ITST at SatiSPeech–IberLEF 2025: Leveraging Transformers for Textual and Multimodal Satire Detection in Spanish

Mario Andrés Paredes-Valverde[1,*], María del Pilar Salas-Zárate[1]

[1] *Tecnológico Nacional de México/I.T.S. Teziutlán, Fracción l y ll SN, 73960 Teziutlán, Puebla, Mexico*

Abstract

This paper presents ITST's participation in the SatiSpeech 2025 shared task on satire detection in Spanish. We propose a modular and efficient pipeline that leverages pretrained transformer models to encode linguistic and prosodic features for binary satire classification. For Task 1 (text-only), we extract contextual sentence embeddings using the Spanish RoBERTa-BNE model and train a Support Vector Machine (SVM) for classification. For Task 2 (multimodal), we integrate acoustic information by concatenating Wav2Vec 2.0 audio embeddings with RoBERTa text features. Both models are evaluated on an internal validation split and performed strongly on the official leaderboard, achieving second place in Task 1 and fifth place in Task 2. Our results demonstrate that transformer-based embeddings, even when combined through simple early fusion, can deliver robust performance in both textual and multimodal satire detection.

Keywords

Satire Speech Recognition, Automatic Emotion Recognition, Natural Language Processing, Transformers, SVM

1. Introduction

Satire constitutes a complex and context-dependent form of communication, posing significant challenges to traditional content classification systems. Unlike conventional humor, satire operates through implicit critique, employing rhetorical devices such as irony, exaggeration, and parody. These strategies often rely on a shared understanding between speaker and audience, making satire inherently ambiguous and susceptible to misinterpretation. Accurately interpreting satire often requires understanding sociocultural context and speaker intent. These layers of meaning are conveyed not only through textual content but also via prosodic features in speech, including intonation, rhythm, and pitch variation. In multimodal contexts—where text and audio interact—these interpretive demands increase, necessitating models that can process both modalities in tandem to detect satire effectively [1, 2].

In complex multimodal scenarios where text, audio, and visual elements converge, interpretive demands increase due to the interplay of linguistic and paralinguistic cues. As a result, satire detection needs models capable of integrating and reasoning over multiple modalities simultaneously. Multimodal deep learning, particularly models that jointly process textual and auditory signals, has emerged as a promising approach to meet this challenge [3, 4].

Interest in automatic satire detection has grown, motivated by its potential to combat misinformation, support content moderation, and enhance media analysis. On digital platforms, satirical content is frequently misinterpreted as factual news, increasing the risk of misleading interpretations and the spread of false narratives. Consequently, reliable satire detection systems are essential, particularly in multilingual and culturally diverse environments.

Previous research has primarily focused on text-based satire detection, with transformer-based models achieving strong performance on datasets drawn from news articles and social media [5]. However, satire is also prevalent in spoken media—such as television programs, podcasts, and online videos—where vocal delivery plays a critical role. Despite this, benchmarks for multimodal satire

IberLEF 2025, September 2025, Zaragoza, Spain

[*] Corresponding author.

mario.pv@teziutlan.tecnm.mx (M. A. Paredes-Valverde); maria.sz@teziutlan.tecnm.mx (M. d. P. Salas-Zárate)
000-0001-9508-9818 (M. A. Paredes-Valverde); 0000-0003-1818-3434 (M. d. P. Salas-Zárate)

classification remain scarce. Related work on sarcasm and irony detection has demonstrated that multimodal models, integrating text and audio features, consistently outperform text-only baselines [6].

To address this gap, the SatiSpeech [7] shared task at IberLEF 2025 [8] introduces a benchmark for multimodal satire detection in Spanish. The task comprises two subtasks: (1) satire classification based on text alone, and (2) multimodal classification using aligned text and audio segments.

Our participation in the SatiSpeech at IberLEF 2025 [7] shared task involved the development of customized models for both tasks. To this end, we based our approach on the methodology used in the EmoSpeech competition at IberLEF 2024 [9], which also included a multimodal task involving both audio and text.

For Task 1, we employed a text-only architecture using the RoBERTa-base-bne model for Spanish (MarIA) [10]. Sentence embeddings were extracted from the [CLS] token of the final hidden state and used as input to a Support Vector Machine (SVM) classifier. We optimized model performance through grid search over kernel types, regularization parameters, and gamma values.

For Task 2 (multimodal classification), we extended this approach by incorporating audio-based features derived from the Wav2Vec 2.0 [11] model for Spanish. Specifically, we extracted embeddings from the first hidden state vector of each audio sample and concatenated them with the corresponding RoBERTa-based text embeddings. These multimodal vectors were then used to train a second SVM classifier, with hyperparameters optimized through the same grid search strategy as in Task 1. This combined representation allowed the model to capture both semantic content and prosodic cues critical to satire detection.

