UMU-Ev at SatiSPeech-IberLEF 2025: Exploring Multimodal Satire Detection with Efficient Audio-Text Representations

Eduardo Valero-Vilella[1]

[1]*Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain*

Abstract

This paper presents the system developed by the UMU-Ev team for the SatiSPeech 2025 challenge organized within IberLEF, focused on satire detection in Spanish. The proposed approach integrates textual and acoustic modalities using computationally lightweight representations and classifiers, prioritizing efficiency in resource-constrained environments while maintaining competitive performance. For the text modality, models such as RoBERTa-bne and FastText were evaluated, while for audio, HuBERT-based representations were used in combination with MFCCs and prosodic features. The modalities were integrated through different fusion strategies: concatenation, averaging, weighted sum, and attention.

The system attained top-tier rankings in both tasks. In the monomodal text task, it reached a *Macro F1-Score* of 0.8445, ranking third. In the multimodal task, it achieved a *Macro F1-Score* of 0.8834, obtaining the first position in the official ranking. These results prove that it is possible to achieve strong performance using lightweight architectures and systematic experimentation.

Keywords

Satire Detection, Multimodal Classification, Transformers, Spanish NLP, MFCC, HuBERT, RoBERTa, FastText,

1. Introduction

In today's digital media landscape, disinformation and satire increasingly intermingle. This convergence is especially prominent on social networks and audiovisual platforms. Developing technologies that can properly interpret user-consumed and shared messages has thus become crucial. Distinguishing satirical from ironic content is particularly relevant for key tasks. These include intent detection, discourse analysis, and automatic content classification in both informative and humorous contexts [1]. Multimodal analysis combining textual and acoustic signals offers a promising solution. This approach effectively addresses language complexity and its expressive nuances.

Automated satire detection presents major challenges for NLP systems and audio analysis. This discourse type frequently employs irony, exaggeration, and ambiguity, along with linguistic devices that resist formalization [2]. Real-world contexts like interviews, talk shows, or online humor introduce additional complexity, where subtle prosodic cues often accompany satirical elements. These include intonation, pauses, and emphasis patterns that escape text-only models [3]. Consequently, exploring approaches that integrate textual and acoustic information becomes essential. Such integration would enhance satire detection models' performance.

Traditionally, satire detection relied on textual analysis. Statistical models or transformer architectures like BETO and Spanish-trained RoBERTa variants were commonly used. These approaches succeeded in related tasks like emotion and irony detection. Several recent studies in the EmoSPeech framework demonstrate this effectiveness [4]. However, performance declines in ambiguous scenarios. This is especially true when acoustic signals contribute meaningful nuances. Multimodal approaches are now gaining momentum. They combine text and audio representations [5]. Despite progress in recent competitions, robust modality integration remains challenging. The issue persists particularly for

IberLEF 2025, September 2025, Zaragoza, Spain
eduardo.valero@um.es (E. Valero-Vilella)
https://github.com/valevil27 (E. Valero-Vilella)

languages like Spanish. Although increasingly represented, Spanish resources remain incomparable to those in English.

Our participation proved valuable for the SatiSPeech 2025 competition [6]. Organized within the IberLEF [7] forum, it enabled development and evaluation of monomodal and multimodal configurations. We used a realistically annotated corpus for this purpose. The final system employed a late fusion architecture. This combined two data streams: Audio embeddings from HuBERT [8] plus prosodic features (MFCCs [9], pauses, intensity, and pitch). Textual representations came from RoBERTa-bne [10]. Integration used a late fusion strategy. Final classification relied on a Support Vector Machine (SVM). Our system achieved top results in the competition's multimodal task. We ranked first on the official leader board. This outcome supports an important hypothesis: Competitive performance in complex tasks remains achievable. It holds true even with limited computational resources and base models. Furthermore, our simplified approach enables new research pathways. Future work can explore more complex models.

2. State of the Art

Automated satire detection remains a significant challenge in natural language processing. This discursive phenomenon systematically employs irony, exaggeration, and double meanings to critique social, political, or cultural behaviors [11]. Unlike more straightforward linguistic tasks, satire requires interpreting not only explicit content but also communicative intent, often contrary to literal meaning [12]. This semantic ambiguity complicates automated system design, as even humans show inconsistencies when identifying satirical or sarcastic expressions [13]. Current state-of-the-art models also struggle with figurative language. Their strong dependency on data patterns limits their ability to represent pragmatic and sociocultural context [12, 11].

These difficulties are amplified for Spanish, a language underrepresented in international resources and benchmarks compared to English. Scientific literature on Spanish satire and irony detection remains scarce. Available datasets typically suffer from size limitations, quality issues, or inadequate dialectal coverage [11]. Spanish also exhibits significant regional diversity affecting lexicon, syntactic constructions, and ironic usage patterns. This diversity prevents simple generalization of models trained on single domains or varieties [14]. Additional challenges arise when working with texts from social media or audiovisual sources. These often contain grammatical errors, abbreviations, emojis, and implicit prosodic markers that evade conventional textual analysis [15]. Consequently, researchers must develop more robust linguistic representations and classification strategies adaptable to satire's diverse manifestations across Spanish-speaking contexts [11].

The scenario grows more complex when addressing satire from a multimodal perspective integrating textual and acoustic information. While prosody, tone, rhythm, and speech pauses provide crucial clues about satirical intent, their correct interpretation requires effective integration with textual content [16]. Detecting inconsistencies between verbal content and delivery, such as a seemingly serious comment uttered with mocking intonation, demands models capable of establishing cross-modal relationships. These models must identify ironic patterns not explicitly expressed [17]. However, multimodal satire research in Spanish remains incipient. Existing approaches focus primarily on related tasks like emotion or irony detection. Few annotated corpora combine text and audio with specific satire labels. Effective system development thus requires overcoming both technical barriers and structural limitations from data scarcity [14, 16].

Early satire and irony detection approaches used traditional machine learning methods with manually extracted linguistic features. These works explored lexical and stylistic traits (n-grams, punctuation, PoS tags), semantic indicators (polysemy, polarity contrast), and emotional markers (polarity, categorical/dimensional emotions) [12, 11]. They also examined contextual and extralinguistic properties using tools like the *LIWC* dictionary [18]. Some studies proposed satire-specific features including jargon and offensive language usage [11, 12]. Classifiers like SVMs, decision trees, and ensemble methods showed competitive performance, especially for Spanish social media corpora [16]. While effective

in specific domains, these approaches require intensive feature engineering and generalize poorly to heterogeneous or informal contexts [15, 11].

The deep learning revolution brought models that learn distributed representations directly from text. Notable architectures include BiLSTM with attention mechanisms [15], hierarchical networks combining phrase- and document-level analysis [11, 15], and CNN-based models for detecting ironic tones [17]. Researchers also explored *Transformer* models [17] like BERT and RoBERTa, plus multi-view systems such as MvAttLSTM [11]. The latter integrate multiple information sources (linguistic, dense, contextualized) via multi-head attention. These models demonstrate a marked capacity to encode complex semantic and pragmatic subtleties inherent in satirical discourse. Some frameworks even unify sarcasm detection in multimodal environments incorporating text, audio, and visual signals. However, their effectiveness heavily depends on corpus quality/diversity and modality fusion strategies [11, 17, 15].

Fusion strategy represents a key dimension in multimodal systems. Integration approaches for text, audio, and image vectors significantly impact model performance [16]. Common techniques include Early Fusion (direct concatenation of raw representations/embeddings) and Late Fusion (separate modality processing with later output combination) [19, 17]. In satire detection, researchers observe interesting interactions between semantic and prosodic cues: When ironic meaning is textually clear, tone/intonation influence diminishes. Conversely, in ambiguous cases, prosodic features (F0 modulation, duration, amplitude) become decisive for correct intent interpretation [20, 19].

Recent competitions have catalyzed advances in this field. Notable examples include the EmoSPeech challenge [5] at IberLEF 2024, focused on Spanish emotion recognition from real-world data. This competition introduced the Spanish MEACorpus 2023 [2], over 13 hours of manually annotated audio using Ekman's taxonomy. This multimodal corpus from spontaneous YouTube situations has gained wide adoption. The BSC-UPC team [4] achieved top results by integrating pre-trained text (RoBERTa-bne) and audio (XLSR-wav2vec 2.0) representations. Their architecture used Attention Pooling and dense classification networks in a voting ensemble. The attention mechanism reduced embedding dimensionality efficiently, particularly valuable in resource-limited contexts.

Collectively, these studies formed the foundation for our experimental design. They guided our selection of promising textual/acoustic representations and fusion techniques compatible with modest computational infrastructure. We prioritized simplicity, efficiency, and performance balance.

3. Objectives

The primary objective of this work is to develop a system capable of detecting satire in Spanish using multimodal signals. This system will combine textual and acoustic information within the framework of the SatiSPeech 2025 challenge [6]. To achieve this overarching goal, we define the following specific objectives:

- Evaluate various representation strategies for Spanish text and audio. This includes both pre-trained models and classical techniques, with emphasis on solutions offering optimal performance-computational cost balance.
- Compare different monomodal configurations (processing text and audio separately). This comparison will establish a robust baseline for developing our multimodal solution.
- Design an effective multimodal fusion strategy. The solution must be compatible with modest training infrastructure constraints.
- Implement and evaluate diverse classification approaches. We will consider both simple neural networks and traditional classifiers including SVM and Random Forest.

