UCO-UC-CICESE-Plenitas Team at SpeechMATICS: Speech Multimodal Audio-Text Irony Classification in Spanish

Yoan Martínez-López[1,2,*], Yanaima Jauriga[3], Ireimis de las Mercedes Leguen de Varona[3], Miguel Bethencourt[3], Julio Madera[3], Ansel Rodríguez-González[6], Kristell Aguilar Alarcón[1,2], José Carlos Arévalo Fernández[2], Carlos de Castro Lozano[1,2] and José Miguel Ramírez Uceda[1,2]

[1]*EATCO. Universidad de Córdoba, Córdoba, Spain*
[2]*Plénitas, C/ Le Corbusier s/n, 14005 Córdoba, Spain*
[3]*Universidad de Camaguey, Circunvalación Norte, Camino Viejo Km 5 y 1/2, Camaguey, Cuba*
[6]*Centro de Investigación Científica y de Educación Superior de Ensenada, Unidad Académica Tepic, México*

Abstract

This paper describes the participation of the UCO-UC-CICESE-Plenitas team in SpeechMATICS 2026 (Speech Multimodal Audio-Text Irony Classification in Spanish), a shared task at IberLEF 2026 devoted to irony detection in non-formal spoken Spanish. The task is framed as a binary classification problem and organised into two subtasks: a text-only setting (Subtask 1) and a multimodal setting that combines transcriptions and audio (Subtask 2). We propose a single configurable pipeline that integrates three complementary text components—a TF-IDF representation, multilingual sentence embeddings coupled to a class-balanced linear classifier, and a few-shot large language model (LLM) acting as a training-free classifier—together with an acoustic branch based on Mel-Frequency Cepstral Coefficients (MFCCs) for the multimodal subtask. Component predictions are combined through majority voting. An internal comparison of LLM backbones and sentence-embedding models shows that the LLM backbone is the dominant factor, that a smaller well-aligned model (gemma4:31b) outperforms a substantially larger one (qwen3.5:397b), and that the best configuration pairs gemma4:31b with multilingual-e5-large-instruct. On the official evaluation, the system ranks third in Subtask 1 (macro F1 of 0.641, within 0.018 of the best system) and fourth in Subtask 2 (macro F1 of 0.641), outperforming the official baseline in both. A key negative finding is that the MFCC-based feature concatenation yielded no net gain over the textual signal, whereas competing systems improved by 0.038–0.050 macro F1 with audio.

Keywords

irony detection, spoken Spanish, multimodal classification, sentence embeddings, large language models, MFCC, ensemble learning

1. Introduction

Irony is one of the most complex and pervasive phenomena in human communication, characterized by a deliberate mismatch between a speaker's literal words and their intended meaning. Unlike more explicit forms of figurative language, irony frequently depends on subtle and context-dependent cues—prosody, intonation, timing, and shared expectations—that make it difficult to identify even for human listeners. This difficulty becomes considerably more pronounced in spoken interaction, where meaning emerges not from lexical content alone but from the interplay between what is said and how it is said. The failure to correctly interpret irony can produce misunderstandings in everyday exchanges and poses a substantial obstacle for automatic language understanding systems[1, 2, 3].

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
yoan.martinez@plenitas.com (Y. Martínez-López); yanaima.jauriga@reduc.edu.cu (Y. Jauriga); ireimis.leguen@reduc.edu.cu (I. d. l. M. L. d. Varona); betamiguel00@gmail.com (M. Bethencourt); julio.madera@reduc.edu.cu (J. Madera); ansel@cicese.edu.mx (A. Rodríguez-González); kristell.aguilar@plenitas.com (K. A. Alarcón); josecarlos@plenitas.com (J. C. A. Fernández); carlosdecastrolozano@gmail.com (C. d. C. Lozano); miguel@plenitas.com (J. M. R. Uceda)
0000-0002-1950-567X (Y. Martínez-López); 0009-0000-6891-0068 (Y. Jauriga); 0000-0002-1886-7644 (I. d. l. M. L. d. Varona); 0000-0001-8873-6802 (M. Bethencourt); 0000-0001-5551-690X (J. Madera); 0000-0001-9971-0237 (A. Rodríguez-González); 0000-0001-8913-8388 (K. A. Alarcón); 0009-0004-4155-957X (J. C. A. Fernández); 0000-0001-6603-843X (C. d. C. Lozano); 0000-0002-5027-7521 (J. M. R. Uceda)

