A Multimodal Late-Fusion System for Irony Detection in Spontaneous Spoken Spanish

Carlos Díez-Fenoy*

Machine Learning for Data Science Group (ML4DS), Signal Theory and Communications Department (University Carlos III of Madrid), Leganés, 28911, Madrid, Spain

Abstract

This paper describes the system submitted to the SpeechMATICS 2026 shared task on multimodal irony detection in spontaneous spoken Spanish, part of IberLEF 2026. The task is structured into two subtasks: Task 1 requires binary irony classification from transcriptions alone, while Task 2 additionally provides the audio recording. For Task 1, a Spanish RoBERTa model (`roberta-base-bne`) is fine-tuned using layer-wise learning rate decay, a freeze-then-unfreeze training strategy, label smoothing, and decision threshold optimization, achieving first place with a macro F1-score of 0.659. For Task 2, a multimodal fusion network (FusionNet) integrates three parallel streams—contextual text embeddings from the Task 1 encoder, utterance-level representations from a Spanish Wav2Vec2 model, and a 56-dimensional vector of hand-crafted prosodic and spectral features—through a late-fusion MLP with a residual connection. The FusionNet output is combined with the Task 1 predictions via logistic stacking, achieving second place with a macro F1-score of 0.697. An exploratory data analysis conducted prior to modeling reveals that ironic utterances tend to be shorter, are associated with specific evaluative lexical markers, and exhibit higher onset strength in the acoustic signal—the single statistically significant acoustic discriminator identified ($p = 0.022$, $\eta^2 = 0.357$). The improvement of 3.8 percentage points from Task 1 to Task 2 provides empirical evidence that the acoustic modality contributes discriminative information beyond what is recoverable from the transcript alone. The code is publicly available at https://github.com/CDF51342/SpeechMATICS-dev.

Keywords

irony detection, spoken Spanish, multimodal classification, transformer fine-tuning, wav2vec2, prosodic features, late fusion, IberLEF 2026

1. Introduction

Irony is a pervasive rhetorical device in human communication, characterized by the intentional expression of a meaning that is opposite to, or systematically incongruent with, the literal content of an utterance [1]. Its interpretation depends not only on lexical and syntactic cues but also on prosodic patterns, contextual knowledge, and shared background between speaker and listener. From a pragmatic standpoint, irony has been analyzed as a form of *echoic mention* [2], in which the speaker implicitly signals dissociation from an expectation or a prior utterance. This mechanism makes irony particularly difficult to detect automatically, since any system relying solely on surface-level textual features risks ignoring the prosodic and paralinguistic signals that are central to its expression in spoken language.

The challenge is compounded in Spanish, where spoken irony frequently exploits intonation patterns—such as sustained high pitch or exaggerated pitch range—as well as paralinguistic cues such as voice quality modifications and specific rhythmic patterns [3]. These signals may reinforce, contradict, or even replace the textual signal, meaning that a system operating on transcripts alone may systematically fail on instances where irony is encoded primarily in the acoustic channel.

The automatic detection of irony in spoken Spanish has concrete societal relevance. In opinion mining and sentiment analysis, irony is a major source of misclassification: a statement such as *"qué maravilla de servicio"* conveys a strongly negative opinion that surface-level polarity classifiers label as positive [4]. Irony detection also matters for platform content moderation, where ironic formulations are

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

carlosdi@pa.uc3m.es (C. Díez-Fenoy)

0009-0008-5225-3575 (C. Díez-Fenoy)

deliberately used to circumvent filters [5], and for dialogue systems, where irony leads to pragmatically incoherent responses.

Despite growing interest in irony detection in Spanish [5], the field has been historically dominated by text-only approaches due to the scarcity of annotated multimodal corpora. Early shared tasks such as IroSvA at IberLEF 2019 [6] addressed irony detection in written Spanish but left the spoken dimension unexplored. The SpeechMATICS 2026 shared task, part of IberLEF 2026 [7], addresses this gap by introducing a multimodal benchmark grounded in the MESironic corpus [8]—the first publicly available multimodal dataset for irony detection in spontaneous spoken Spanish.

This paper describes the system developed for both subtasks. For Task 1, a Spanish-language RoBERTa model is fine-tuned on transcriptions. For Task 2, this textual encoder is combined with a large pre-trained speech model and hand-crafted acoustic features within a multimodal fusion architecture (FusionNet), whose output is ensembled with the Task 1 predictions via logistic stacking. The system achieves first place in Task 1 and second place in Task 2. The full code is publicly available at https://github.com/CDF51342/SpeechMATICS-dev.

The paper is organized as follows. Section 2 describes the task, metric, and dataset. Section 3 presents the exploratory data analysis. Section 4 details the models. Section 5 reports the results. Section 6 concludes.

2. Task Description

2.1. Task Overview

SpeechMATICS 2026 [9] is a shared task within the IberLEF 2026 evaluation forum [7], co-located with SEPLN 2026. It addresses irony detection in spontaneous spoken Spanish through a two-task framework. Both tasks require binary classification of each instance as *ironic* or *non-ironic*.

Task 1: Text-Only Irony Detection. Participants receive the manual transcription of each recording and must classify it without access to the audio signal. The goal is to establish a strong textual baseline and quantify how much ironic intent is recoverable from the transcript alone.