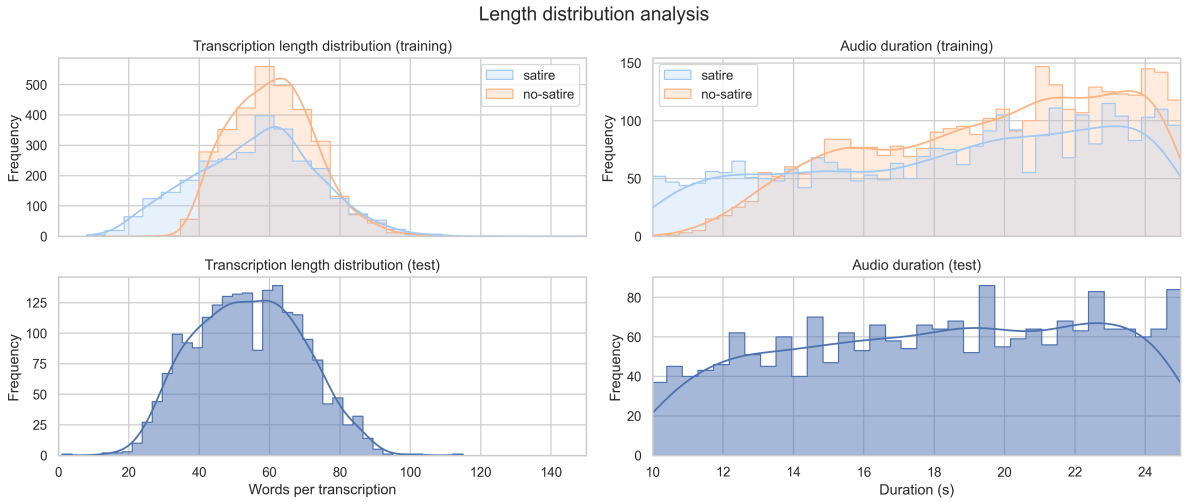

Figure 1: Distribution of transcription lengths and audio durations in the SatirA corpus, differentiated by partitions. The training partition is further split by class.

4. Methodology

This section describes the proposed system for the task of satire detection in Spanish using multimodal signals. The system comprises several independent modules for processing textual and acoustic modalities, whose outputs are integrated in a fusion stage prior to classification. The strategies employed in each component are detailed below.

4.1. Dataset

The official data provided by the organizers of the multimodal task in the SatiSPeech 2025 challenge [6] were used for system development and evaluation. The corpus, named SatirA, consists of Spanish audio clips extracted from YouTube videos of satirical programs such as *El Intermedio*, *Zapeando*, *Homo-Zapping*, and *El Mundo Today*, as well as non-satirical news sources like *Antena 3 Noticias*, *El Mundo*, and *BBC News*. The dataset includes content from various Spanish-speaking regions, ensuring significant linguistic and cultural diversity while minimizing regional bias.

Videos were segmented automatically with diarization tools [21, 22], and clips exceeding 25 seconds in duration were discarded. Automatic transcriptions were generated using Whisper [23]. Content annotation as satirical or non-satirical followed a semi-automatic approach: first, automatic classification techniques were applied, followed by manual validation by three experts.

The dataset includes approximately 25 hours of recordings, divided into a training set with 6,000 samples and a test set with 2,000 samples. The test set contains hidden labels as it was used for the official evaluation of participant submissions. The final system evaluation was conducted through the Codalab platform, used by the organizers to host and manage the competition.

Analysis of class distribution in the training set reveals a slight imbalance: approximately 52.8% of samples are labeled as non-satirical and 47.2% as satirical.

Figure 1 shows the distribution analysis of transcription lengths and audio durations, differentiated by class in the training set (`satire` / `no-satire`) and aggregated in the test set. Regarding text, non-satirical transcriptions tend to be slightly longer with a more concentrated distribution centered around 55 words. Satirical transcriptions, conversely, show greater dispersion.

For audio, non-satirical segments consistently exhibit longer durations. These differences may reflect distinct structural patterns in satirical versus informative or descriptive discourse. The test set distributions align with those of the training set, suggesting good data coherence, though models should avoid relying exclusively on these formal differences for class inference.

4.2. Text Representation

Various text representation strategies were explored, ranging from classical approaches to pre-trained language models. First, traditional vectorization techniques like `CountVectorizer` and `TfidfVectorizer` were applied, with vocabulary limited to the 1,000 most frequent words and removal of Spanish stop-words. Basic preprocessing included lowercase conversion and removal of URLs, mentions, and punctuation marks.

Concurrently, dense representations from pre-trained models were evaluated. For `Word2Vec` and `FastText`, average vectors were generated from words in each text. For Transformer-based models like `RoBERTa-bne` and `XML-RoBERTa`, Hugging Face implementations with `AutoTokenizer` and `AutoModel` were used. Table 1 summarizes the dimensions of the most effective representations identified during experimentation for both textual and acoustic modalities.

4.2.1. Experimental Procedure

Internal testing employed cross-validation on different subsets of the training set. The primary evaluation metric was *Macro F1-Score*, following challenge guidelines. *Macro F1-Score* is the unweighted average of F1-scores across all classes, calculated individually per class and then averaged.

Experimentation was divided into two phases. An initial exploratory phase used reduced training subsets (1,000 samples for training, 200 for validation) to rapidly test multiple representation-classifier combinations. The most promising configurations were selected from these tests.

In the second phase, complete embeddings of selected models were precomputed and stored in `.npy` format to accelerate training and evaluation. Subsequently, hyperparameter search techniques were used to optimize classifiers:

- `GridSearchCV` from `sklearn` for logistic regression and support vector machines.
- `Keras Tuner` for dense neural networks, testing different architectures, layer sizes, and dropout rates.

To ensure experiment reproducibility, the parameter `random_state=420` (or equivalents) was fixed for all applicable procedures, including data splitting, cross-validation strategies, and classifier initialization. Evaluation results along with optimal hyperparameter combinations were saved in `.json` files for subsequent analysis.

4.2.2. Notable Representations and Classifiers

Three representations demonstrated particularly competitive performance during experimentation:

- `RoBERTa-base-bne`: A `RoBERTa` version trained on 570 GB of peninsular Spanish text compiled by the National Library of Spain. This model is specifically adapted for Spanish tasks and has shown outstanding performance in text classification, entity recognition, and question answering [24].
- `XML-RoBERTa`: A multilingual `RoBERTa`-based model trained on data from over 100 languages, demonstrating strong contextual text handling capabilities [25].
- `FastText`: Dense vector representations based on sub-words, proving especially robust for informal or diverse linguistic contexts [26].
- `Word2Vec`: An unsupervised learning model generating dense word representations in continuous vector space, capturing semantic and syntactic relationships through contextual occurrence in large text corpora [27].

Among evaluated classifiers, the following stood out:

- **Logistic regression**: Linear model for binary classification that fits a logistic function to estimate probabilities and assign classes [28].

- **Support vector machines (SVM)**: Classifier identifying the hyperplane that maximizes the margin between classes, using only the closest points (support vectors) to define boundaries [28].
- **Fully connected dense neural networks (DNN)**: Model composed of interconnected neuron layers where each node applies a nonlinear transformation to a linear combination of its inputs [28].

Table 1
Dimensions of representations used for text and audio

Text		Audio	
Representation	Dimension	Representation	Dimension
RoBERTa-bne (CLS)	768	MFCC (statistics)	26
XML-RoBERTa (CLS)	768	MFCC + prosodic	31
FastText (mean)	300	MFCC + prosodic + deltas	83
Word2Vec (mean)	300	Wav2Vec2-base (mean/CLS)	768
		HuBERT-base (mean/CLS)	768

4.3. Audio Representation

Two audio representation strategies were explored: traditional MFCC-based feature extraction, and pre-trained self-supervised representation models:

- MFCCs *(Mel-Frequency Cepstral Coefficients)*: Coefficients representing the short-term power spectrum of an audio signal, modeling human sound perception via mel scales [29]. Additional features were incorporated:
  - *Prosodic features*: Reflect suprasegmental speech aspects like intonation, rhythm, energy, and duration, associated with emotional expression or emphasis [30].
  - *Deltas and delta-deltas*: First and second-order temporal derivatives of acoustic features (MFCCs in our case), capturing dynamic changes over time such as spectral parameter velocity and acceleration [29].
- Wav2Vec2: Self-supervised model based on contrastive learning that acquires speech representations directly from unlabeled audio. It uses a convolutional network to encode signals and a transformer network for contextualization, trained to distinguish true from negative representations [31].
- HuBERT: Self-supervised model predicting hidden phonetic units obtained through clustering on acoustic representations, combining segmentation learning and content prediction. This enables learning hierarchical structures without manual transcriptions [8].

For both self-supervised models, base versions were used, extracting two representation types: mean pooling and the special CLS token (the output vector's first position, considered a condensed representation of processed audio clip content) [31, 8].

4.3.1. Experimental Procedure

The experimentation process mirrored the textual case, divided into two main phases:

An initial exploratory phase used a reduced training subset (1,000 training samples, 200 validation samples), enabling rapid evaluation of multiple acoustic representation-classifier combinations without hyperparameter tuning. The most promising combinations were selected from these experiments.

In the second phase, complete representations of the training set were precomputed and stored in `.npy` format for reuse. Using these representations, more complex models were explored alongside identical hyperparameter search techniques: `GridSearchCV` and `Keras Tuner`.