The ability to detect and interpret irony has therefore become increasingly relevant for a range of applications, including conversational agents, social media analysis, sentiment understanding, and spoken human–computer interaction[4, 5]. Yet most prior research has approached irony as a text-only problem, relying on lexical and syntactic markers while disregarding the acoustic dimension of spoken language. Voice-based approaches, in turn, have been constrained by the scarcity of suitable annotated resources, and this limitation is especially acute for languages other than English. As a result, multimodal irony detection—particularly for spoken Spanish—remains a largely underexplored area of research. Multimodal methods that jointly exploit textual transcriptions and speech signals offer a promising route toward capturing the nuanced, context-sensitive nature of ironic communication, since ironic intent is often conveyed primarily through prosodic or acoustic cues rather than explicit lexical signals. Motivated by this gap, we present SpeechMATICS (Speech Multimodal Audio-Text Irony Classification in Spanish), a shared task organized within IberLEF 2026 [6, 7]. SpeechMATICS addresses the automatic detection of irony in spoken Spanish by combining speech and textual modalities, and is framed as a binary classification problem in which spoken dialogue fragments are labeled as ironic or non-ironic. This formulation allows participants to investigate which features—linguistic, acoustic, or a combination of both—are most indicative of ironic intent. By introducing the first benchmark for voice-based irony detection in Spanish, SpeechMATICS aims to advance research on multimodal language understanding and to encourage the development of methods capable of modeling the complex interaction between speech and language in ironic communication.

2. Methodology

2.1. Subtasks

The SpeechMATICS shared task frames irony detection in non-formal spoken Spanish as a binary classification problem, where each instance is labelled as ironic (1) or non-ironic (0). The task is organised into two complementary subtasks that probe the contribution of different modalities to the same underlying decision. Subtask 1 (text-only irony classification) requires systems to decide whether a transcription conveys ironic content using textual cues exclusively; acoustic information is not permitted. This subtask establishes the upper and lower bounds of purely text-based irony detection and serves as a controlled baseline against which the multimodal setting can be compared. Subtask 2 (multimodal irony classification) allows systems to exploit both the manually reviewed transcription and the corresponding audio signal, so that the additive value of prosodic and acoustic cues over the textual signal can be assessed. Our system addresses both subtasks within a single, configurable pipeline: the text-only configuration feeds Subtask 1, while a text–audio fusion configuration feeds Subtask 2. Predictions for both subtasks are emitted jointly in the required submission format (id, `label_t1`, `label_t2`).

2.2. Evaluation Protocol

Following the official protocol, systems are ranked by the macro-averaged F1-score, with macro precision and macro recall reported as complementary measures. Macro-averaging weights both classes equally regardless of their relative frequency in the test set, which is appropriate given the near-balanced but not identical class distribution of the corpus. The official scores are obtained through bootstrapping with replacement: rather than evaluating a single pass over the test set, multiple resampled subsets are drawn, the metrics are computed on each resample, and the reported leaderboard value is the mean across iterations, yielding more stable and robust performance estimates. For internal model selection and to mitigate over-fitting to the public leaderboard, all feature-based components are tuned with stratified 5-fold cross-validation on the training partition, optimising the macro F1-score. Hyperparameters are searched through grid search over the cross-validation folds, and the resulting cross-validated macro F1 is used to compare candidate embedding models, classifiers, and feature configurations before any submission is generated. This separation between internal validation and the official bootstrapped

evaluation ensures that the design choices reported below were made without leaking information from the test set.