Task 2: Multimodal Irony Detection. Participants receive both the transcription and the corresponding `.flac` audio recording, and may exploit acoustic, prosodic, and paralinguistic information. Task 2 is the primary challenge of the shared task.

2.2. Evaluation Metric

Both subtasks are evaluated using the macro-averaged F1-score:

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{C} \sum_{c=1}^{C} \frac{2 \cdot P_c \cdot R_c}{P_c + R_c} \tag{1}$$

where C is the number of classes and P_c, R_c are the per-class precision and recall computed in a one-vs-rest fashion [10]. The macro average weights all classes equally regardless of frequency, preventing trivial majority-class solutions.

2.3. Dataset

The dataset is MESironic [8], a multimodal corpus for irony detection in spontaneous spoken Spanish. Each instance consists of a short utterance, its manual transcription, and a binary label. The utterances come from naturalistic conversational speech, implying considerable variability in speaking style, recording conditions, and explicitness of the ironic signal—a substantially harder setting than acted or prompted speech corpora.

The training partition contains 5,100 labeled instances: 2,587 ironic (50.7%) and 2,513 non-ironic (49.3%), yielding a near-perfectly balanced distribution (Figure 1). The unlabeled test partition was used for the official evaluation.

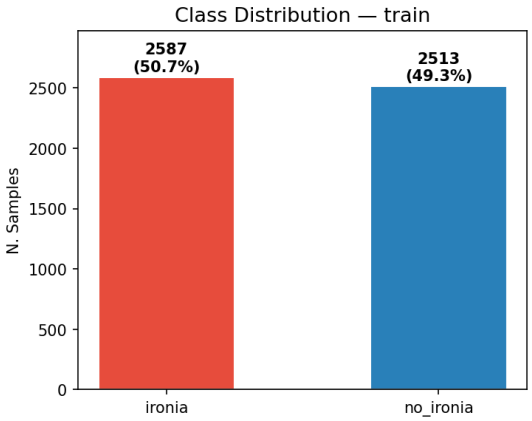

Figure 1: Class distribution in the training set.

3. Exploratory Data Analysis

A systematic exploratory analysis was conducted on both modalities before designing the classification systems. The objectives were: (1) to characterize distributional properties of the corpus relevant to model design, (2) to identify features that differ between classes and could inform feature engineering for Task 2, and (3) to assess whether the ironic signal is recoverable from the acoustic channel in isolation.

3.1. Text Analysis

Ironic utterances are more concentrated in the short-length range, while non-ironic utterances show a higher proportion of longer texts (Figure 2). This asymmetry suggests that irony in spontaneous speech tends to be expressed concisely—a short incongruous remark is often more effective than an extended one [11].

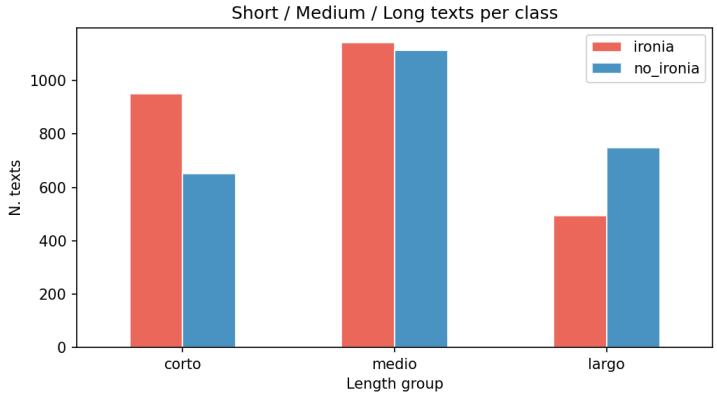

Figure 2: Utterance length distribution (short/medium/long) by class.

After stopword removal, ironic utterances are disproportionately associated with evaluative expressions (e.g., *claro*, *igual*) and second-person forms (*sabes*, *tienes*), which are characteristic of direct-address ironic constructions in colloquial Spanish. The most discriminating bigram, *claro claro*, is a canonical marker of sarcastic agreement [3] (Figures 3 and 4).

A log-odds TF-IDF analysis [12] confirms that the most discriminative ironic terms are contextually specific and long-tailed (Figure 5), which motivates the use of contextual language models over bag-of-words representations.

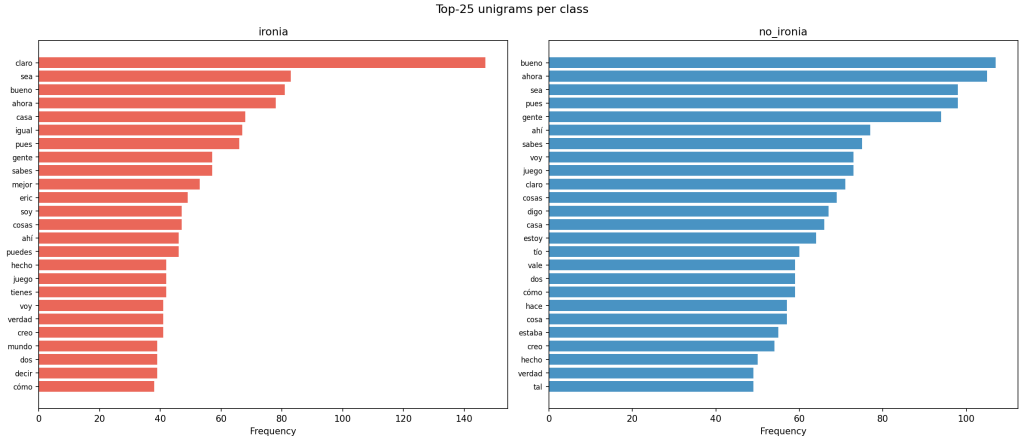

Figure 3: Top-25 unigrams per class (stopwords removed).