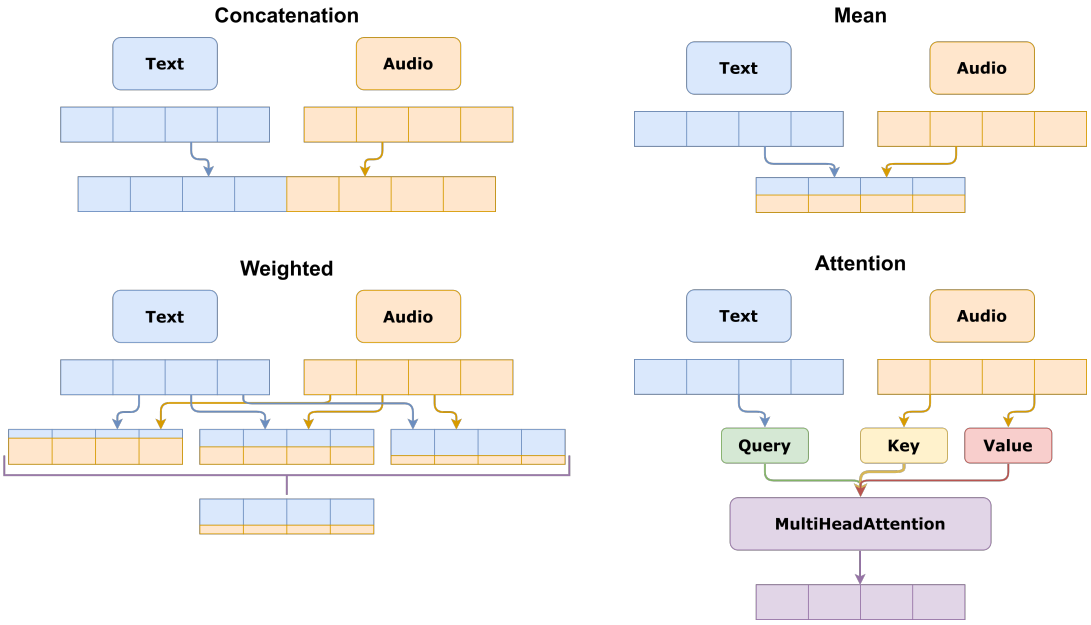

Figure 2: Schematic representation of multimodal fusion mechanisms used in the proposed system.

4.3.2. Notable Representations and Classifiers

After completing experimentation, the following acoustic representations proved most effective:

- MFCCs with prosodic features
- MFCCs with prosodic features and deltas
- HuBERT with mean pooling extraction

Notably, while performance differences between MFCCs with and without deltas (always including prosodic features) were modest, including deltas provided consistent slight improvements across most tested models. Thus, this variant was retained for final experimentation.

Regarding classifiers used with these representations, support vector machines and dense neural networks achieved the best results, mirroring their effectiveness in text classification, unlike logistic regression, which underperformed in this modality.

4.4. Multimodal Fusion

Multimodal experimentation was designed using classifiers and representations that performed best in unimodal text and audio tasks. Specifically, SVM and DNN classifiers were selected, excluding logistic regression due to its inferior audio performance. The following representations were used:

- **Text**: RoBERTa-bne (CLS) and FastText (mean).
- **Audio**: HuBERT (mean), MFCCs with prosodic features, and MFCCs with prosodic features and delta-deltas.

Combinations were generated using different vector fusion methods for each text-audio representation pair, as illustrated in Figure 2. The goal was to evaluate cross-modal complementarity and analyze its impact on classifier performance. Concatenation was used for same-modality representations due to its simplicity and speed.

The experimental procedure replicated earlier phases: first, reduced training subsets (1,000/200) were used for exploratory testing of combinations, fusion methods, and classifiers. Subsequently, the most promising configurations were retrained on the full training set, applying hyperparameter searches to optimize DNNs and SVMs.

4.4.1. Fusion Methods

The following fusion methods were evaluated to combine text and audio representations:

- **Concatenation**: Baseline method joining both vectors into a single extended vector by appending the audio vector to the text vector (or vice versa). Preserves all information from both modalities but increases classifier input dimensionality.
  Two independent vectors (one per modality) are pre-normalized with `StandardScaler` to ensure comparable scales. The `NumPy` `concatenate` function is then applied, forming an input vector with total dimensionality equal to the sum of both representations.
- **Weighted sum**: Vectors are combined through weighted summation, multiplying each vector by a weight (between 0 and 1) before summing. This method adjusts each modality's relative influence but requires identical dimensions. For mismatched dimensions, the longer vector was truncated.
  A logistic regression model validated weight searches due to its speed and low computational cost. Weights were tested in 0.1 increments using a validation subset for each weight combination. Note that weights were optimized for this lightweight model and may not maximize performance for subsequent SVM/DNN classifiers.
- **Mean**: Equivalent to weighted fusion with equal weights (`0.5/0.5`). This specific variant was included for its simplicity and balanced integration without additional tuning.
- **Attention**: This method adapts the attention architecture by Vaswani et al. [32] for text-audio fusion. Specifically, the text vector serves as `query`, while the audio vector simultaneously acts as `key` and `value`. This configuration allows the attention mechanism to identify the audio aspects most relevant to textual content.
  Implementation used TensorFlow's `MultiHeadAttention` class [33], which computes attention as a weighted combination of value vectors using weights derived from query-key similarity. Since original vectors don't represent temporal sequences, an extra dimension was added to simulate this, a technical requirement for this layer. However, this simplification may limit the mechanism's capacity to model complex relationships observed in real sequences. Attention output is normalized and fed to the classifier.
  For parameterization, `num_heads=4` was selected to balance expressiveness and computational cost. This allows attention space division into multiple subspaces without excessive overfitting or parameter growth, particularly relevant given the embeddings' relatively low dimensionality. The `key_dim` parameter (key dimension per head) was defined as `dim_model//num_heads`, ensuring concatenated attention output matches input embedding dimensionality, a standard practice guaranteeing model coherence.

4.5. Experimental Environment

All experiments were conducted on a personal ASUS ROG Strix G531GT laptop with an Intel Core i7-9750H processor (2.60GHz), 16 GB RAM, and an NVIDIA GeForce GTX 1650 Mobile GPU (4 GB VRAM). The operating system was Ubuntu 24.04 LTS, and development used a Python 3.12.6 virtual environment. Notebooks and scripts were primarily coded and executed in Visual Studio Code.

Key libraries included: `TensorFlow 2`, `NumPy`, `torch`, `torchaudio`, `HuggingFace Transformers`, `scikit-learn`, `librosa`, keras_tuner and `FastText`.

System computational limitations influenced certain experimental decisions. Thus, embedding models were used exclusively in feature extraction mode, with generated embeddings saved to disk to accelerate subsequent processes. Computing audio representations for the 6,000 training samples required approximately one hour per configuration.

5. Results

This section presents the results obtained during the second phase of the experimental process. We begin with monomodal models applied separately to text and audio, followed by multimodal experiments combining both modalities. Finally, we analyze the results obtained in the official competition.

Experiments were conducted on the training set using five-fold cross-validation, with a split of 5,500 samples for training and 500 for validation in each fold. The primary evaluation metric was *Macro F1-Score*, following the competition's established criteria. Precision and recall were also calculated but were consulted only occasionally as auxiliary metrics.

5.1. Monomodal Results: Text

Figure 3 shows results for the monomodal satire classification task using only transcribed text. Various combinations of representations (`RoBERTa-bne`, `FastText`, and `Word2Vec`) and classifiers (DNN, SVM, and logistic regression) were evaluated. Hyperparameter search configurations can be seen in Table 2.

The best performance was achieved with `RoBERTa-bne` using an SVM classifier, with its `classification_report` shown in Table 3. This embedding model performed best both individually and when combined with `FastText`, significantly outperforming `FastText` alone, reinforcing the value of contextualized models as base representations. For these two top embedding models, SVM was the highest-performing classifier, with dense neural networks and logistic regression slightly below and showing similar performance. Notably, SVM's hyperparameter search time was the longest at nearly one minute per iteration, while logistic regression was ten times faster.

Regarding representations, `FastText` performed reasonably well as a standalone model, especially with SVM and DNN, though not reaching `RoBERTa`'s level. `Word2Vec` showed the lowest performance both independently and when combined with `RoBERTa`, and was subsequently excluded.

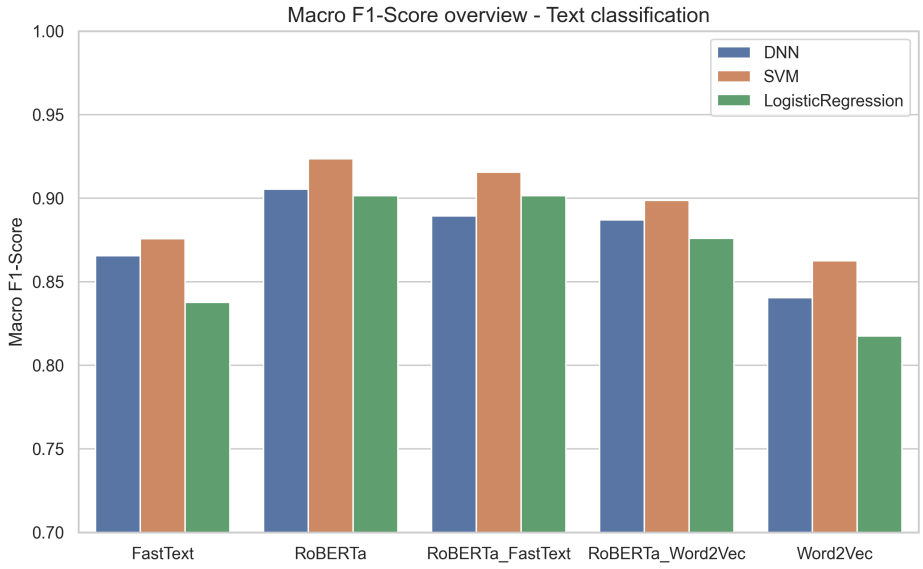

Figure 3: Cross-validation results for textual representations and classifiers evaluated (*Macro F1-Score*).

5.2. Monomodal Results: Audio

Acoustic representations `HuBERT` and `Wav2Vec2` (base versions) were evaluated with two extraction strategies (mean pooling and `CLS` token), both isolated and combined with `MFCCs`. Results are illustrated

in Figure 4. Hyperparameter search configurations and optimal settings are shown in Table 4.