2.3. Dataset

All experiments use the MESironic corpus [8], the first multimodal dataset specifically designed for irony detection in spoken Spanish. The corpus contains 6,003 samples of non-formal Peninsular Spanish, amounting to 5 hours and 49 minutes of audio drawn from publicly available sources, namely podcasts, livestreams, and stand-up comedy fragments. Each sample provides an audio fragment together with a transcription that was generated automatically and subsequently reviewed by hand by a graduate in Hispanic Studies. Annotation follows a rhetorical definition of irony: an utterance is considered ironic when the speaker's intended meaning contradicts the literal meaning of the words. Every sample was labelled independently by three annotators — a linguistics graduate specialised in colloquial Spanish, a computer-science graduate, and one of ten volunteers from diverse academic backgrounds — and the final label was obtained by simple majority vote. The corpus is divided into a training set of 5,100 samples and a test set of 903 samples; both partitions preserve the original label distribution of 50.72% ironic and 49.28% non-ironic instances. The near balance between classes justifies the use of class-balanced classifiers and macro-averaged metrics throughout. In our pipeline, the transcriptions are read from the textual field of the provided CSV files and the corresponding audio files are loaded from the FLAC archives by sample identifier, so that the textual and acoustic views of every instance remain aligned.

2.4. Sentence Embeddings

The core of our text representation is built on multilingual sentence embeddings[9, 10, 11]. Each transcription is encoded into a fixed-length dense vector using a pre-trained sentence-transformer model, which captures semantic and contextual regularities that sparse representations cannot. We surveyed a range of multilingual and Spanish-oriented encoders and selected the final model on the basis of cross-validated macro F1. The candidates included intfloat/multilingual-e5-large-instruct and intfloat/multilingual-e5-large, BAAI/bge-m3, google/embeddinggemma-300M, nomic-ai/nomic-embed-text-v1, and sentence-transformers/paraphrase-multilingual-mpnet-base-v2, among others [12, 13, 14]. The instruction-tuned E5-large variant was retained as the primary encoder, with the E5 family encoded under the recommended query prefix and the remaining models used in their native form. Embeddings are L2-normalised at encoding time, which improves the behaviour of the downstream linear and distance-based classifiers. The resulting embedding matrix is standardised and passed to a conventional classifier. We treat the choice of classifier as a tunable component and consider logistic regression, support-vector machines, multilayer perceptrons, random forests, and gradient-boosted trees, each wrapped in a grid search over its principal hyperparameters with the macro F1-score as the selection criterion and class weights balanced to reflect the corpus distribution. A regularized linear classifier on standardized embeddings provided the most robust trade-off between accuracy and stability across folds and was adopted as the default head for the embedding branch. As a sparse counterpart and lower baseline, we also retain a TF-IDF representation (word n-grams with sub-linear term-frequency scaling) coupled to the same classifier factory, which lets us quantify the benefit of dense semantic embeddings over surface lexical features.

2.5. Large Language Models

In parallel with the embedding branch, we incorporate a generative large language model (LLM) as a complementary, training-free classifier [14, 15, 16, 17]. The LLM is prompted to decide, for each transcription, whether the fragment is ironic under the same rhetorical definition of irony used during annotation. We use a few-shot configuration in which a small, class-balanced set of short training examples is selected and presented as alternating user–assistant turns before the query instance, so that the model is conditioned on canonical examples of each class while keeping the prompt compact and

model-agnostic. The prompt design enforces a strict output contract: the model must emit its decision inside a delimited tag, choosing between exactly two normalised labels. This contract makes the output reliably parseable across heterogeneous models, from small instruction-tuned checkpoints to large reasoning models. A robust parser removes any intermediate reasoning blocks emitted by reasoning-oriented models before extracting the label, and falls back to the conservative non-ironic class when no valid label can be recovered, which avoids spurious positive predictions from malformed generations. Decoding uses a temperature of zero for determinism and a short generation budget with the closing tag as a stop sequence. The pipeline is provider-agnostic: it supports locally served open-weight models as well as a hosted inference endpoint, and the same prompt operates unchanged across model families (for example llama, gemma, and reasoning models such as the qwen family)[18, 19, 17, 16]. Because current text-only LLMs cannot ingest the raw acoustic signal, the LLM contributes to the textual decision in both subtasks rather than processing audio directly.