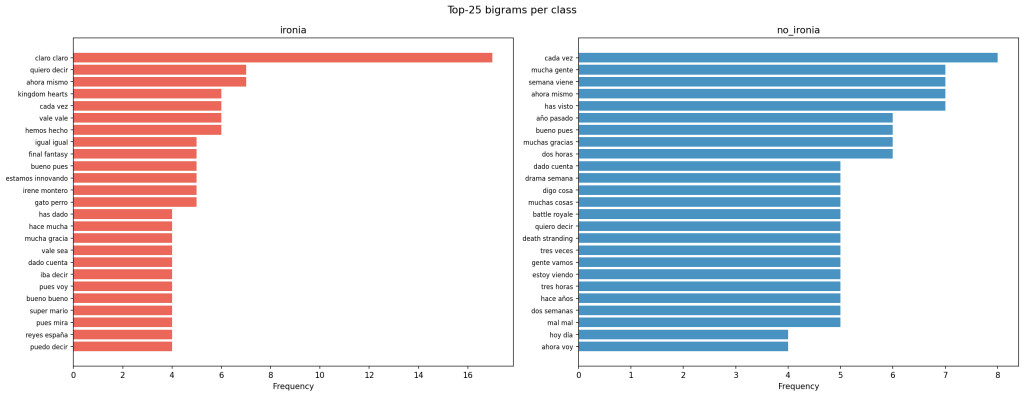

Figure 4: Top-25 bigrams per class (stopwords removed).

A particularly informative subset consists of 61 minimal pairs—instances sharing the same transcription but carrying different labels—where the ironic interpretation depends entirely on the acoustic realization. This subset is used as a controlled experiment for isolating acoustic contributions, and all acoustic analyses in Section 3.2 are performed on it to eliminate lexical confounds.

3.2. Acoustic Feature Analysis

Acoustic features were extracted using `librosa` [13] at 16 kHz, comprising temporal descriptors (duration, RMS energy, zero-crossing rate, silence ratio), F0 descriptors (mean, median, standard deviation, range, and end-slope), spectral descriptors (centroid, rolloff, bandwidth, flatness), 13 MFCCs and their first-order deltas, and rhythmic descriptors (tempo, onset strength mean and standard deviation).

Figure 6 shows the point-biserial correlation of each feature with the class label across the full training set. The feature most positively correlated with the ironic class is *mfcc_1* ($r \approx 0.15$), followed by *rms* ($r \approx 0.10$) and *zcr* ($r \approx 0.05$). Features negatively correlated with irony—and therefore more associated with the non-ironic class—include *duracion_s*, *mfcc_4_std*, and *silence_ratio*. These correlations are modest in magnitude, which is consistent with the inherent difficulty of detecting irony from acoustic features alone in naturalistic speech: no single descriptor constitutes a reliable classifier, but the feature set as a whole encodes complementary discriminative signal across energy, spectral shape, and temporal dynamics.

The correlation matrix in Figure 7 reveals the internal structure of the feature set. The strongest positive correlation is between *rms* and *mfcc_1* ($r = 0.65$), reflecting the well-known relationship between signal energy and the zeroth cepstral coefficient. A strong negative correlation is observed