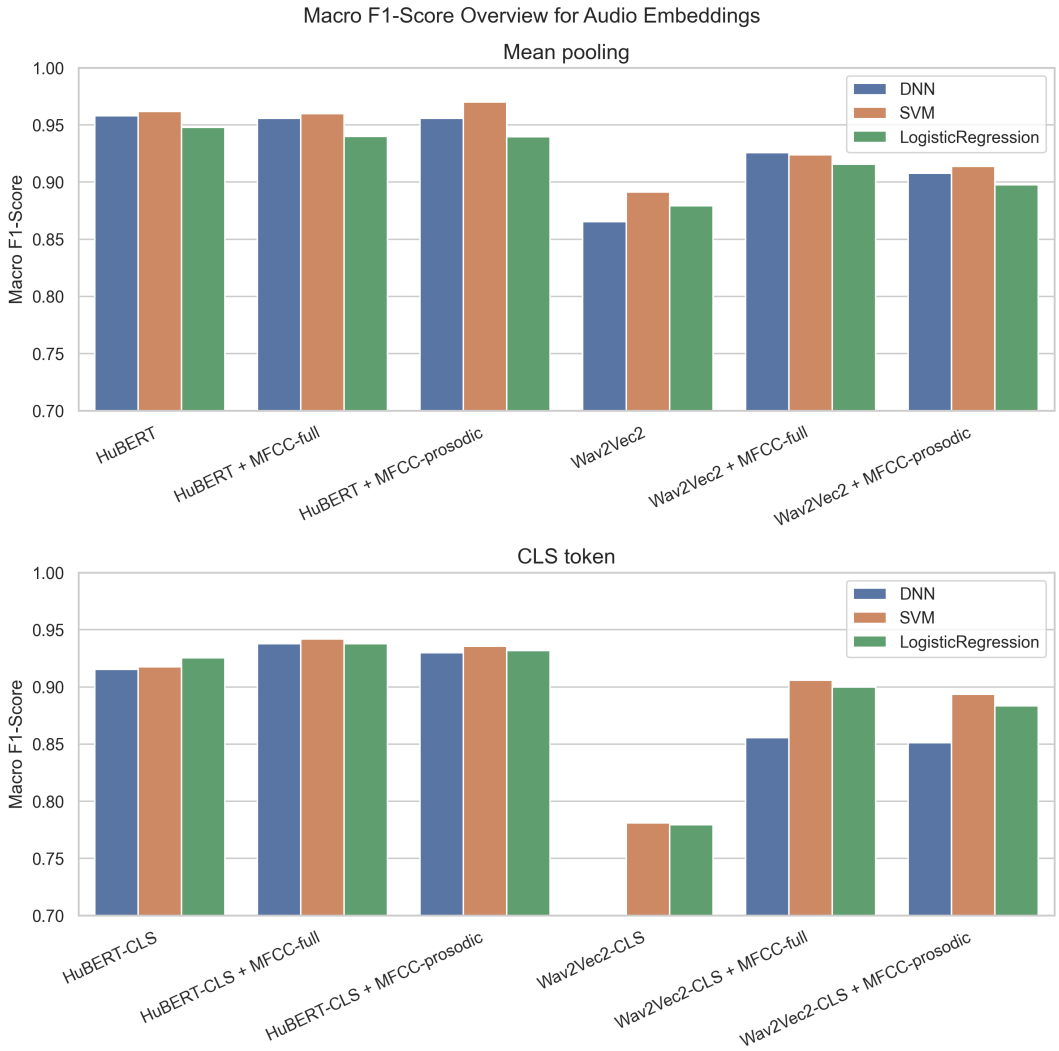

Figure 4: Performance comparison for monomodal audio classification across acoustic representations and classifiers.

HuBERT consistently showed superior performance when using mean pooling rather than the CLS token, with an average difference of 3%. This gap was particularly pronounced with DNN and SVM classifiers. The best configuration combined `HuBERT-mean` and prosodic MFCCs with SVM, as shown in Table 5.

Similar to HuBERT, the mean-based representation systematically outperformed CLS for `Wav2Vec2`, with a much larger difference. The best result used `Wav2Vec2-mean` and MFCCs with prosodic features and delta-deltas with DNN (*Macro F1-Score* = 0.9258). Logistic regression again performed best with the CLS token representation, except when `Wav2Vec2` was concatenated with full MFCCs.

In all configurations, HuBERT clearly outperformed `Wav2Vec2`, both in standalone versions and when combined with MFCCs. The average performance gap was 5% for mean representations and 8% for the CLS token. Consequently, `HuBERT-mean` was selected as the audio representation model for multimodal combinations, both standalone and combined with MFCCs.

SVM and DNN were the most competitive classifiers for this task. Logistic regression, which showed competitive performance in text and other audio embeddings, lagged notably with `HuBERT-mean`. This limitation motivated its use as an auxiliary method for weight estimation in weighted fusion, where its low computational cost enables rapid combination exploration.

Table 2
Hyperparameter search ranges and optimal configurations for text classifiers.

Model	Search Range	Optimal Configuration
SVM	• C ∈ {0.1, 1, 10} • `kernel` ∈ {linear, rbf} • `gamma` ∈ {scale, auto}	• C = 10 • `kernel` = rbf • `gamma` = scale • Representation: `RoBERTa-bne`
Logistic Regression	• C ∈ {0.01, 0.1, 1, 10} • `penalty` ∈ $\{\ell_1, \ell_2\}$ • `solver` = liblinear • `max_iter` ∈ {100, 200}	• C = 0.01 • `penalty` = ℓ_2 • `solver` = liblinear • `max_iter` = 100 • Representation: `RoBERTa-bne+FastText`
DNN (Keras)	• `num_layers` ∈ {1, 2, 3} • `units` ∈ {32, 64, ..., 256} • `activation` ∈ {relu, tanh} • `dropout` ∈ {0.1, ..., 0.5} • `learning_rate` ∈ [1e-4, 1e-2] (`log-scale`)	• Layers: – `Dense(64,relu)` + `Dropout(0.4)` – `Dense(160,relu)` + `Dropout(0.2)` – `Dense(96,relu)` + `Dropout(0.2)` • `learning_rate` = 8.31×10^{-4} • Representation: `RoBERTa-bne`

Table 3
`classification_report` for best text model (SVM on `RoBERTa-bne`).

Class	Precision	Recall	F1-Score	Support
Non-Satirical (0)	0.9154	0.9432	0.9291	264
Satirical (1)	0.9342	0.9025	0.9181	236
Macro Average	0.9248	0.9229	0.9236	500

SVM Model Hyperparameters:

- C: 10
- `kernel`: `rbf`
- `gamma`: scale
- `Search time per iteration`: 61.77 seconds

Table 4
Hyperparameter search ranges and optimal configurations for audio classifiers.

Model	Search Range	Optimal Configuration
SVM	• C ∈ {0.1, 1, 10} • kernel ∈ {linear, rbf} • gamma ∈ {scale, auto}	• C = 10 • kernel = rbf • gamma = scale • Representation: HuBERT+MFCCs prosodic
Logistic Regression	• C ∈ {0.01, 0.1, 1, 10} • penalty ∈ $\{\ell_1, \ell_2\}$ • solver = liblinear • max_iter ∈ {100, 200}	• C = 0.1 • penalty = ℓ_2 • solver = liblinear • max_iter = 100 • Representation: HuBERT
DNN	• num_layers ∈ {1, 2, 3} • units ∈ {32, 64, ..., 256} • activation ∈ {relu, tanh} • dropout ∈ {0.1, ..., 0.5} • learning_rate ∈ [1e-4, 1e-2] (log-scale)	• Layers: – Dense(256,relu) + Dropout(0.5) – Dense(32,relu) + Dropout(0.5) • learning_rate = 3.40×10^{-4} • Representation: HuBERT

Table 5
classification_report for best audio model (SVM on HuBERT+MFCCs-prosodic).

Class	Precision	Recall	F1-Score	Support
Non-Satirical (0)	0.9698	0.9735	0.9716	264
Satirical (1)	0.9702	0.9661	0.9682	236
Macro Average	0.9700	0.9698	0.9699	500

SVM Model Hyperparameters:

- C: 10
- kernel: rbf
- gamma: scale
- Search time per iteration: 44.94 seconds

5.3. Multimodal Results

This final phase analyzed multiple text-acoustic representation combinations using different fusion methods and classifiers, evaluated via cross-validation on the training set. Results shown in Figures 5, 6, and 7 provide an overview, with key configurations summarized in Table 6.

Table 6
Summary of top and bottom-performing combinations.

