classification.

2.3.2. Convolutional Neural Network (CNN) to extract features from Mel Frequency Cepstral Coefficients (MFCCs)

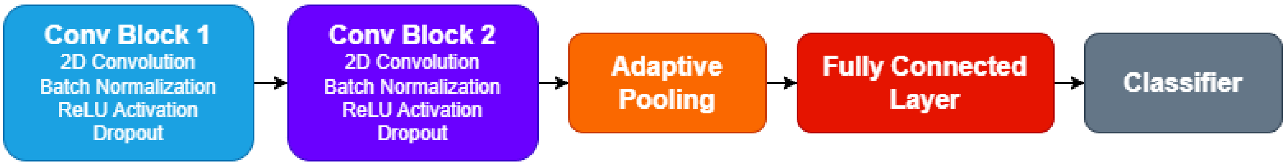

Figure 2.3.2: Diagram of Convolutional Neural Network (CNN) to extract features

To capture relevant acoustic cues for satire detection, we employed a Convolutional Neural Network (CNN) architecture to process Mel Frequency Cepstral Coefficients (MFCCs), a widely adopted representation of speech signals. MFCCs offer a perceptually grounded, compact encoding of the short-term spectral envelope of audio, effectively capturing timbral and prosodic nuances that are potentially informative for detecting satirical tone.

Our CNN architecture is designed to extract robust feature representations from the MFCC matrices, treating them as 2D inputs over time and frequency. The network consists of two convolutional blocks, each composed of a 2D convolutional layer followed by batch normalization and ReLU activation. The first block uses 16 filters while the second uses 32, allowing the model to progressively learn more abstract and complex patterns in the spectro-temporal domain. These blocks are followed by an adaptive average pooling layer, which reduces the spatial dimensions to a fixed-size 4×4 output, regardless of the input length — a crucial design choice to accommodate variable-length utterances during inference.

The pooled feature maps are then flattened and passed through a fully connected (linear) layer with 64 hidden units, acting as a feature bottleneck. Finally, a classification head maps the 64-dimensional representation to a two-class output (satire vs. non-satire). This design effectively balances model complexity and computational efficiency, enabling the system to learn discriminative audio features without overfitting.

By learning directly from low-level MFCC inputs, the CNN is capable of capturing non-trivial acoustic cues — such as exaggerated prosody, irregular rhythm, or tonal patterns — that may signal satirical intent. These audio-based features were later integrated with text-based features in our multimodal system to further enhance satire detection performance.

2.3.3. Classification Model

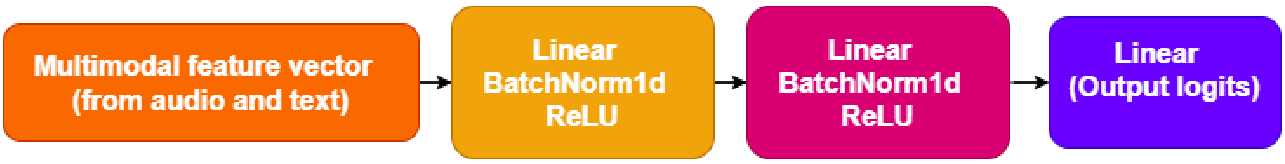

Figure 2.3.3: Classification model based on Multi-Layer Perceptron

The final multimodal feature vector, formed by concatenating audio and text embeddings, is passed to a classification model to predict whether an utterance is satirical or not. We explored three classification approaches:

- **Multi-Layer Perceptron (MLP):** This model consists of a two-layer feedforward neural network. Each hidden layer includes a linear transformation, followed by Batch Normalization and ReLU activation. The final layer is a linear projection to the output space. This architecture, shown in Figure 2.3.3, is designed to capture non-linear relationships between the multimodal features.
- **Logistic Regression:** As a simpler alternative, we also experimented with a linear logistic regression classifier directly applied to the concatenated multimodal vector. This model serves as a lightweight yet effective approach in cases where the feature space is already well-structured.
- **Support Vector Machine with RBF Kernel:** To model more complex decision boundaries, we employed a non-linear SVM with a radial basis function (RBF) kernel. This method is particularly suited for handling the fused feature space where interactions between modalities may not be linearly separable.