The remainder of this paper is organized as follows: Section 3 provides an overview of the shared task and dataset; Section 2 reviews previous work on satire detection in both textual and multimodal contexts; Section 4 describes our modeling approaches for both the unimodal and multimodal configurations; Section 5 presents experimental results and performance comparisons; and Section 6 concludes with key findings and future research directions.

2. Related Work

The task of satire detection has gained growing attention in natural language processing due to its relevance for media analysis, misinformation detection, and content moderation. Satirical content, while often humorous, relies on subtle rhetorical devices such as irony, parody, and exaggeration, which present unique challenges for automatic classification systems.

Early approaches to satire detection focused predominantly on handcrafted linguistic features, including lexical and syntactic patterns, sentiment polarity mismatches, and stylistic cues [12, 13]. However, these approaches struggled to generalize across domains and genres, particularly in multilingual and culturally diverse settings.

With the advent of transformer-based models, substantial improvements have been reported in satire and humor detection tasks. Studies such as [14] demonstrate the effectiveness of models like BERT in capturing contextual and pragmatic nuances of satirical texts, especially when trained on domain-specific corpora. These models leverage self-attention mechanisms to represent implicit relationships between tokens, which are critical for decoding irony and sarcasm.

Despite these advances, most research has remained focused on the textual modality. Recent work has begun to explore multimodal approaches that incorporate acoustic and visual information to improve satire detection, particularly in spoken or audiovisual content. For example, [6] proposed a multimodal satire detection framework combining textual and visual cues, while [4] applied audio-text fusion techniques for emotion and satire classification in Spanish. These studies have shown that prosodic features such as intonation, rhythm, and speech tempo provide complementary information for disambiguating satirical intent, especially when textual signals are ambiguous.

Moreover, the integration of multimodal transformer encoders (e.g., Wav2Vec 2.0 for speech and ViLT for vision) has opened new possibilities for cross-modal learning. Yet, challenges remain in effectively fusing these heterogeneous representations, as naive concatenation may fail to capture complex interdependencies between modalities.

In this context, the SatiSpeech 2025 shared task [7] represents a significant step forward by providing a benchmark dataset for multimodal satire detection in Spanish. It enables systematic evaluation of models that jointly process text and speech, fostering the development of more robust and culturally aware satire detection systems.

3. Task Description

The SatiSpeech Shared Task, organized as part of IberLEF 2025, targets the automatic detection of satirical content in Spanish across both textual and multimodal (text + audio) inputs. The task reflects the inherent complexity of satire, which relies on implicit rhetorical strategies and context, making it a valuable but challenging target for computational methods in media analysis, misinformation detection, and discourse understanding.

The task is divided into two tasks:

- **Task 1: Text-Based Satire Detection.** Participants are required to classify whether a given transcript represents satire using only textual information, leveraging features such as lexical choice, syntactic structure, and figurative language like irony or exaggeration.
- **Task 2: multimodal Satire Detection.** This subtask introduces additional complexity by requiring the integration of aligned speech and transcription data. Systems must incorporate both semantic (textual) and prosodic (acoustic) features—such as pitch, rhythm, and emphasis—for binary satire classification.

3.1. Dataset

Participants were provided with the *SatirA* dataset, a curated collection of Spanish-language speech segments collected primarily from YouTube. Satirical examples were sourced from television and web-based comedy programs like *El Intermedio*, *Zapeando*, *Homo-Zapping*, and *El Mundo Today*. Non-satirical samples were drawn from journalistic content published by outlets including *Antena 3 Noticias*, *El Mundo*, and *BBC News*. The dataset includes a variety of dialects and regional accents to encourage the development of robust models that generalize across linguistic variation.

Speech segments were automatically extracted using diarization tools, limited to a maximum length of 25 seconds. Transcriptions were generated using Whisper ASR [15], and a semi-supervised labeling process was used: automatic predictions were reviewed and corrected by expert annotators to ensure high-quality annotations.

The released training set contains approximately 25 hours of labeled content. Our modeling pipeline uses the entire provided training data for model training. For evaluation and hyperparameter tuning, we held out 10% of this dataset to create an internal validation set. Table 1 summarizes the distribution of samples and transcript lengths across both splits.

Split	Class	Samples	Avg. Length	Std. Dev.
Train	Non-satirical	3168	60.83	12.01
Train	Satirical	2832	55.95	17.60
Validation	Non-satirical	329	60.34	11.44
Validation	Satirical	271	56.80	16.85

Table 1
Distribution of training and validation samples by class. Avg. Length refers to the word count of transcriptions.