#	Text	Audio	Fusion	Model	Macro F1-Score
1	FastText	HuBERT+MFCC-prosodic	attention	SVM	0.9670
2	FastText	HuBERT+MFCC-prosodic	mean	SVM	0.9693
3	RoBERTa+FastText	HuBERT+MFCC-prosodic	attention	SVM	0.9693
4	RoBERTa	HuBERT+MFCC-prosodic	weighted	SVM	0.9692
5	RoBERTa	HuBERT+MFCC-prosodic	mean	SVM	0.9687
6	FastText	HuBERT+MFCC-prosodic	weighted	SVM	0.9685
7	RoBERTa+FastText	HuBERT+MFCC-prosodic	concatenation	SVM	0.9678
8	RoBERTa+FastText	HuBERT+MFCC-prosodic	weighted	SVM	0.9672
9	RoBERTa+FastText	HuBERT+MFCC-prosodic	mean	SVM	0.9671
10	RoBERTa	HuBERT+MFCC-prosodic	concatenation	SVM	0.9668
11	FastText	HuBERT+MFCC-prosodic	concatenation	SVM	0.9665
12	RoBERTa	HuBERT+MFCC-prosodic	attention	SVM	0.9660
13	FastText	HuBERT+MFCC-full	mean	DNN	0.9657
14	FastText	HuBERT+MFCC-full	weighted	DNN	0.9657
15	FastText	HuBERT+MFCC-prosodic	attention	DNN	0.9650
58	RoBERTa	HuBERT+MFCC-full	attention	SVM	0.9199
59	RoBERTa	HuBERT+MFCC-full	weighted	SVM	0.9199
60	FastText	HuBERT+MFCC-full	mean	SVM	0.9199
61	RoBERTa	HuBERT+MFCC-full	mean	SVM	0.9199
62	RoBERTa+FastText	HuBERT+MFCC-full	concatenation	SVM	0.9199
63	RoBERTa+FastText	HuBERT	mean	DNN	0.9198
64	RoBERTa+FastText	HuBERT	concatenation	DNN	0.9198
65	RoBERTa	HuBERT+MFCC-prosodic	weighted	DNN	0.9197
66	RoBERTa+FastText	HuBERT	attention	DNN	0.9197
67	RoBERTa+FastText	HuBERT+MFCC-prosodic	attention	DNN	0.9150
68	RoBERTa	HuBERT	concatenation	DNN	0.9148
69	FastText	HuBERT	attention	DNN	0.9147
70	RoBERTa	HuBERT	attention	DNN	0.9147
71	FastText	HuBERT	mean	DNN	0.8948
72	FastText	HuBERT	weighted	DNN	0.8897

Overall multimodal combination performance was high and relatively homogeneous, with most *Macro F1-Scores* between 0.92 and 0.97. Regarding representations, best combinations used HuBERT acoustic embeddings with prosodic features (Figures 5a and 6b). All text variants showed similar performance when combined with audio representations.

Fusion methods yielded consistent results across concatenation, attention, mean, and weighted sum, with no dominant strategy. Attention and weighting achieved the highest scores in several combinations (Figure 5b). However, Figure 7a shows no clearly dominant method, with all displaying similar distributions and slight advantages for attention at the upper end.

Regarding classifier performance (Figure 6), SVM was the most robust option across most configurations, systematically outperforming dense networks. The exception occurred with full MFCCs combinations, where DNN achieved better results. DNN also showed lower score dispersion (Figure 7b), suggesting greater stability despite lower average scores.

An exception occurred with HuBERT combined with full MFCCs, where DNN performed better. Additionally, DNN exhibited lower variance, suggesting greater stability across runs compared to SVM's

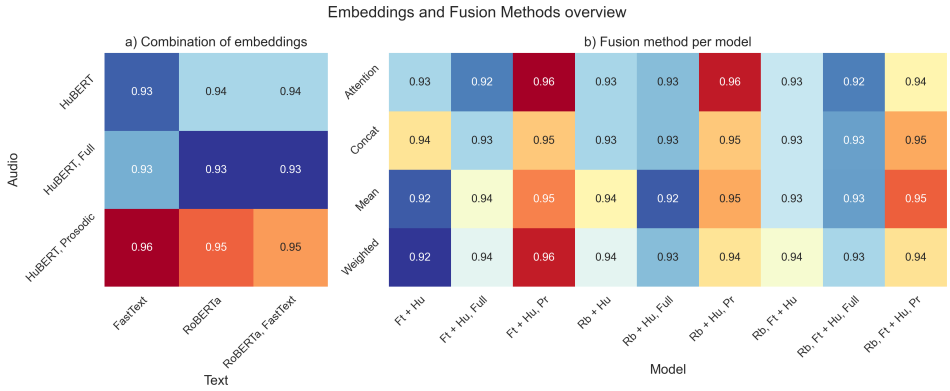

Figure 5: Overview of embedding combinations and fusion methods. a) Average *Macro F1-Score* by text-audio embedding combination. b) Results per embedding combination by fusion method.

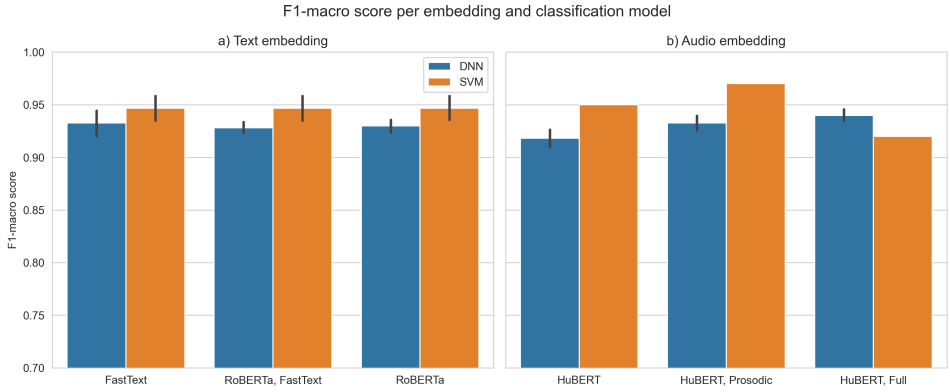

Figure 6: *Macro F1-Score* by representation combination and classifier for monomodal text (a) and audio (b) tasks.

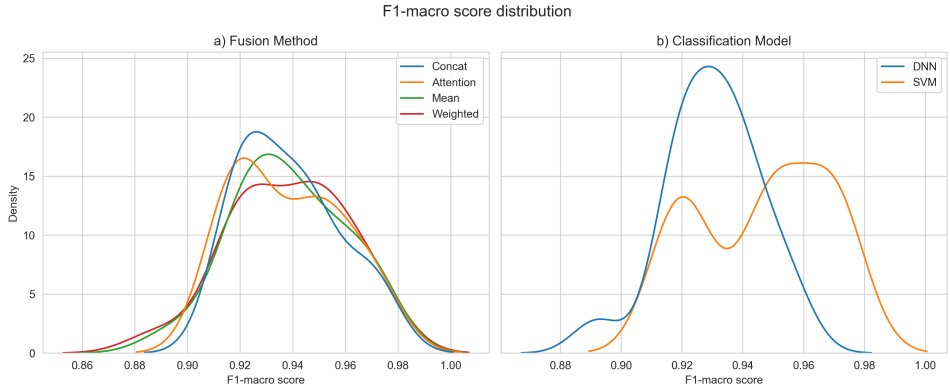

Figure 7: Result distributions. a) By text-audio fusion method. b) By classifier.

wider dispersion.

Notably, while top multimodal configurations did not significantly outperform standalone acoustic models as they did with text models, they clearly improved result consistency (Figure 8). Monomodal audio models showed greater dispersion across runs and combinations, while multimodal systems sustained high performance across configurations. Hyperparameter search configurations and optimal settings are shown in Table 7.

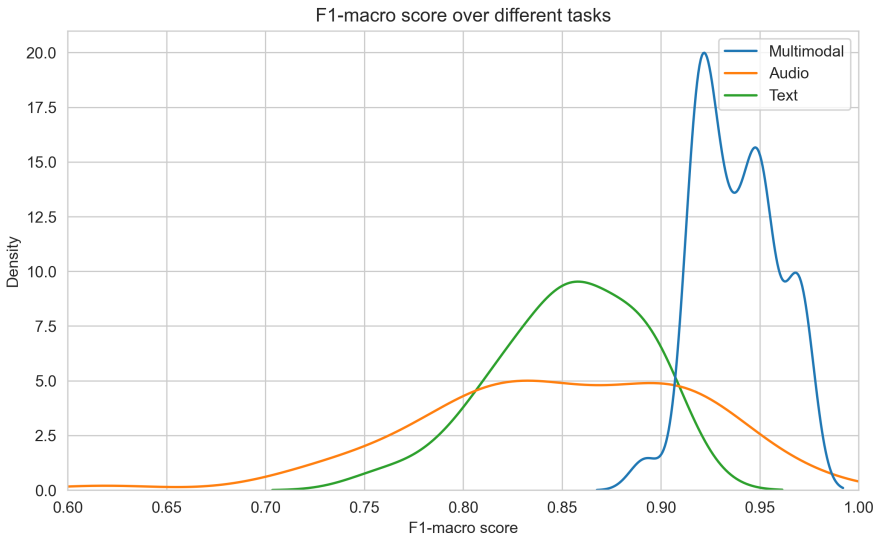

Figure 8: *Macro F1-Score* distribution across explored tasks (text, audio, multimodal). Multimodal modality shows lower dispersion and higher concentration at high values.

5.4. Competition Results

We present results from evaluation on the competition's hidden-label test set and analyze submitted configurations for both the monomodal text task and the multimodal task.

5.4.1. Monomodal Task: Text

For the monomodal text task, five configurations combining `RoBERTa-bne` and `FastText` with different classifiers were tested. Results are shown in Table 9.

The best result used `RoBERTa-bne` with SVM, achieving a *Macro F1-Score* of 0.8446 (Experiment 2). This aligns with cross-validation results, confirming the classifier's consistency across experimental conditions. The second-best configuration combined `RoBERTa-bne` and `FastText` with SVM (Experiment 4, 0.8417), indicating that adding `FastText` did not improve prediction performance.

The `RoBERTa-bne + FastText` with logistic regression configuration (Experiment 5) achieved a *Macro F1-Score* of 0.8191, notable for its simplicity and low computational cost. Dense neural networks scored lower with both `RoBERTa` (Experiment 1, 0.8122) and `FastText` (Experiment 3, 0.8058), suggesting reduced generalization capacity compared to traditional classifiers.

Table 7
Hyperparameter search ranges and optimal configurations for multimodal classifiers.