2.6. System Architecture

The system is implemented as a single configurable pipeline that orchestrates feature extraction, classification, and prediction fusion for both subtasks; the overall workflow is summarised in Figure 1. After loading the aligned text and audio views of the corpus, the pipeline routes each instance through one or more of three text components — the TF-IDF representation, the sentence-embedding branch, and the few-shot LLM classifier — and, for the multimodal subtask, an acoustic branch. The acoustic branch extracts Mel-Frequency Cepstral Coefficients (MFCCs) from each audio file. For every recording we compute 20 MFCCs and summarise them across time by their mean and standard deviation, producing a compact fixed-length acoustic descriptor that captures the spectral envelope of the utterance; the extractor is extensible to additional descriptors such as chroma, spectral contrast, and tonnetz, and it degrades gracefully by zero-filling any file that cannot be processed so that the train/test alignment is preserved. For the multimodal subtask, the text features and the MFCC descriptor are concatenated into a single joint representation that is standardized and classified with the same tuned head used in the text branch, allowing prosodic and lexical–semantic evidence to be combined at the feature level. Final predictions are produced by an ensemble layer that combines the available components through majority voting. In the text-only subtask, the votes of the TF-IDF, embedding, and LLM classifiers are aggregated; in the multimodal subtask, the votes of the text-plus-MFCC fusion configurations are aggregated. Majority voting was chosen for its robustness: it tolerates the occasional failure of any single component (for example a parsing fall-back from the LLM or a corrupted audio file) without destabilizing the overall decision. The class-balanced ensemble outputs are mapped back to the official label strings and written to the submission file, with the text-only decision populating the `label_t1` field and the multimodal decision populating the `label_t2` field, before the file is compressed into the archive expected by the evaluation platform.

3. Results

Table 1 reports the macro F1-score for every combination of LLM backbone and sentence-embedding model in the ensemble.

Three trends emerge. First, the LLM backbone is the dominant factor: averaged over embeddings, gemma4:31b (≈0.630) clearly outperforms qwen3.5:397b (≈0.604) and llama3.1:8b (≈0.478). Notably, the substantially larger qwen3.5:397b model is consistently surpassed by the smaller gemma4:31b, indicating that model family and language alignment matter more than raw parameter count for irony detection in colloquial Spanish. Second, the sensitivity to the embedding choice grows as the backbone weakens: under gemma4 the top four embeddings cluster tightly within 0.633–0.641, whereas under llama3.1:8b the score spans 0.330–0.535 and degenerates with nomic-embed-text-v1 (0.330, near-random). Third, the best configuration, gemma4:31b combined with multilingual-e5-large-instruct, reaches 0.64126, virtually tied with embeddinggemma-300M (0.64101). This internal estimate closely matches our official Task 1 macro F1 (0.641), confirming that model selection did not overfit the test set.