2.3.4. Ensemble Model

To enhance the robustness and overall performance of our system, we adopted a hard voting ensemble strategy that combines the predictions of three independent classifiers: a Multi-Layer Perceptron (MLP), Logistic Regression, and Support Vector Machine with RBF kernel. All models operate on the same multimodal input, which is constructed by concatenating text embeddings from BETO and audio embeddings extracted via a CNN.

A key aspect of our ensemble design is the training procedure for the CNN module. While the MLP is trained using a standard schedule of 20 epochs, the CNN, which is responsible for extracting audio features, is fine-tuned for a much longer period-specifically, 100 epochs, which is five times more than the MLP. In the ensemble setup described for Task 2, this extended training of the CNN was carried out independently to obtain more refined and stable audio feature representations. The main motivation for this longer training is to improve the quality of the audio embeddings and to enhance the contribution of the audio stream within the final hard voting ensemble.

Once all models are trained, final predictions are obtained using a hard voting scheme, where each classifier casts a vote, and the majority class is selected as the output. This ensemble approach not only stabilizes the predictions but also yielded performance that exceeded our initial expectations on the evaluation set.

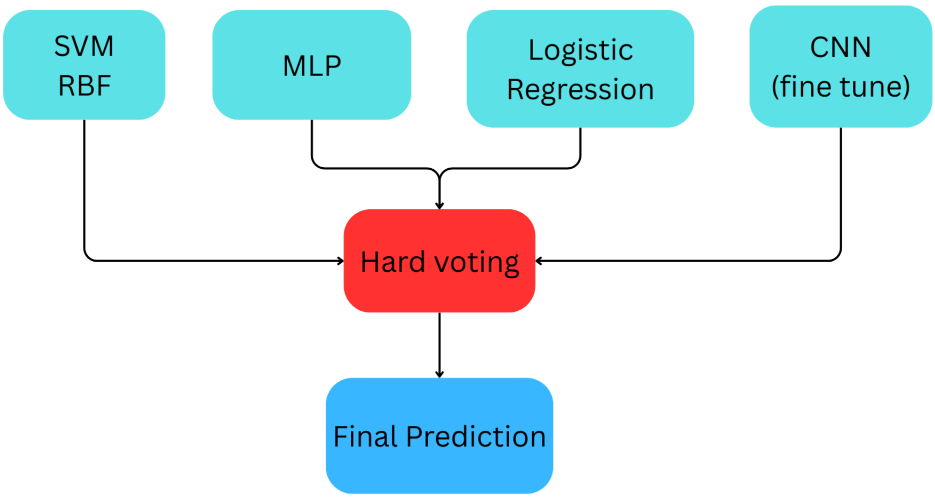

Figure 2.3.4: Hard Voting Technique for Task 2 powered by BETO

3. Experimental Setup

3.1. Datasets and Evaluation Metrics

3.1.1. Task 1 - Text Satire Detection

For Task 1, we used the official datasets provided by the organizers to train our models. To facilitate a comprehensive understanding of the data, we present both a table outlining the data distribution and a diagram illustrating the sequence lengths. Table 1 presents the data distribution for the datasets used in Task 1. The data is divided into a training set (6,000 samples), a validation set (384 samples), and a test set (2,000 samples). The task involves classifying samples as either "Satire" or "Non-satire." In the training set, there are 2,832 satire samples and 3,168 non-satire samples, while the validation set contains 178 satire and 206 non-satire samples. The test set consists of 2,000 samples, with label distributions not disclosed. Although the training and validation sets exhibit a moderate class imbalance, both satire and non-satire instances are well-represented, which is important for effective model training and evaluation. These distributions in table 1 play a crucial role in training and fine-tuning our models, as well as addressing potential data-related challenges.

Besides, Figure 3.1.1 depicts the distribution of sequence length, that is, the number of words within a sequence, for the two distinct categories in the datasets. The violin plots reveal the density distributions for both "satire" and "no-satire" samples. Overall, both categories exhibit similar sequence length patterns, with most samples concentrated in the 40-70 word range. However, the "satire" category displays a broader distribution with some sequences extending significantly longer, reaching approximately 220 words at the upper extreme. The "no-satire" category appears slightly more compact in its distribution, with fewer extremely short or long samples. Interestingly, both categories show similar median sequence lengths as indicated by the internal box plots, but the "satire" class demonstrates greater variability, particularly in the upper range. This pattern suggests that satirical content occasionally requires more elaborate expression, while non-satirical content tends to maintain more consistent length parameters.