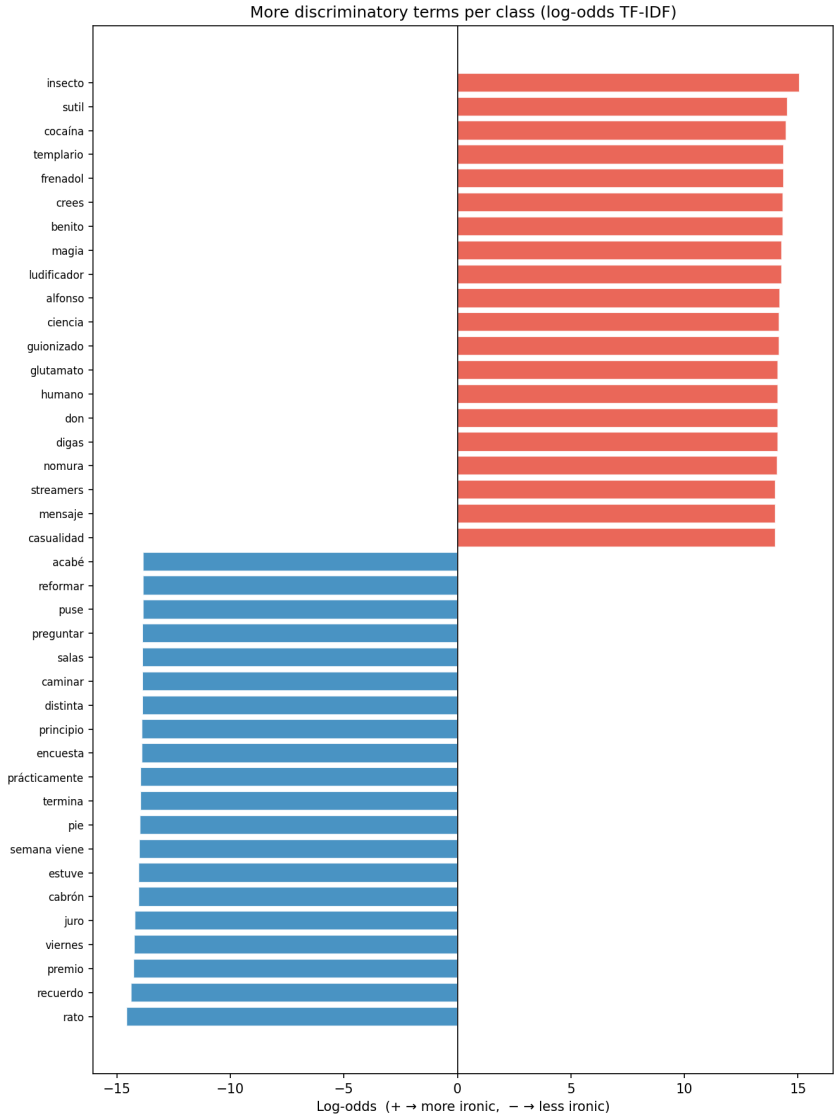

Figure 5: Most discriminative terms by class under log-odds TF-IDF weighting.

between *zcr* and *mfcc_2* ($r = -0.70$), consistent with the spectral distribution encoded by low-order MFCCs. Additional moderate correlations exist between *mfcc_3_std* and *mfcc_2_std* ($r = 0.52$), and between *silence_ratio* and *mfcc_1* ($r = -0.63$). These interdependencies confirm that the raw feature vector contains redundant dimensions, motivating the use of a `RobustScaler` [14] prior to fusion to reduce the influence of outliers and decorrelate feature scales.

3.3. Summary and Design Implications

Three findings drive the system design. First, discriminative lexical cues in the text are sparse and domain-specific, favoring contextual transformers over bag-of-words models. Second, 61 instances in the corpus share identical transcriptions across ironic and non-ironic recordings, establishing a lower bound on the proportion of examples that cannot be disambiguated from text alone and for which the audio signal is strictly necessary. Third, the acoustic analysis shows that no single feature is a strong classifier in isolation—the highest correlation with the label is $r \approx 0.15$ for *mfcc_1*—but that energy, spectral, and temporal descriptors collectively encode complementary discriminative information. Together, these three findings justify a multimodal architecture that integrates both modalities rather than relying on either one alone.

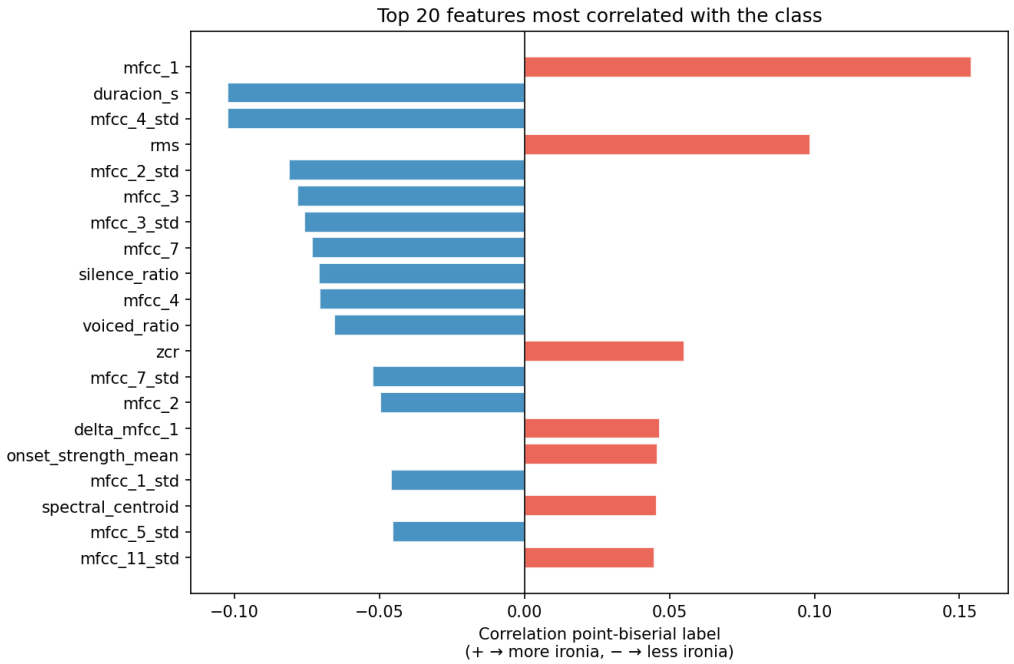

Figure 6: Point-biserial correlation of the top-20 acoustic features with the class label on the full training set. Positive values indicate association with the ironic class; negative values with the non-ironic class.