Model	Search Range	Optimal Configuration
SVM		
	• C ∈ {0.1, 1, 10} • `kernel` ∈ {linear, rbf} • `gamma` ∈ {scale, auto}	• C = 10 • `kernel` = `rbf` • `gamma` = scale • Representation: `FastText + HuBERT + prosodic MFCCs (attention)`
DNN		
	• `num_layers` ∈ {1, 2, 3} • `units` ∈ {32, 64, ..., 256} • `activation` ∈ {relu, tanh} • `dropout` ∈ {0.1, ..., 0.5} • `learning_rate` ∈ [1e-4, 1e-2] (`log-scale`)	• Layers: – `Dense(256,relu)` + `Dropout(0.5)` – `Dense(32,relu)` + `Dropout(0.2)` – `Dense(160,relu)` + `Dropout(0.1)` • `learning_rate` = 8.53×10^{-4} • Representation: `FastText + HuBERT + full MFCCs (mean)`

Table 8
`classification_report` for best multimodal model (SVM on `FastText + HuBERT + MFCCs-prosodic` with attention).

Class	Precision	Recall	F1-Score	Support
Non-Satirical (0)	0.9902	0.9528	0.9712	265
Satirical (1)	0.9490	0.9894	0.9688	235
Macro Average	0.9696	0.9711	0.9700	500

SVM Model Hyperparameters:

- C: 10
- `kernel`: `rbf`
- `gamma`: scale

Table 9
Results in SatiSPeech 2025 competition for monomodal text task.

Representation	Classifier	Macro F1-Score (test)
`RoBERTa-bne`	DNN	0.8122
`RoBERTa-bne`	SVM	**0.8446**
`FastText`	DNN	0.8058
`RoBERTa-bne+FastText`	SVM	0.8417
`RoBERTa-bne+FastText`	Logistic Regression	0.8191

5.4.2. Multimodal Task

Table 10 shows test set results for five configurations selected after multimodal analysis. All use HuBERT and prosodic MFCCs audio representations combined with different text representations and classifiers using various fusion methods.

The best result combined this acoustic representation with RoBERTa classified via SVM after concatenation fusion (Experiment 1), achieving a *Macro F1-Score* of 0.8834. This configuration reflects SVM's strength with high-dimensional balanced representations from direct embedding fusion.

The second-highest score used the same acoustic representation with `FastText`, fused via attention and classified with DNN (Experiment 2, 0.8803). This validates the attention mechanism's effectiveness for integrating heterogeneous modalities without explicit sequential structure.

Third place used `RoBERTa + FastText` fused via mean and classified with SVM (Experiment 3, 0.8798). This configuration stood out for its simplicity and stability, balancing representations without complex tuning.

The last two experiments showed slight performance drops: `RoBERTa + FastText` with DNN and concatenation fusion (Experiment 5, 0.8746), and `FastText` with SVM and weighted sum fusion (Experiment 4, 0.8747). Though competitive, they didn't surpass other configurations.

Table 10
Results in SatiSPeech 2025 competition for multimodal text + audio task.

Audio	Text	Classifier	Fusion	Macro F1-Score (test)
`HuBERT + Prosodic`	`RoBERTa`	SVM	Concatenation	**0.8834**
`HuBERT + Prosodic`	`FastText`	DNN	Attention	0.8803
`HuBERT + Prosodic`	`RoBERTa + FastText`	SVM	Mean	0.8798
`HuBERT + Prosodic`	`FastText`	SVM	Weighted sum	0.8747
`HuBERT + Prosodic`	`RoBERTa + FastText`	DNN	Concatenation	0.8746

5.5. Error Analysis

For error evaluation, we took the best-performing model on the training set, defining a 500-sample holdout set. We identified 9 misclassifications: 7 false positives and 2 false negatives, shown in Table 11. Errors fall into these groups:

- **False positives: Latin American news**
  Errors in non-satirical samples (`3db0b886`, `eebdb176`) correspond to segments from Latin American news channels, presented in neutral accents with background music and sound effects characteristic of this format. Despite serious content and monotone delivery, these were misclassified as satire.
- **False negatives: satire mimicking news**
  A second group includes satirical samples misclassified as non-satirical, such as segments from *El Mundo Today* (`8f7e5f21`, `e29b87d0`) and *El Intermedio* (`607a93b9`, `5b97747c`). Speakers use completely serious tones mimicking traditional news/report formats. Satire is conveyed primarily through semantic content requiring contextual and pragmatic interpretation beyond acoustic cues and literal text. When present, prosodic cues appear subtly or late, potentially escaping model detection.
- **False negatives: satire cues only in audio**
  Some samples contain typical satirical elements (laughter, onomatopoeia, exaggerated tone) not reflected in transcriptions, as in `ab2e18c4` or `ca6bda3a`. These highlight limitations of text models and certain acoustic representations in capturing paralinguistic information essential for satire detection.

- **Contextual ambiguity and labeling errors**
  Sample `e1a1a563` is particularly ambiguous: a social media-style intervention with *reel*/*short* background music where the tone is informal but content isn't clearly satirical without speaker/-context knowledge. This ambiguity poses challenges even for human annotators and suggests some errors may stem from label quality issues or lack of explicit contextual information.

Table 11
Classification errors of best model on validation set.

ID	Error Type	Transcription
`3db0b886-434a22fd`	False positive	El presidente de Estados Unidos, Barack Obama, prometió un año de acción con el fin de revertir la desigualdad social e impulsar la economía en su país. Durante su discurso sobre el Estado de la Unión, Obama también defendió su programa de salud y dijo que Estados Unidos debe dejar de estar en pie de guerra.
`8f7e5f21-d3edaf72`	False negative	Forzado por su familia, un nostálgico se ha decidido al fin a reciclar el botellín de cerveza que estaba bebiendo cuando Iniesta marcó el gol del Mundial. Natalio Felote llevaba desde 2010 guardando el botellín como si fuera la mismísima Copa del Mundo.
`e29b87d0-628166cf`	False negative	El gobierno admite que los test del coronavirus que consisten en lanzar una moneda al aire son sólo fiables al 50%. Sanidad pagó 10 euros por cada moneda de euro de este test del coronavirus. El fabricante, de origen chino, no contaba con la homologación y era el mismo al que se compraron los 3 millones de bolsas de plástico para la cabeza que se usaron cuando no había suficientes respiradores.
`607a93b9-4a3b098e`	False negative	En fin, amigos, debatir es bueno. Y quizá sea hora de que como país empecemos a hacerlo de forma sosegada sobre algunas tradiciones un tanto bochornosas y caducas, digamos. Es más, puestos a soñar, me gustaría sentar en la mesa de este programa a un toro y a alguien que disfruta torturándolo para que lleguen a un entendimiento, digamos.
`ca6bda3a-3cfca4c5`	False negative	Y como a Andrea Chávez no pueden pasar ni 24 horas sin hacer el ridículo frente a todos, inmediatamente en las redes se le fueron a reclamar a Maguicha que hace unas semanas había dicho que el Poder Legislativo estaba plagiado de nepotismo. Hija, ¿con qué cara dices eso si tú tienes al mismísimo Botijai junto a ti?
`ab2e18c4-0100df14`	False negative	O de... Exacto, pero bueno. Así es las cosas y bueno, van a seguir con estos careos y van a seguir con el juicio donde le piden 200 millones de pesos a Carlos Loret por un video que injustamente puso en donde se ve al hermano del presidente recibiendo sobres de dinero. Va a haber un corte comercial y regresa. Ustedes, hermanita, no se muevan porque pasó todavía más cosa.
`e1a1a563-73ebf7f1`	False negative	A ver, sacarse unas oposiciones en sí es muy complicado. Es mucho esfuerzo, es agotador, súmale el baile. Tengo un entrenador personal para esto. Me considero un poco la Nati Peluso de los opositores a administración del Estado.
`5b97747c-a4aca1b4`	False negative	¿Cómo se consigue que un niño sonría? Pues con una consola que es carísima. Pero, ¿opina lo mismo en un barrio rico que en un barrio obrero? ¿Piensan como yo que el dinero da la felicidad? Según dos premios Nobel en Economía, dicen que para ser feliz hay que ganar 100.000 euros al año. Teniendo en cuenta ese dato de 100.000 euros al año, ¿es usted feliz?
`eebdb176-23f41919`	False positive	La Corte Internacional de la Haya resolvió la disputa marítima entre Chile y Perú ratificando la línea fronteriza existente hasta las 80 millas, basado en el hito 1 entre ambos países. La Corte también estableció una nueva frontera equidistante a partir de ese punto hasta las 200 millas marítimas.

5.5.1. Common Errors

Applying the same procedure to other top-performing multimodal models, we observed that fragments `ab2e18c4` and `5b97747c` were misclassified by all evaluated models. In both cases, satire manifests primarily in the acoustic channel while transcriptions contain no clear clues.

- `ab2e18c4` contains laughter and onomatopoeia not reflected in text.
- `5b97747c` relies on ironic intonation for satire interpretation, especially in the opening phrase. Additionally, knowledge of the broadcast format (political comedy programs) is crucial for correct interpretation.

These errors highlight models' difficulty capturing implicit satire, which depends on both subtle acoustic elements and external contextual information. From this analysis, we derive:

- Errors are systematic rather than random, especially when satire:
  - Is conveyed through subtle intonation/prosody.
  - Is absent from transcriptions.
  - Requires external knowledge.
- Models struggle with implicit satire lacking explicit textual cues.
- Certain sound effects (e.g., background music) may induce false positives if erroneously correlated with satire during training.

5.6. Interpretation and Analysis

5.6.1. Classifiers

Comparative analysis shows relevant performance differences across classifiers, representations, and fusion strategies. Generally, support vector machines (SVM) demonstrated more robust behavior, systematically outperforming dense neural networks (DNN) in most configurations. This may stem from SVMs' capacity to maximize inter-class separation margins in high-dimensional spaces, particularly effective with our representations. DNNs, though slightly less accurate on average, showed lower variance between runs, indicating greater result stability.