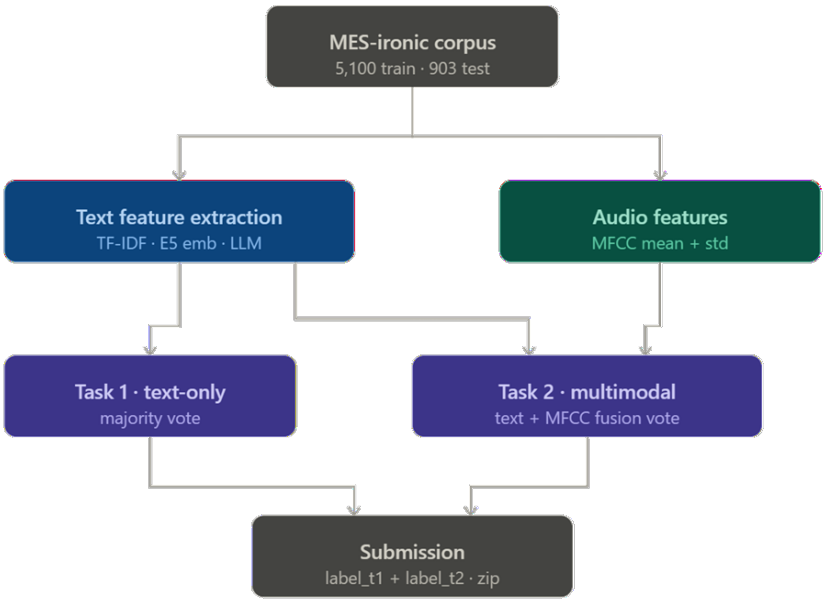

Figure 1: Overview of the proposed SpeechMATICS pipeline.

Table 1
Internal comparisons

Method	Score
gemma4:31b-cloud+intfloat/multilingual-e5-large-instruct	0.64126
gemma4:31b-cloud+sentence-transformers/paraphrase-multilingual-mpnet-base-v2	0.59827
gemma4:31b-cloud+intfloat/multilingual-e5-large	0.63778
gemma4:31b-cloud+google/embeddinggemma-300M	0.64101
gemma4:31b-cloud+BAAI/bge-m3	0.63324
gemma4:31b-cloud+nomic-ai/nomic-embed-text-v1	0.62911
qwen3.5:397b-cloud+intfloat/multilingual-e5-large-instruct	0.60000
qwen3.5:397b-cloud+sentence-transformers/paraphrase-multilingual-mpnet-base-v2	0.59242
qwen3.5:397b-cloud+multilingual-e5-large	0.59604
qwen3.5:397b-cloud+google/embeddinggemma-300M	0.62018
qwen3.5:397b-cloud+BAAI/bge-m3	0.59203
qwen3.5:397b-cloud+/nomic-embed-text-v1	0.62504
llama3.1:8b+intfloat/multilingual-e5-large-instruct	0.51799
llama3.1:8b+sentence-transformers/paraphrase-multilingual-mpnet-base-v2	0.53487
llama3.1:8b+multilingual-e5-large	0.44962
llama3.1:8b+embeddinggemma-300M	0.53487
llama3.1:8b+BAAI/bge-m3	0.50078
llama3.1:8b+/nomic-embed-text-v1	0.33012

In the text-only subtask (Table 2), UCO-UC-CICESE-Plenitas ranks third with a macro F1 of 0.641, within 0.018 of the top score (carlosdf, 0.659) and clearly ahead of HammerTeam (0.622) and the baseline (+0.041, a 6.8% relative improvement). Unlike the other systems, whose precision, recall, and F1 are nearly identical, our system exhibits a monotonic P>R>F1 pattern (0.654/0.646/0.641). Since the macro F1 falls below recall, the gap stems from an uneven per-class precision/recall balance rather than from the textual representation—our precision lies just 0.006 below the leader—pointing to threshold calibration as a direct route to closing the distance to the podium.

In the multimodal subtask (Table 3), UCO-UC-CICESE-Plenitas drops to fourth place with a macro F1

Table 2
Task 1 official results.

Rank	Participant	Run ID	Precision	Recall	F1-score
1	carlosdf	727495	**0.660**	**0.659**	**0.659**
2	migueluc3m	732058	0.658	0.658	0.658
3	**UCO-UC-CICESE-Plenitas**	724787	0.654	0.646	0.641
4	HammerTeam	657292	0.623	0.623	0.622
–	Baseline	754029	0.602	0.602	0.600

of 0.641, identical to its text-only result. Although it still surpasses the baseline (+0.032), the gap to the leader widens to 0.067 (migueluc3m, 0.708)—nearly four times the Task 1 margin.