4. Methodology

This section describes the approach we used for both tasks of the SatiSpeech Shared Task: text-based satire detection (Task 1) and multimodal satire detection (Task 2). Instead of end-to-end neural fine-tuning, we adopted a feature-based strategy that leverages pretrained transformer models for Spanish, followed by traditional SVM classifiers. This design offered a balance between efficiency, interpretability, and performance.

Our system was developed in Python using `PyTorch`, `Transformers`, and `scikit-learn`. For Task 1, we extracted sentence-level embeddings from a RoBERTa model trained on Spanish corpora. For Task 2, we extended the pipeline by incorporating acoustic embeddings extracted from Wav2Vec 2.0. The final multimodal architecture is illustrated in Figure 1.

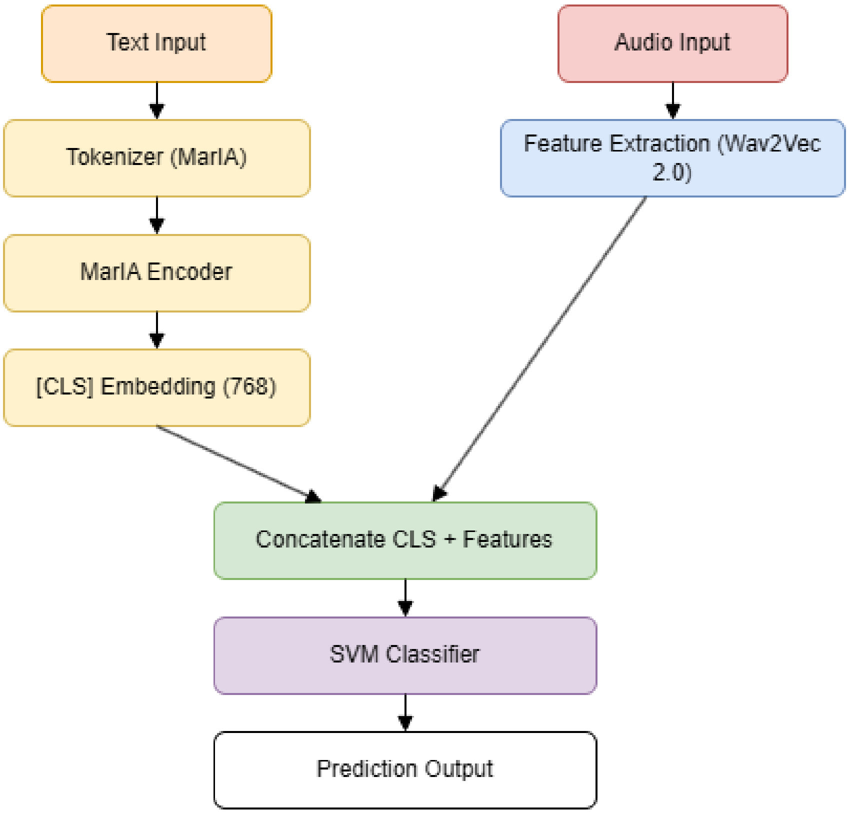

Figure 1: Multimodal classification pipeline combining RoBERTa-based text embeddings with Wav2Vec 2.0 audio embeddings.

4.1. Text-Based Satire Detection (Task 1)

For the first subtask, we used the pretrained `PlanTL-GOB-ES/roberta-base-bne` model to encode Spanish-language transcriptions. Each text was tokenized to a maximum length of 512 tokens with padding and truncation. From the final hidden layer of RoBERTa, we extracted the embedding corresponding to the `[CLS]` token, which resulted in a 768-dimensional fixed-length sentence representation.

These embeddings were then used to train an SVM classifier. We conducted a grid search over kernel types (`rbf`, `poly`), regularization parameter C, and kernel coefficient γ. The classifier was trained on the full dataset provided, using a 10% internal validation split to select the best hyperparameters. Classification performance was evaluated using the macro-averaged F1-score to account for class imbalance.

4.2. Multimodal Satire Detection (Task 2)

In the second subtask, we incorporated speech-based features into the model. We used the `facebook/wav2vec2-large-xlsr-53-spanish`[1] model to extract a 1024-dimensional embedding

[1] https://huggingface.co/facebook/wav2vec2-large-xlsr-53-spanish

from each audio clip. This embedding was obtained by computing the first token vector from the final hidden layer of the model after processing the waveform at 16kHz.