4. Models and Approaches

The system consists of a fine-tuned text classifier for Task 1 and a multimodal fusion network (FusionNet) for Task 2, combined via logistic stacking.

4.1. Task 1: Text-Only Irony Classifier

4.1.1. Model Selection

The backbone is `IsGarrido/roberta-base-bne`[1], a RoBERTa-base model [15] pre-trained on 570 GB of Spanish text under the MarIA initiative [16]. Its broad domain coverage was preferred over Twitter-specific alternatives such as `robertuito` [17], as the MESironic utterances come from conversational speech rather than social media; a preliminary validation confirmed this choice empirically.

4.1.2. Fine-Tuning Strategy

A linear classification head is appended on top of the `[CLS]` representation. Four complementary techniques improve generalization. **Layer-wise learning rate decay** [18] assigns smaller learning rates to lower layers, preserving generic linguistic representations while allowing task adaptation in upper layers [19]. **Freeze-then-unfreeze training** first optimizes only the classification head, then fine-tunes the full model end-to-end, reducing the risk of catastrophic forgetting [18]. **Label smoothing** ($\varepsilon = 0.1$) [20] softens one-hot targets to reduce overconfidence on the inherently ambiguous boundary between ironic and non-ironic. **Decision threshold tuning** optimizes the classification threshold on the validation set by exhaustive search over $[0.3, 0.7]$ (50 steps) to maximize the macro F1-score. Training uses AdamW [21] with a cosine learning rate schedule [22], weight decay $\lambda = 0.01$, a maximum of 10 epochs, and early stopping with patience 3.

[1] `IsGarrido/roberta-base-bne` is a community mirror of *PlanTL-GOB-ES/roberta-base-bne*, no longer available on the Hugging Face Hub. https://huggingface.co/IsGarrido/roberta-base-bne

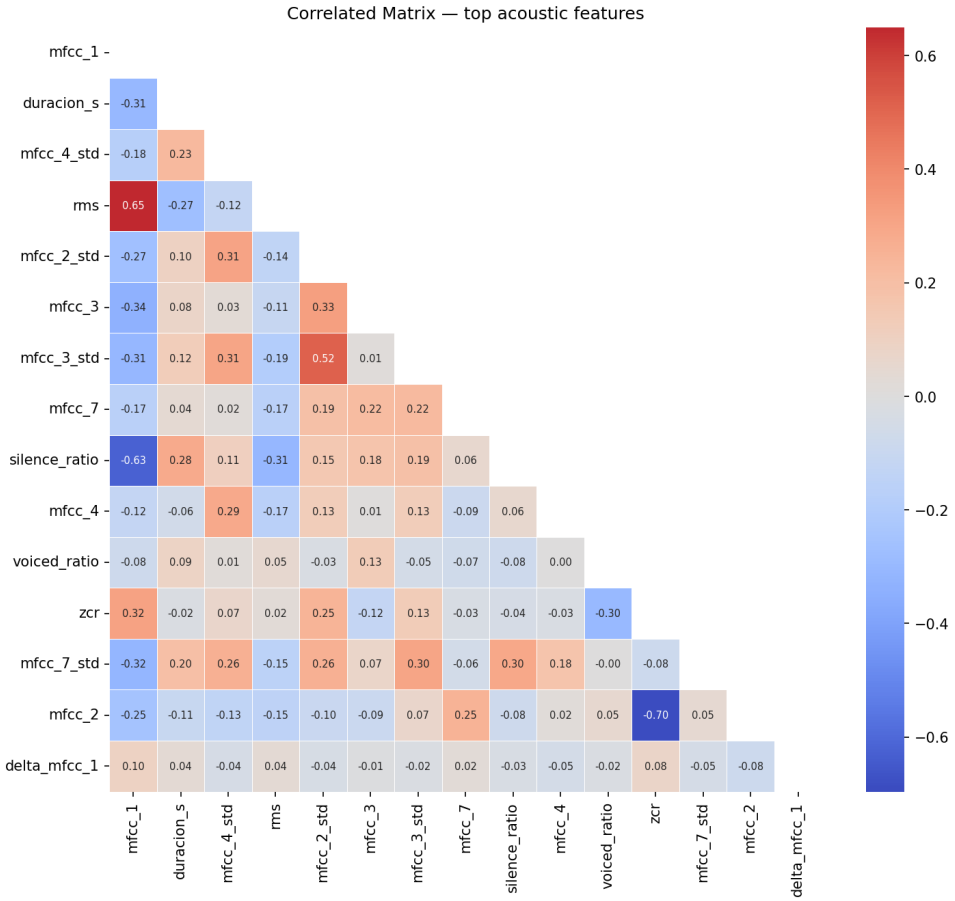

Figure 7: Correlation matrix of the top acoustic features extracted from the training set.

4.2. Task 2: Multimodal FusionNet

FusionNet processes three parallel streams that are projected to a common dimensionality, concatenated, and classified via a two-block MLP with a residual connection (Figure 8).