This behaviour is associated with the fact that the system's F1 remains unchanged when audio is added (0.641→0.641), whereas every competitor improves substantially (e.g. migueluc3m 0.658 →0.708), indicating that the MFCC-based feature concatenation provided no net discriminative value. The rank change therefore reflects the absence of a multimodal gain rather than any degradation. The $P > R > F1$ imbalance persists and is slightly amplified (0.662/0.649/0.641), yet precision stays competitive (within 0.007 of the third-ranked HammerTeam), localising the gap in two separable factors: an uninformative acoustic branch and a recall-reducing class bias.

Table 3
Task 2 official results.

Rank	Participant	Run ID	Precision	Recall	F1-score
1	migueluc3m	732058	**0.710**	**0.709**	**0.708**
2	carlosdf	724959	0.697	0.697	0.697
3	HammerTeam	657286	0.669	0.669	0.668
4	**UCO-UC-CICESE-Plenitas**	713344	0.662	0.649	0.641
–	Baseline	754029	0.610	0.609	0.609

Across both official subtasks, UCO-UC-CICESE-Plenitas consistently outperforms the baseline, demonstrating the robustness of the proposed ensemble across evaluation settings. Beyond the competitive rankings obtained in the shared task, two observations emerge from the results.

First, the multimodal extension fails to improve over the text-only configuration. While competing systems achieved noticeable gains when incorporating audio information, our performance remained unchanged. This suggests that the adopted MFCC-based representation and feature-level fusion strategy did not provide additional discriminative information beyond that already captured by the textual components. Several factors may explain this outcome. First, MFCCs mainly characterize the spectral envelope of speech and may not adequately capture prosodic phenomena that are particularly relevant for irony, such as intonation patterns, pitch dynamics, speaking rate, or expressive timing. Second, the acoustic information was reduced to global mean and standard deviation statistics, which removes temporal information that may be crucial for identifying ironic delivery. Third, the majority-voting fusion strategy treats all components equally and does not account for differences in confidence or reliability across modalities, potentially allowing the stronger textual components to dominate the final decision. Finally, the absence of probability calibration may further limit the effective integration of acoustic and textual evidence. The result highlights that simply adding acoustic descriptors is not sufficient to guarantee multimodal gains and that the way textual and acoustic information are represented and combined may be more important than the availability of an additional modality itself.

Second, the system consistently exhibits higher precision than recall in both subtasks. Since precision remains competitive with the top-ranked systems, the remaining performance gap appears to be associated primarily with recall balance rather than with the quality of the textual representations. This observation suggests that future improvements may be obtained through better calibration and fusion strategies rather than through replacing the core textual components.

4. Conclusions

We have presented the UCO-UC-CICESE-Plenitas system for SpeechMATICS 2026, an IberLEF shared task on irony detection in non-formal spoken Spanish. The proposed pipeline combines TF-IDF features, multilingual sentence embeddings, and a few-shot LLM classifier through majority voting, while the multimodal variant additionally incorporates MFCC-based acoustic descriptors. The system outperformed the official baseline in both subtasks, ranking third in the text-only setting and fourth in the multimodal setting.

Two findings emerge from this participation. First, the choice of LLM backbone had a substantial impact on performance: the smaller gemma4:31b model consistently outperformed the considerably larger qwen3.5:397b model, suggesting that language alignment and model behaviour may be more important than parameter count alone for irony detection in spoken Spanish. Second, the multimodal extension did not improve over the text-only configuration. While competing systems benefited from incorporating audio information, the MFCC-based representation and feature-level fusion strategy used in our approach did not provide measurable gains.