Logistic regression performed notably worse in demanding scenarios, especially with HuBERT-based tasks, due to its linear nature. However, its low computational cost made it useful as an auxiliary classifier for tasks like weight estimation in weighted fusion, prioritizing efficiency over final performance.

5.6.2. Text Representations

For text representations, results confirm contextualized models' superiority over static alternatives. Specifically, `RoBERTa-bne` delivered the best representations for satire detection, significantly outperforming classical models like `FastText` or `Word2Vec`. This was expected since `RoBERTa` captures full utterance context, including ironic structures and pragmatic nuances difficult to represent with non-contextual models.

`FastText` showed competitive standalone performance, especially with SVM or DNN classifiers, benefiting from its ability to handle rare/out-of-vocabulary words via sublexical components. However, its combination with `RoBERTa` yielded no substantial improvements, suggesting contextualized representations already contained most useful signals.

Conversely, `Word2Vec` showed the worst performance both standalone and combined, reinforcing that static representations, even enriched through aggregation techniques, are inadequate for capturing satire's linguistic mechanisms.

5.6.3. Audio Representations

For audio, a similar pattern emerged with clearly superior performance from advanced model representations. HuBERT consistently outperformed Wav2Vec2 by average margins of $5 \sim 8\%$ in *Macro F1-Score* depending on extraction strategy. This advantage may stem from its masked phonetic unit prediction training, favoring capture of prosodic/phonological structures relevant to this task.

In both cases, mean temporal representation (*mean pooling*) outperformed the CLS vector, suggesting global aggregation preserves information distributed throughout the signal, including intonational and temporal patterns key to identifying ironic/satirical content.

Moreover, combining these representations with traditional acoustic descriptors like MFCCs and prosodic features (pauses, energy, pitch) further improved performance. The optimal configuration fused HuBERT-mean with prosodic MFCCs using an SVM classifier. This demonstrates effective complementarity: self-supervised models provide rich contextual information while handcrafted descriptors incorporate timbre/prosody nuances enriching the joint representation. Attributes like voice melody, pause duration, or intensity modulation, fundamental for oral satire expression, are thus better reflected in the final feature space.

5.6.4. Fusion Methods

Regarding multimodal fusion strategies, the evaluated methods (concatenation, mean, weighted sum, attention) yielded similar overall performance, suggesting the key lies in joint modality exposure rather than specific integration techniques. No clearly superior technique emerged, though concatenation and attention slightly led at the distribution's upper end.

Concatenation achieved the highest *Macro F1-Score* in official evaluation when used with SVM. This strategy, preserving all information from both modalities in an extended vector, allowed SVM to effectively leverage generated feature richness, showing particular robustness in high-dimensional, well-balanced input spaces.

Attention fusion showed equally competitive performance. Its design dynamically prioritized the most relevant acoustic components based on textual content, without requiring explicit sequential structure. This mechanism, implemented via multi-head attention with text as query, proved capable of modeling useful cross-modal relationships, emphasizing intonational/rhythmic aspects associated with verbal irony.

The parity between attention and concatenation indicates both strategies effectively exploit modal complementarity, either letting the classifier discover relevant interactions or guiding integration through the fusion mechanism. Even simpler approaches like arithmetic mean yielded competitive results, reinforcing that both modalities provide coherent complementary information usable with minimally parametrized techniques.

Overall, results demonstrate multimodal combination not only improves average scores but significantly reduces run-to-run dispersion, providing greater robustness and stability than monomodal configurations regardless of the specific fusion method applied.

5.6.5. System Limitations

Beyond aggregate results, error analysis reveals recurring patterns explaining current system limitations.

First, errors aren't randomly distributed but concentrate in specific sample types with particular characteristics. Among false negatives, prominent cases involve satire expressed through subtle prosodic elements or requiring implicit contextual knowledge for correct interpretation. Segments with formal tone or news structure from sources like *El Mundo Today* or *El Intermedio* were misclassified as non-satirical despite clearly satirical discourse content. These examples highlight models' difficulty capturing communicative intent when explicit markers are absent from textual or acoustic signals.

Additionally, false positives occurred in fragments from Latin American news, where background music, sound effects, or marked accents may have induced misclassification through associations learned during training. Such errors suggest deceptive correlations in data, where secondary elements like

delivery style, background music, or certain acoustic features may have been learned as indirect satire indicators without being inherently so.

These errors systematically arise in underrepresented genres or contexts where satire relies on implicit prosodic cues or cultural knowledge absent from training data. Overcoming these may require incorporating external context information and improving acoustic representations' sensitivity to fine prosodic nuances still partially missed by current models.

6. Conclusions

6.1. Proposed System

The proposed system addresses automatic satire detection in Spanish through a multimodal approach, integrating textual and acoustic information via dense representations and efficient classifiers.

For the textual modality, various pre-trained models were evaluated, with RoBERTa-bne, specifically trained for Spanish, standing out, alongside representations based on FastText and Word2Vec constructed through average aggregation. For the audio signal, both traditional acoustic features (MFCCs and prosodic traits like pauses, intonation, or energy) and embeddings from self-supervised models (HuBERT and Wav2Vec2) were explored, using both CLS token extraction and mean pooling.

To fuse representations from both modalities, we evaluated several strategies: concatenation, arithmetic mean, weighted sum, and multi-head attention mechanisms. For classification, support vector machines (SVM) and dense neural networks (DNN) were primarily used, with logistic regression discarded in advanced stages due to inferior performance.

Systematic exploration of these combinations identified highly competitive configurations during internal validation, particularly attention-based fusion between FastText and HuBERT+MFCCs with prosodic features using SVM, though differences among top models were minimal. However, in the official evaluation on the blind test set, the highest score was achieved by combining RoBERTa-bne and HuBERT+MFCCs with prosodic features, using concatenation fusion and SVM classification, securing first place in the multimodal task of the SatiSPeech 2025 competition. In the monomodal text task, the system also performed exceptionally, ranking third using SVM on RoBERTa-bne.

6.2. System Limitations

Despite its performance, the developed approach has limitations stemming from computational constraints and architectural decisions. Hardware restrictions and processing time limited the use of pre-trained models to feature extraction mode only, preventing *fine-tuning* on architectures like RoBERTa-bne or HuBERT. This hindered adaptation to the satire detection domain, likely affecting the capture of subtle pragmatic or prosodic patterns. The inability to train task or context-specific representations also precluded exploring larger, deeper, or specialized architectures.

Architecturally, we opted for late multimodal integration based on concatenation or direct combination of independently extracted embeddings, followed by standalone classification. Though computationally efficient, this approach cannot explicitly model cross-modal interactions during training, potentially limiting the system's ability to capture complex multimodal signals or dependencies between textual content and acoustic signals. More sophisticated alternatives, like end-to-end architectures or cross-modal attention were excluded due to these constraints.

Finally, some components exhibit limited linguistic coverage. The acoustic model HuBERT, while high-performing, was not originally trained on Spanish data, potentially hindering its ability to capture phonological nuances, regional accents, or language-specific prosodic patterns. This lack of linguistic specialization, both textual and acoustic, may have reduced the system's sensitivity to culturally marked or subtle satire.

6.3. Impact on Results

These limitations directly impacted system behavior and results. Without fine-tuning, pre-trained models could not adapt to task-specific nuances of text/audio representations, likely affecting the capture of subtle semantic or prosodic patterns. For text, this lack of specialization may explain why performance, though competitive, did not surpass other solutions likely using fine-tuning on similar models.

Similarly, for audio and multimodal configurations, using generic embeddings combined with late fusion strategies like concatenation may have limited the system's ability to distinguish ambiguous or complex examples. Qualitative analysis of mispredictions suggests that without explicit context modeling, the system relies on superficial audio cues: certain voice tones, paralinguistic elements (e.g., laughter), or background sound effects that, due to their recurrence in satirical training examples, may have been misinterpreted as reliable satire indicators.

This dependence on deceptive correlations or unintended artifacts compromised robustness for out-of-distribution cases. Additionally, the absence of joint fusion architectures prevented full exploitation of text-audio complementarity. In examples where irony manifests only through intonation or content form contrast, the system lacks mechanisms to effectively integrate and enhance these signals.

In summary, while the approach delivered solid performance in both challenge tasks, computational and architectural constraints likely limited its potential. Incorporating fine-tuned models and deeper, jointly trained fusion strategies would better capture pragmatic and multimodal satire nuances, reduce errors, and improve generalization to new domains or discourse styles.

6.4. Ethical Considerations

Developing automated systems for complex linguistic phenomena like satire entails ethical implications beyond technical dimensions. Unlike objective tasks (e.g., topic classification or basic sentiment analysis), satire involves pragmatic and culturally loaded components that vary significantly across contexts, speakers, and communities [13, 12]. Integrating such tasks into sensitive applications thus requires critical reflection on potential risks, especially when deployed in real-world scenarios where decisions may have tangible impact [34, 35].

A representative risk emerges when using satire detectors as auxiliary tools in misinformation detection systems [36, 37, 14, 35]. While distinguishing humor from deliberate deception is legitimate, avoiding penalization of legitimate satire, ambiguities persist. A system might mislabel satirical content as fake news, censuring or discrediting humor that serves critical public discourse functions. Similarly, disinformation agents could exploit this inverse bias by superficially incorporating ironic elements to evade detection. Both scenarios illustrate how overreliance on satirical indicators as misinformation filters may yield consequential false positives/negatives.

Moreover, the strong contextual dependence of humor and irony must be considered. Satire relies on cultural conventions, shared references, and linguistic codes that are not always equivalent across communities [11]. Systems trained on specific corpora may reflect a partial view of the phenomenon, erratically classifying expressions deviating from dominant patterns. Such biases could invisibilize legitimate satire from underrepresented groups or perpetuate inequalities through systematically skewed classifications [35, 13].