Text stream. The frozen Task 1 encoder extracts a 768-dimensional `[CLS]` representation, projected to $d_{\text{txt}} = 128$ via a linear layer with layer normalization and dropout ($p = 0.5$). Freezing preserves the discriminative representations from Task 1 and reduces trainable parameters.

Audio stream. `jonatasgrosman/wav2vec2-large-xlsr-53-spanish` [23]—built on wav2vec 2.0 [24] and cross-lingually pre-trained across 53 languages [25]—produces frame-level representations that are mean-pooled [26] into a 1,024-dimensional utterance embedding, projected to $d_{\text{aud}} = 256$ with layer normalization and dropout ($p = 0.55$). Wav2Vec2 weights are also kept frozen.

Hand-crafted acoustic stream. A 56-dimensional vector of prosodic and spectral features (13 MFCC means, 13 MFCC standard deviations, 13 delta-MFCC means, RMS energy, ZCR, silence ratio, voiced ratio, pitch mean/median/std/range/end-slope, spectral centroid/rolloff/bandwidth/ flatness, tempo, onset strength mean/std) directly encodes the features identified as most discriminative in the EDA—particularly onset strength and pitch descriptors—which wav2vec 2.0, optimized for speech recognition, may not represent explicitly. Features are standardized with a `RobustScaler` [14] and projected to $d_{\text{acust}} = 128$ with layer normalization and dropout ($p = 0.5$).

Fusion and classification. The three representations are concatenated:

$$\mathbf{h}_{\text{fused}} = [\mathbf{h}_{\text{txt}} \parallel \mathbf{h}_{\text{aud}} \parallel \mathbf{h}_{\text{acust}}] \in \mathbb{R}^{512} \tag{2}$$

Late fusion [27] is preferred over early or cross-attention-based intermediate fusion [28], as it allows modality-specific representations to develop independently and avoids the overfitting risk of attention

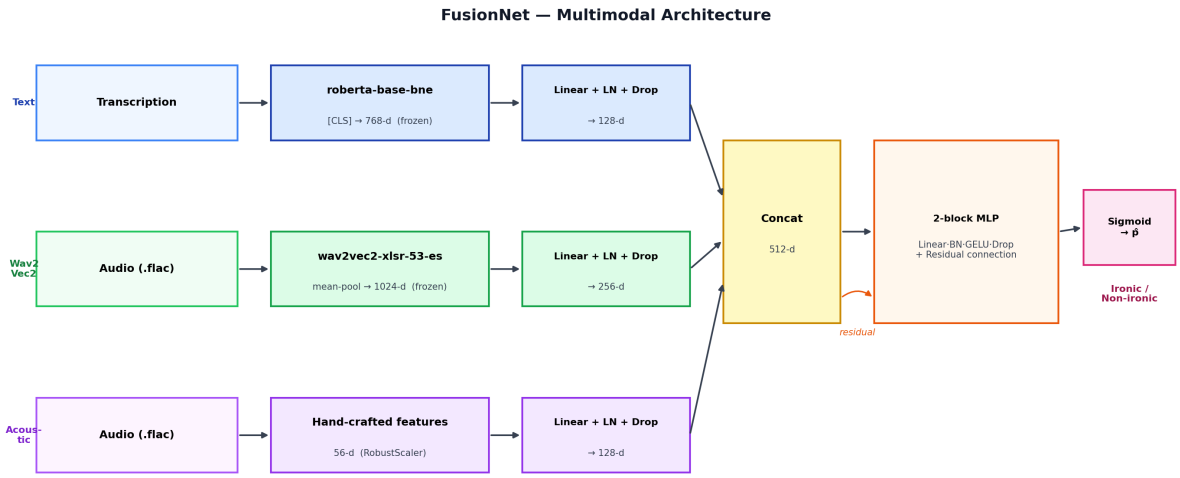

Figure 8: FusionNet architecture. Three parallel streams are projected, concatenated, and fused through a two-block MLP with residual connection.

mechanisms on a moderate-sized dataset. The fused vector is passed through a two-block MLP (linear, batch normalization, GELU [29], dropout) with a residual connection:

$$\mathbf{h}_{\text{out}} = \text{MLP}(\mathbf{h}_{\text{fused}}) + \mathbf{W}_{\text{res}}\,\mathbf{h}_{\text{fused}} \tag{3}$$

The residual connection [30] facilitates gradient flow and allows the model to approximate a linear solution when appropriate. Training uses binary cross-entropy, AdamW, cosine schedule, weight decay $\lambda = 0.1$, and threshold optimization on validation, yielding $\tau^* = 0.496$. The best checkpoint is at epoch 37 ($\text{F1}^{\text{val}}_{\text{macro}} = 0.6705$).

4.3. Logistic Stacking Ensemble

A logistic regression meta-learner [31] is trained on out-of-fold predictions $[\hat{p}_{\text{T1}}, \hat{p}_{\text{T2}}]$ of the Task 1 and Task 2 models, with L_2 regularization. This is preferred over majority voting or averaging because the meta-learner can learn asymmetric weights reflecting complementary error patterns between the two base models, and produces a probabilistic output amenable to a final threshold optimization.