These findings indicate two main directions for future work. The first is to improve decision calibration and class balance in order to increase recall while preserving competitive precision. The second is to explore alternative multimodal strategies, including richer prosodic representations and more effective fusion mechanisms, to better exploit the acoustic information available in spoken irony.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Opus 4.7) and Grammarly for language revision, grammar and spelling correction, and assistance in the preparation of figures.

References

[1] A. J. Barceló Milkova, Más allá de las palabras: Detección de ironía en voz y texto con inteligencia artificial, Trabajo final de máster, Universitat Politècnica de València, 2025. URL: https://riunet.upv.es/handle/10251/228652, repositorio institucional RiuNet.

[2] V. González Verdegais, Creación de un corpus multimodal anotado para la detección de la ironía y su uso en la inteligencia artificial, Trabajo final de máster, Universitat Politècnica de València, 2025. URL: https://riunet.upv.es/handle/10251/228533, repositorio institucional RiuNet.

[3] M. Malik, D. Tomás, P. Rosso, How challenging is multimodal irony detection?, in: E. Métais, F. Meziane, V. Sugumaran, W. Manning, S. Reiff-Marganiec (Eds.), Natural Language Processing and Information Systems, Springer Nature Switzerland, 2023, pp. 18–32.

[4] A. P. Kostenko, The concept of irony in the study of literature, Science and Education a New Dimension. Philology II (2014) 9–13.

[5] A. Reyes, P. Rosso, D. Buscaldi, From humor recognition to irony detection: The figurative language of social media, Data & Knowledge Engineering 74 (2012) 1–12. URL: https://doi.org/10.1016/j.datak.2012.02.005. doi:10.1016/j.datak.2012.02.005.

[6] 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 Workshop Proceedings, CEUR-WS.org, 2026.

[7] 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.

[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, 2026. To appear.

[9] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic BERT sentence embedding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2022, pp. 878–891. doi:10.18653/v1/2022.acl-long.62.

[10] M. Kusner, Y. Sun, N. Kolkin, K. Weinberger, From word embeddings to document distances, in: Proceedings of the 32nd International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research, PMLR, 2015, pp. 957–966.

[11] Q. Liu, M. J. Kusner, P. Blunsom, A survey on contextual embeddings, arXiv preprint arXiv:2003.07278 (2020). URL: https://arxiv.org/abs/2003.07278. doi:10.48550/arXiv.2003.07278.

[12] J. P. Gujjar, H. P. Kumar, Text similarity technique using BGE-M3, in: 2025 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE), IEEE, 2025, pp. 1–7.

[13] Z. Nussbaum, J. X. Morris, B. Duderstadt, A. Mulyar, Nomic embed: Training a reproducible long context text embedder, arXiv preprint arXiv:2402.01613 (2024). URL: https://arxiv.org/abs/2402.01613. doi:10.48550/arXiv.2402.01613.

[14] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models, ACM Transactions on Intelligent Systems and Technology 15 (2024) 1–45. doi:10.1145/3641289.

[15] A. K. Joshi, Natural language processing, Science 253 (1991) 1242–1249. doi:10.1126/science.253.5025.1242.

[16] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao, Large language models: A survey, arXiv preprint arXiv:2402.06196 (2024). URL: https://arxiv.org/abs/2402.06196. doi:10.48550/arXiv.2402.06196.

[17] A. Zubiaga, Natural language processing in the era of large language models, Frontiers in Artificial Intelligence 6 (2024) 1350306. doi:10.3389/frai.2023.1350306.

[18] Gemma Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Iqbal, et al., Gemma 3 technical report, arXiv preprint arXiv:2503.19786 (2025). doi:10.48550/arXiv.2503.19786.

[19] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, Z. Qiu, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025). URL: https://arxiv.org/abs/2505.09388. doi:10.48550/arXiv.2505.09388.

A. Online Resources

The results of the SpeechMATICS 2026 are available via:

- SpeechMATICS,
- SpeechMATICS 2026 Results
- SpeechMATICS 2026 Code