We concatenated the 1024-dimensional audio embedding from Wav2Vec 2.0 with the 768-dimensional RoBERTa text embedding, forming a unified 1792-dimensional multimodal feature vector. This fused representation was then used as input to a second SVM classifier, also tuned using grid search with the same hyperparameter configuration as in Task 1.

This approach allowed us to integrate both semantic (textual) and prosodic (acoustic) signals relevant to the detection of satire. By separating the embedding extraction and classification stages, the architecture remains modular and interpretable. Furthermore, the pipeline is efficient to train, GPU-compatible, and scalable to larger or multilingual datasets.

As in Task 1, model selection was based on macro-averaged F1-score using the internal validation split. The overall architecture remained consistent across tasks, with the only difference being the additional acoustic branch in the multimodal setup.

5. Results

During development, we experimented with multiple configurations of the SVM classifier, varying the kernel type (RBF vs. polynomial), regularization parameter (C), and the inclusion of normalization techniques. Among these, the RBF kernel with C=1.0 and gamma='scale' yielded the best macro F1 score. Other configurations showed reduced recall on the satirical class, indicating sensitivity to hyperparameter choices.

For Task 1, our RoBERTa-based model trained on Spanish transcriptions achieved a macro F1 score of **0.9394** on the internal validation set. This highlights the strength of monolingual transformers like RoBERTa-BNE in capturing linguistic markers of satire such as hyperbole, irony, and rhetorical shifts.

For Task 2, we extended the model by integrating speech representations extracted via Wav2Vec 2.0. The resulting system achieved a slightly higher macro F1 of **0.9412**. This marginal improvement supports the hypothesis that acoustic signals such as intonation, rhythm, and pitch help reinforce textual signals in satire detection. However, the performance gain remains modest, indicating that early fusion through concatenation does not fully leverage the expressive richness of audio features.

Table 2
Macro-averaged precision (M-P), recall (M-R), and F1-score (M-F1) for both tasks on the validation set.

Model	M-P	M-R	M-F1
Task 1			
RoBERTa-BNE + SVM	0.9397	0.9391	**0.9394**
Task 2			
RoBERTa-BNE + Wav2Vec 2.0 + SVM	0.9406	0.9419	**0.9412**

Tables 3 and 4 present the official rankings for SatiSpeech 2025. Team ITST placed **2nd** in Task 1 with a macro F1 of **84.55**, and **5th** in Task 2 with a macro F1 of **83.27**, confirming the competitiveness of our approach despite its relative simplicity and efficiency.

Since test set gold labels were not released, we created a stratified 20% validation split to assess model performance. Table 5 shows the macro-averaged metrics for both tasks.

A class-wise analysis reveals strong and balanced performance for both the `satire` and `no-satire`. Specifically, RoBERTa-BNE embeddings alone yielded high precision and recall across both classes, indicating that textual markers of satire are robustly captured. The addition of Wav2Vec 2.0 slightly improved recall on satirical samples, suggesting the model effectively captured prosodic features like exaggerated intonation or timing. So, while the performance improvement was not substantial, it confirms the utility of acoustic features and highlights the potential for more advanced fusion mechanisms in future work.

Table 3
Official leaderboard for Task 1

#	Team Name	M-F1
1	UPV-ELiRF	85.64
2	**ITST**	**84.55**
3	UMU-Ev	84.46
4	nguyenminhbao5032	83.27
5	Ferrara	83.21

Table 4
Official leaderboard for Task 2

#	Team Name	M-F1
1	UMU-Ev	88.34
2	UPV-ELiRF	86.44
3	Ferrara	83.70
4	nguyenminhbao5032	83.27
5	**ITST**	**83.27**

Table 5
Macro-averaged classification metrics on the internal validation set

Task	Precision	Recall	F1-score
Task 1 (Text)	0.9397	0.9391	0.9394
Task 2 (Text + Audio)	0.9406	0.9419	0.9412

6. Conclusion

In this work, we presented a lightweight yet effective pipeline for satire detection in Spanish, developed as part of the SatiSpeech 2025 shared task. Our approach relies on pretrained transformer models—RoBERTa-BNE for textual embeddings and Wav2Vec 2.0 for audio features—combined through early fusion and classified using SVMs. The design is modular, reproducible, and computationally efficient, enabling fast experimentation and high generalization capacity.

Our models achieved competitive performance on both tasks, securing second place in Task 1 and fifth place in Task 2 on the official leaderboard. The results confirm that monolingual language models are highly effective in identifying satirical patterns in text, and that the inclusion of audio features provides additional, though marginal, improvements.