5. Experiments and Results

5.1. Experimental Setup

All experiments use the MESironic training partition [8] with a stratified 85/15 train/validation split. The random seed is fixed for reproducibility. Audio features and Wav2Vec2 embeddings are extracted offline and cached. Models are implemented in PyTorch [32] with the HuggingFace Transformers library [33] and trained on a single GPU.

5.2. Training Dynamics

Figure 9 shows the FusionNet training dynamics. Validation macro F1 improves steadily through the first 25 epochs and then stabilizes, reflecting cosine schedule convergence. The best checkpoint is reached at epoch 37 ($\text{F1}^{\text{val}}_{\text{macro}} = 0.6705$). Training loss decreases monotonically throughout, indicating that the combination of dropout, weight decay, and frozen encoders prevents severe overfitting despite the moderate dataset size.

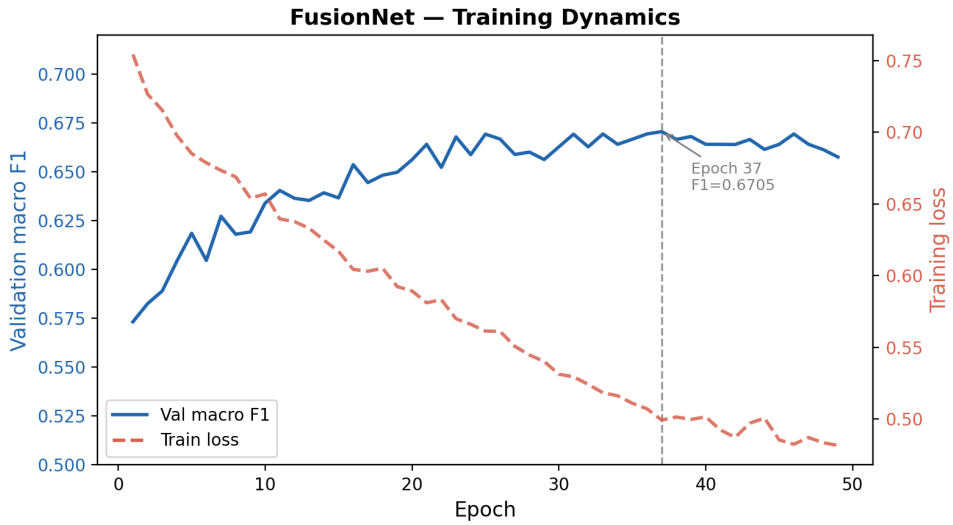

Figure 9: FusionNet training dynamics. Validation macro F1 (left axis) and training loss (right axis) over epochs. The dashed line marks the best checkpoint (epoch 37, $F1^{val}_{macro} = 0.6705$).

5.3. Official Competition Results

Results are reported on the blind test sets released by the SpeechMATICS 2026 organizers, evaluated in terms of macro-averaged precision, recall, and F1-score.

In Task 1, the text-only classifier achieves a macro F1-score of **0.659** (precision 0.660, recall 0.659), ranking first among all participating systems and exceeding the official baseline by 9.8 percentage points (baseline F1 = 0.600). The near-identical precision and recall confirm that the decision threshold optimization effectively balances the two error types across classes.

In Task 2, the multimodal FusionNet with logistic stacking achieves a macro F1-score of **0.697** (precision 0.697, recall 0.697), ranking second overall and exceeding the official baseline by 8.8 percentage points (baseline F1 = 0.609). The improvement of 3.8 percentage points over the Task 1 result obtained by the same system provides direct empirical evidence that the acoustic modality contributes discriminative information beyond what is recoverable from the transcript alone—a finding consistent with the EDA result that onset strength significantly differs between classes ($p = 0.022$, $\eta^2 = 0.357$) and that pitch descriptors show consistent directional trends ($\eta^2 \approx 0.25$–0.30).

A cross-task comparison reveals a ranking inversion between Task 1 and Task 2: the submitted system leads in the text-only setting but is surpassed in the multimodal setting by a margin of 1.1 percentage points. This inversion suggests that the textual encoder is particularly strong, while the multimodal fusion leaves room to better exploit the acoustic signal—consistent with the observation that the acoustic discriminative signal in this corpus is statistically modest in isolation.

6. Conclusions

This paper described the system submitted to SpeechMATICS 2026, a shared task on multimodal irony detection in spontaneous spoken Spanish within IberLEF 2026 [7]. The proposed approach combines a fine-tuned `roberta-base-bne` classifier for Task 1 with a three-stream FusionNet for Task 2, integrating text embeddings, Wav2Vec2 speech representations, and hand-crafted prosodic features through a late-fusion MLP with a residual connection, ensembled with the Task 1 output via logistic stacking.

The system achieves first place in Task 1 (F1 = 0.659, +9.8 pp over the baseline) and second place in Task 2 (F1 = 0.697, +8.8 pp over the baseline). The 3.8 percentage point gain from Task 1 to Task 2 confirms that the acoustic modality contributes discriminative information beyond the text, even though most individual acoustic features do not reach statistical significance in isolation. This result underscores the value of multimodal approaches for irony detection in naturalistic speech and motivates

the development of richer audio representations specifically tailored to prosodic and paralinguistic cues.