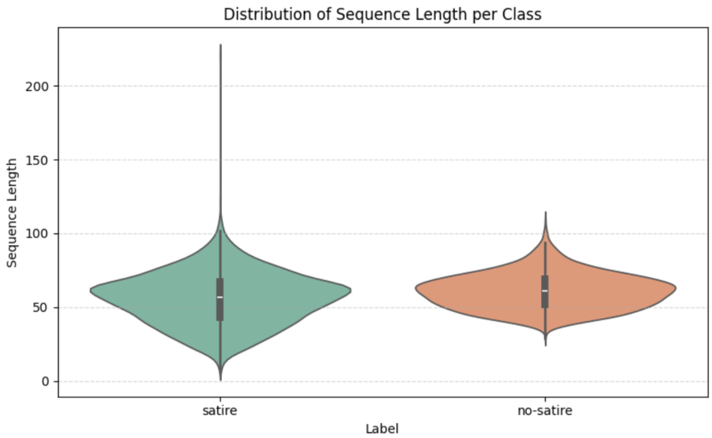

Figure 3.1.1: Sequence length distribution for Satire vs. Non-satire samples

3.1.2. Task 2 - Multimodal Satire Detection

For Task 2, we extended the satire detection framework to a multimodal setting by integrating raw audio features with the textual data described in Task 1. While the dataset splits and label distributions remained identical to Task 1 (see Table 1), each sample now included a paired audio clip. Text preprocessing followed the same pipeline as Task 1, involving tokenization and standardization, while audio features such as Mel-frequency cepstral coefficients (MFCCs) were extracted directly from the raw, unprocessed audio waveforms. This approach preserved subtle prosodic cues-such as exaggerated intonation or sarcastic pacing-that are critical to satirical speech but sensitive to distortion from aggressive normalization. System performance was evaluated using precision, recall, and macro F1-score, mirroring Task 1's evaluation protocol to ensure consistent and objective comparison between the unimodal and multimodal approaches. The macro F1-score, which equally weights both classes, served as the primary metric to account for the dataset's slight class imbalance.

3.2. System Settings

For the text modality, we used two types of vector representations: BETO embeddings with 768 dimensions and TF-IDF vectors reduced to 300 dimensions using Singular Value Decomposition (SVD). These representations were used consistently across all text-based models.

For shallow classifiers such as Logistic Regression and RBF-SVM, we employed Optuna to perform automatic hyperparameter optimization on the validation set.

For the audio modality, the CNN used to generate audio embeddings was trained for 20 epochs with a batch size of 16. The downstream MLP classifier was trained for 20 epochs with a batch size of 32. Both the CNN and MLP were optimized using the Adam optimizer, with a fixed learning rate of 0.001.

In the ensemble setup described for Task 2, we further fine-tuned the CNN independently for 100 epochs—five times longer than the MLP training schedule—to obtain more refined and stable audio feature representations.

This extended training was motivated by the need to enhance the contribution of the audio stream within the final hard voting ensemble.

4. Experiment Results and Discussion

4.1. Task 1 - Text Satire Detection

No	Model	F1 Task 1
1	Logistic Regression (TF-IDF)	0.8047
2	SVM (TF-IDF)	0.7929
3	Logistic Regression (BETO)	0.8291
4	XGBoost (BETO)	0.8053
5	**Final Ensemble (method 1+3+4 with Soft Voting)**	**0.8345**

Table 4.1: F1-scores of individual models and the final soft voting ensemble for Task 1.

As shown in Table 4.1, the use of TF-IDF embeddings combined with Logistic Regression achieved an F1-score of 0.8047, outperforming SVM (0.7929). This result suggests that even simple linear classifiers, when combined with sparse lexical representations, can capture a significant portion of satire-related patterns. However, these models may still be limited in capturing nuanced contextual information, which is often essential in satire.

Where BETO embeddings were used, we observed an improvement in performance. Logistic Regression with BETO achieved the highest F1-score of 0.8291, while XGBoost slightly lagged at 0.8053. This indicates that the BETO embeddings encode useful semantic and syntactic signals for satire detection, and that a simple linear decision boundary is already effective when applied to rich contextual representations. The relatively smaller gain from using XGBoost suggests that BETO embeddings are already linearly separable to a large extent.