While our simple fusion method proved sufficient to outperform several end-to-end architectures, future work will explore more advanced integration strategies, such as late fusion, attention-based weighting, or modality-specific fine-tuning. Furthermore, domain adaptation techniques may help reduce the gap between validation and test performance.

Overall, our findings support the utility of transformer-based feature extraction for multimodal satire detection and highlight the potential of hybrid pipelines combining pretrained models with traditional classifiers.

Future work will explore late fusion strategies where separate classifiers for text and audio are combined via ensemble techniques. Attention-based fusion could dynamically weigh modalities depending on input characteristics. Modality-specific fine-tuning may enhance performance by aligning internal representations with domain-specific cues, while domain adaptation strategies such as adversarial training or corpus alignment could reduce performance drops when transferring to out-of-domain satire. We also plan to investigate the integration of Large Language Models (LLMs), such as LLaMA variants and Qwen, to capture high-level pragmatic cues and sociocultural context that are often critical in satirical expression. Recent studies have demonstrated the efficacy of LLMs in complex classification tasks, such as hate speech detection, where nuanced intent and linguistic subtleties are critical [16, 17].

Acknowledgments

We are grateful to the Tecnológico Nacional de Mexico (TecNM, by its Spanish acronym) for supporting this work. This research was also sponsored by Secretariat of Science, Humanities, Technology and Innovation (Secihti).

Declaration on Generative AI

During the preparation of this work, the author(s) used DeepL in order to Grammar and spelling check.

References

[1] T. Jiang, H. Li, Y. Hou, Cultural differences in humor perception, usage, and implications, Frontiers in Psychology 10 (2019). URL: https://api.semanticscholar.org/CorpusID:59307773.

[2] M. Dynel, Beyond a joke: Types of conversational humour, Language and linguistics compass 3 (2009) 1284–1299.

[3] T. Baltrusaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2019) 423–443. URL: https://doi.org/10.1109/TPAMI.2018.2798607. doi:10.1109/TPAMI.2018.2798607.

[4] R. Pan, J. A. García-Díaz, M. Á. Rodríguez-García, R. Valencia-García, Spanish meacorpus 2023: A multimodal speech–text corpus for emotion analysis in spanish from natural environments, Computer Standards & Interfaces 90 (2024) 103856. URL: https://www.sciencedirect.com/science/article/pii/S0920548924000254. doi:https://doi.org/10.1016/j.csi.2024.103856.

[5] 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.

[6] L. Li, O. Levi, P. Hosseini, D. Broniatowski, A multi-modal method for satire detection using textual and visual cues, in: G. Da San Martino, C. Brew, G. L. Ciampaglia, A. Feldman, C. Leberknight, P. Nakov (Eds.), 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. URL: https://aclanthology.org/2020.nlp4if-1.4/.

[7] 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).

[8] 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.

[9] R. Pan, J. A. García Díaz, M. Á. Rodríguez García, F. García Sánchez, R. Valencia García, Overview of emospeech at iberlef 2024: Multimodal speech-text emotion recognition in spanish, Procesamiento del lenguaje natural 73 (2024) 359–368.

[10] 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.

[11] A. Baevski, H. Zhou, A. rahman Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, ArXiv abs/2006.11477 (2020). URL: https://api.semanticscholar.org/CorpusID:219966759.

[12] C. Burfoot, T. Baldwin, Automatic satire detection: Are you having a laugh?, in: K.-Y. Su, J. Su, J. Wiebe, H. Li (Eds.), 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/.

[13] A. Reyes, P. Rosso, T. Veale, A multidimensional approach for detecting irony in twitter, Language Resources and Evaluation 47 (2013) 239–268. URL: https://api.semanticscholar.org/CorpusID:580274.

[14] A. Mohan, A. M. Nair, B. Jayakumar, S. Muraleedharan, Sarcasm detection using bidirectional encoder representations from transformers and graph convolutional networks, Procedia Computer Science 218 (2023) 93–102. URL: https://www.sciencedirect.com/science/article/pii/S1877050922024991. doi:https://doi.org/10.1016/j.procs.2022.12.405, international Conference on Machine Learning and Data Engineering.

[15] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 28492–28518. URL: https://proceedings.mlr.press/v202/radford23a.html.

[16] R. Pan, J. A. García-Díaz, R. Valencia-García, Optimizing few-shot learning through a consistent retrieval extraction system for hate speech detection, Procesamiento del Lenguaje Natural 74 (2025) 241–252.

[17] J. A. García-Díaz, R. Pan, R. Valencia-García, Leveraging zero and few-shot learning for enhanced model generality in hate speech detection in spanish and english, Mathematics 11 (2023) 5004.