Several limitations point to concrete directions for future work. The freeze-then-train paradigm prevents joint end-to-end optimization of the full pipeline, which may limit the capacity to learn representations optimally suited to the irony detection objective. Cross-modal attention mechanisms [28] could capture semantic incongruities between acoustic delivery and textual content more explicitly than the late-fusion strategy employed here. Finally, acoustic representations trained directly on the irony detection objective—rather than on speech recognition as in wav2vec 2.0—may better capture the distributed nature of the prosodic irony signal observed in the exploratory analysis.

Notwithstanding these limitations, the results demonstrate that a carefully designed multimodal system, combining strong pre-trained representations for both modalities with task-specific fine-tuning and ensemble methods, achieves competitive performance on a challenging naturalistic benchmark, and that the acoustic modality provides a consistent and measurable contribution over a text-only baseline.

Declaration on Generative AI

During the preparation of this work, the author used Claude (Anthropic) for grammar and spelling checking and text revision. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] H. P. Grice, Logic and conversation, in: P. Cole, J. L. Morgan (Eds.), Syntax and Semantics, Vol. 3: Speech Acts, Academic Press, 1975, pp. 41–58.

[2] D. Sperber, D. Wilson, Irony and the use–mention distinction, in: P. Cole (Ed.), Radical Pragmatics, Academic Press, 1981, pp. 295–318.

[3] L. Ruiz Gurillo, La ironía en español, Iberoamericana Vervuert, Madrid, 2012.

[4] A. Joshi, P. Bhattacharyya, M. J. Carman, Automatic sarcasm detection: A survey, ACM Computing Surveys 50 (2017) 1–22.

[5] D. I. H. Farías, V. Patti, P. Rosso, Irony detection in Twitter: The role of affective content, ACM Transactions on Internet Technology 16 (2016) 1–24.

[6] R. Ortega-Bueno, F. Rangel, D. I. Hernández Farías, P. Rosso, M. Montes-y Gómez, J. E. Medina Pagola, Overview of the IroSvA 2019 shared task, in: Proceedings of IberLEF 2019, volume 2421 of *CEUR Workshop Proceedings*, 2019.

[7] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS. org, 2026.

[8] V. Ahuir, A. J. Barceló-Milkova, A. Casamayor-Segarra, V. González Verdegais, L.-F. Hurtado, C. Periñán-Pascual, M. J. Castro-Bleda, MESironic: A Multimodal Dataset and Benchmark for Irony Detection in Spontaneous Spoken Spanish, Procesamiento del Lenguaje Natural 77 (2026). To appear.

[9] V. Ahuir, A. Barceló-Milkova, A. Casamayor-Segarra, M. J. Castro-Bleda, L. F. Hurtado, Overview of SpeechMATICS at IberLEF 2026: Speech Multimodal Audio-Text Irony Classification in Spanish, Procesamiento del Lenguaje Natural 77 (2026). To appear.

[10] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing & Management 45 (2009) 427–437.

[11] S. Attardo, Irony as relevant inappropriateness, Journal of Pragmatics 32 (2000) 793–826.

[12] B. L. Monroe, M. P. Colaresi, K. M. Quinn, Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict, Political Analysis 16 (2008) 372–403.

[13] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and

music signal analysis in Python, in: Proceedings of the 14th Python in Science Conference, 2015, pp. 18–25.
[14] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[16] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, MarIA: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). doi:10.26342/2022-68-3.
[17] J. M. Pérez, D. Furman, F. Almeida, F. Luque, RoBERTuito: A pre-trained language model for social media text in Spanish, in: Proceedings of LREC, 2022, pp. 7235–7243.
[18] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of ACL, 2018, pp. 328–339.
[19] I. Tenney, D. Das, E. Pavlick, BERT rediscovers the classical NLP pipeline, in: Proceedings of ACL, 2019, pp. 4593–4601.
[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of CVPR, 2016, pp. 2818–2826.
[21] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: Proceedings of ICLR, 2019.
[22] I. Loshchilov, F. Hutter, SGDR: Stochastic gradient descent with warm restarts, in: Proceedings of ICLR, 2017.
[23] J. Grosman, Fine-tuned XLSR-53 large model for speech recognition in Spanish, 2021. URL: https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-spanish.
[24] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 12449–12460.
[25] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, M. Auli, Unsupervised cross-lingual representation learning for speech recognition, in: Proceedings of Interspeech, 2021, pp. 2426–2430.
[26] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, et al., SUPERB: Speech processing universal performance benchmark, in: Proceedings of Interspeech, 2021, pp. 1194–1198.
[27] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2018) 423–443.
[28] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of ACL, 2019, pp. 6558–6569.
[29] D. Hendrycks, K. Gimpel, Gaussian error linear units (GELUs), 2016.
[30] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of CVPR, 2016, pp. 770–778.
[31] D. H. Wolpert, Stacked generalization, Neural Networks 5 (1992) 241–259.
[32] A. Paszke, S. Gross, F. Massa, et al., PyTorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, volume 32, 2019.
[33] T. Wolf, L. Debut, V. Sanh, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of EMNLP (System Demonstrations), 2020, pp. 38–45.