Finally, the table presents the F1-scores of the three individual models alongside the result of their soft voting ensemble. The ensemble method combines BETO + Logistic Regression, BETO + XGBoost, and TF-IDF + Logistic Regression with respective weights of 0.4, 0.4, and 0.2. The final ensemble achieved the best overall performance with an F1-score of 0.8345, outperforming all individual models. This confirms that the models capture complementary patterns—BETO contributes deep semantic information, while TF-IDF captures surface-level lexical features that may still be informative in certain satirical contexts. The ensemble thus balances linear and non-linear decision boundaries, as well as shallow and deep features, leading to a more robust and generalizable classifier.

4.2. Task 2 - Multimodal Satire Detection

No	Model	F1 Task 2
1	MLP (TF-IDF + SVD)	0.803845
2	SVM RBF (TF-IDF + SVD)	0.811747
3	MLP (BETO)	0.835854
4	SVM RBF (BETO)	0.829076
5	Logistic Regression (BETO)	0.827754
6	CNN (Fine-tuned, 100 epochs)	0.802389
7	**Final Ensemble (method 3+4+5+6 with Hard Voting)**	**0.86164**

Table 4.2: F1-scores of individual models and the final hard voting ensemble for Task 2.

Table 4.2 presents the F1-scores of models using TF-IDF embeddings reduced via SVD. The SVM with RBF kernel outperformed the MLP, achieving an F1-score of 0.811747 compared to 0.803845. This suggests that the SVM model was better suited for handling the reduced-dimensional feature space derived from sparse TF-IDF vectors.

For models using BETO embeddings. The MLP model yielded the highest F1-score of 0.835854, outperforming the SVM RBF model (F1 = 0.829076). This result highlights the MLP's effectiveness in leveraging rich contextual features from the pretrained transformer-based model.

Among the individual models, the combination of BETO embeddings with an MLP classifier yielded the highest performance, achieving an F1-score of 0.8359. Building upon this, the final hard voting ensemble—which integrates predictions from BETO-based models and the independently fine-tuned CNN—achieved the best overall F1-score of 0.86164, demonstrating enhanced robustness and generalization through multimodal model aggregation.

5. Conclusion

This paper presented our approach to the SatiSPeech 2025 shared task on satire detection in Spanish. For Task 1, which focused on text-based satire classification, we adopted a combination of shallow learning techniques and pretrained language models. Specifically, we used BETO embeddings with XGBoost and Logistic Regression to capture deep contextual semantics, while TF-IDF combined with Logistic Regression was used to model broader, surface-level lexical patterns. This dual strategy allowed us to exploit both fine-grained and general textual cues.

In Task 2, which incorporated both audio and text modalities, we used a CNN architecture to extract features from MFCC representations of the audio signal. These were then fused with text features and passed to an MLP for classification. Notably, we took the same CNN used for feature extraction and fine-tuned it independently on audio data for 100 epochs—five times longer than the default MLP schedule. This extended training aimed to enhance the audio representation, and the resulting model was added to the ensemble, contributing significantly to performance.

By combining shallow models, deep embeddings, multimodal fusion, and a custom-trained CNN audio encoder, our system achieved strong performance—ranking 4th in the official evaluation and climbing to 2nd in the post-evaluation phase. The improvement was largely driven by our novel idea of introducing a separately fine-tuned CNN dedicated to the audio stream in the ensemble. This result demonstrates the power of tailored multimodal architectures and ensemble learning in complex tasks like satire detection.

6. Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

[1] 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).
[2] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, arXiv preprint arXiv:2308.02976 (2023).
[3] T. Jiang, H. Li, Y. Hou, Cultural differences in humor perception, usage, and implications, Frontiers in Psychology 10 (2019).
[4] M. Salas-Zárate, G. Alor-Hernández, J. L. Sánchez-Cervantes, M. 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.
[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, 2020, pp. 33–38.
[6] R. Ortega-Bueno, P. Rosso, J. E. M. Pagola, Multi-view informed attention-based model for irony and satire detection in spanish variants, Knowledge-Based Systems 235 (2022) 107597.
[7] H. Bredin, A. Laurent, End-to-end speaker segmentation for overlap-aware resegmentation, in: Interspeech 2021, 2021, pp. 3111–3115.

[8] 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.
[9] 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.
[10] 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: ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2020, pp. 7124–7128.
[11] 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, 2023, pp. 28492–28518.
[12] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631.