Finally, satire detection must not be equated with falsity detection or malicious intent attribution. Not all satire contains false claims, nor do all false claims adopt satirical framing [14, 35]. The line between humor, critique, and manipulation is often blurred, requiring contextual knowledge that current models cannot reliably emulate. As noted in prior work [34, 36, 37], deploying automated systems in sensitive domains like misinformation detection necessitates human oversight, continuous auditing, and training data updates to prevent both uncritical decision automation and the imposition of implicit biases.

6.5. Future Work

Although the proposed system demonstrated competitive performance, several improvements warrant exploration. First, expanding fusion methods to include early attention mechanisms alongside the late-attention approaches used here [13, 17]. This variant could model cross-modal interactions from earlier architectural stages, particularly beneficial when acoustic signals reinforce textual content from sequence onset.

Additionally, since all models remained frozen, partial or full fine-tuning of pre-trained models [38] would allow adapting learned representations to Spanish satire nuances [39, 14], potentially enhancing discriminative capacity. Larger models (e.g., `RoBERTa-large-bne` [24] or `HuBERT-large` [8]) could capture more complex patterns, albeit at higher computational cost [11].

To balance performance and efficiency, efficient adaptation techniques like adapter layers [40] or *LoRA* (Low-Rank Adaptation) [41] are especially promising. These methods fine-tune pre-trained models by training only a small parameter subset while freezing the original weights. Adapter layers insert low-dimensional modules between original model layers, learning task-specific transformations without altering core weights, proven effective across NLP tasks while matching full fine-tuning performance with fewer trained parameters [40].

Similarly, *LoRA* [41] injects low-rank matrices into linear projections, representing necessary updates via efficient decomposition. This adds adaptability without significant inference latency, crucial for computationally constrained applications, and has been validated for text and audio tasks, matching full fine-tuning results at reduced training cost. Integrating these techniques would enable future iterations to incorporate models specifically adapted to Spanish satire without compromising efficiency in modest computing environments.

Declaration on Generative AI

During the preparation of this work, the author used GPT-4o for translation and spell checking. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the final version of the publication.

References

[1] 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. doi:10.1007/s40747-021-00625-1.

[2] 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. doi:https://doi.org/10.1016/j.csi.2024.103856.

[3] B. W. Schuller, Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends, Communications of the ACM 61 (2018) 90–99. doi:10.1145/3129340.

[4] M. Casals-Salvador, F. Costa, M. India, J. Hernando, BSC-UPC at EmoSPeech-IberLEF2024: Attention Pooling for Emotion Recognition, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th SEPLN Conference, CEUR Workshop Proceedings, CEUR-WS.org, Valladolid, Spain, 2024. URL: https://ceur-ws.org/Vol-3756/EmoSPeech2024_paper1.pdf.

[5] 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. doi:10.26342/2024-73-27.

[6] 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).

[7] 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.
[8] 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. doi:10.48550/arXiv.2106.07447.
[9] Z. K. Abdul, A. K. Al-Talabani, Mel Frequency Cepstral Coefficient and its Applications: A Review, IEEE Access 10 (2022) 122136–122158. doi:10.1109/ACCESS.2022.3223444.
[10] PlanTL-GOB-ES, roberta-base-bne: Pretrained RoBERTa model for Spanish from the National Library Corpus, 2022. URL: https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne.
[11] 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. doi:https://doi.org/10.1016/j.knosys.2021.107597.
[12] F. Barbieri, F. Ronzano, H. Saggion, Is This Tweet Satirical? A Computational Approach for Satire Detection in Spanish, Procesamiento del Lenguaje Natural (2015). URL: https://www.redalyc.org/articulo.oa?id=515751524015.
[13] F. Bellido Delgado, F. d. B. Navarro Colorado, D. Tomás Díaz, Generación de ironía multimodal, Trabajo de Fin de Máster en Ciencia de Datos, Universidad de Alicante, 2024.
[14] N. Mafla, M. Flores, S. Castillo-Páez, R. Andrade, Automatic Detection of Fake News in Spanish: Ecuadorian Political Satire, Revista Politécnica 50 (2022) 7–16. doi:10.33333/rp.vol50n3.01.
[15] A. Kamal, M. Abulaish, Jahiruddin, Contextualized Satire Detection in Short Texts Using Deep Learning Techniques, Journal of Web Engineering 23 (2024) 27–52. doi:10.13052/jwe1540-9589.2312.
[16] K. Alnajjar, M. Hämäläinen, ¡Qué maravilla! Multimodal Sarcasm Detection in Spanish: a Dataset and a Baseline, in: Proceedings of the Third Workshop on Multimodal Artificial Intelligence, Association for Computational Linguistics, Mexico City, Mexico, 2021, pp. 63–68. doi:10.18653/v1/2021.maiworkshop-1.9.
[17] P. Bisht, D. Bisht, A. Srivastava, Multimodal Sarcasm Detection Using Transformer- Based Architectures: A Unified Framework for Text, Audio, and Visual Data, International Research Journal of Education and Technology 07 (2025).
[18] M. d. P. Salas-Zárate, M. A. Paredes-Valverde, M. Á. Rodríguez-García, R. Valencia-García, G. Alor-Hernández, Automatic detection of satire in Twitter: A psycholinguistic-based approach, Knowledge-Based Systems 128 (2017) 20–33. doi:https://doi.org/10.1016/j.knosys.2017.04.009.
[19] R. Awasthi, V. Chavan, Sarcasm Detection Based on Sentiment Analysis of Audio Corpus Using Deep Learning, South Eastern European Journal of Public Health (2024) 785–794. doi:10.70135/seejph.vi.1543.
[20] Z. Li, X. Gao, Y. Zhang, S. Nayak, M. Coler, A Functional Trade-off between Prosodic and Semantic Cues in Conveying Sarcasm, in: Interspeech 2024, ISCA, 2024, pp. 1070–1074. doi:10.21437/Interspeech.2024-1962.
[21] H. Bredin, A. Laurent, End-to-end speaker segmentation for overlap-aware resegmentation, 2021. URL: https://arxiv.org/abs/2104.04045, _eprint: 2104.04045.
[22] 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 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7124–7128. doi:10.1109/ICASSP40776.2020.9052974.
[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust Speech Recognition via Large-Scale Weak Supervision, 2022. URL: https://cdn.openai.com/papers/whisper.pdf.
[24] 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). doi:10.26342/2022-68-3, publisher: Sociedad Española para el

Procesamiento del Lenguaje Natural.
[25] 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, 2020. doi:10.48550/arXiv.1911.02116.
[26] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, arXiv preprint arXiv:1607.01759 (2016).
[27] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).
[28] J. VanderPlas, Python Data Science Handbook: Essential Tools for Working with Data, 2nd ed., O'Reilly Media, Sebastopol, CA, 2022. URL: https://github.com/jakevdp/PythonDataScienceHandbook.
[29] T. Bäckström, O. Räsänen, A. Zewoudie, P. Pérez Zarazaga, L. Koivusalo, S. Das, E. Gómez Mellado, M. Bouafif Mansali, D. Ramos, S. Kadiri, P. Alku, M. H. Vali, Introduction to Speech Processing, 2 ed., 2022. doi:10.5281/zenodo.6821775.
[30] Centro Virtual Cervantes, CVC. Diccionario de términos clave de ELE. Prosodia., 2009. URL: https://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/prosodia.htm, publisher: Instituto Cervantes.
[31] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, 2020. doi:10.48550/arXiv.2006.11477.
[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, in: Advances in Neural Information Processing Systems, arXiv, 2017. doi:10.48550/arXiv.1706.03762.
[33] TensorFlow Developers, tf.keras.layers.multiheadattention | tensorflow v2.16.1, 2024. URL: https://www.tensorflow.org/api_docs/python/tf/keras/layers/MultiHeadAttention.
[34] A. K. Ulpo Carangui, S. Cabrera Almeida, R. Pizarro Matamoros, G. Morocho, El análisis de los sentimientos con Inteligencia Artificial como estrategia de las relaciones públicas, Ciencia Latina Revista Científica Multidisciplinar 8 (2024) 7658–7666. doi:10.37811/cl_rcm.v8i5.14174.
[35] L. M. Romero-Rodríguez, A. L. Valle-Razo, B. Castillo-Abdul, Fake news de humor y sátira y actitudes hacia la política: análisis comparativo del realismo percibido y los sentimientos de eficacia, alienación y cinismo en estudiantes de comunicación, OBETS. Revista de Ciencias Sociales 16 (2021) 465. doi:10.14198/OBETS2021.16.2.15.
[36] O. Levi, P. Hosseini, M. Diab, D. A. Broniatowski, Identifying Nuances in Fake News vs. Satire: Using Semantic and Linguistic Cues, in: Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, 2019, pp. 31–35. doi:10.18653/v1/D19-5004.
[37] 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. doi:10.18653/v1/W16-0802.
[38] O. Razuvayevskaya, B. Wu, J. A. Leite, F. Heppell, I. Srba, C. Scarton, K. Bontcheva, X. Song, Comparison between parameter-efficient techniques and full fine-tuning: A case study on multilingual news article classification, PLOS ONE 19 (2024) e0301738. doi:10.1371/journal.pone.0301738.
[39] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. d. Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-Efficient Transfer Learning for NLP, 2019. doi:10.48550/arXiv.1902.00751.
[40] N. Inoue, S. Otake, T. Hirose, M. Ohi, R. Kawakami, ELP-Adapters: Parameter Efficient Adapter Tuning for Various Speech Processing Tasks, 2024. doi:10.48550/arXiv.2407.21066.
[41] 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. doi:10.48550/arXiv.2106.09685.