LACELL at SatiSPeech-IberLEF 2025: Multimodal Linguistic Features plus Embeddings for Satire Identification from YouTube

Ángela Almela[1,*,†], Pascual Cantos-Gómez[1,†], Daniel Granados-Meroño[1,†] and Gema Alcaraz-Mármol[2,†]

[1]Facultad de Letras, Universidad de Murcia, Campus de La Merced, 30001, Murcia (Spain)
[2]Facultad de Educación, Universidad de Castilla-La Mancha, 45004, Toledo (Spain)

Abstract

These working notes summarize the LACELL team's participation in the SatiSPeech 2025 shared task, which focuses on multimodal satire detection, integrating textual and acoustic features to better capture the subtleties of humorous and satirical communication. Its application in Spanish is particularly relevant given the linguistic and cultural diversity of Spanish-speaking communities, where satire often relies on both content and prosody. We participated in the multimodal task using an ensemble approach combining linguistic and prosodic features with sentence embeddings extracted from several fine-tuned LLMs, and achieved the 8th place with a macro F1 score of 81.4695%. This result demonstrates the potential of multimodal strategies for capturing complex communicative intentions such as satire.

Keywords

Linguistic Features, Sentence Embeddings, Multimodal Classification, Satire Classification, Natural Language Processing

1. Introduction

Satire detection is an increasingly important task in Natural Language Processing (NLP) and multimodal analysis, especially in contexts where irony, humor, and exaggeration play a central role in public discourse. Unlike emotion recognition, satire detection poses additional challenges due to its frequent reliance on implicit meaning, ambiguity, and culturally grounded references, making it difficult to identify using surface-level features alone [1]. These challenges are compounded when satire is conveyed through multiple modalities, such as tone of voice and textual cues.

The automatic detection of satire is a challenging task due to its reliance on figurative language, cultural references, and implicit meaning, which are often difficult to capture through surface-level features alone. Prior studies highlight the importance of cultural and contextual understanding in humor and satire perception [2], as well as the role of figurative language in social media discourse [3]. Recent approaches have explored multimodal strategies that combine textual, visual, or acoustic cues to improve satire recognition [4, 5], and even the application of large language models in low-resource scenarios.

The SatiSPeech 2025 shared task [6], organized within IberLEF 2025 [7], aims to advance the field of multimodal satire detection by providing a benchmark dataset in Spanish that combines transcribed

IberLEF 2025, September 2025, Zaragoza, Spain
*Corresponding author.
†These authors contributed equally.
angelalm@um.es (Á. Almela); pcantos@um.es (P. Cantos-Gómez); daniel.granadosm@um.es (D. Granados-Meroño); gema.alcaraz@uclm.es (G. Alcaraz-Mármol)
https://portalinvestigacion.um.es/investigadores/331758/detalle (Á. Almela);
https://portalinvestigacion.um.es/investigadores/330963/detalle (P. Cantos-Gómez);
https://portalinvestigacion.um.es/investigadores/332724/detalle (D. Granados-Meroño);
https://www.researchgate.net/profile/Gema-Alcaraz-Marmol (G. Alcaraz-Mármol)
0000-0002-1327-8410 (Á. Almela); 0000-0001-6329-2352 (P. Cantos-Gómez); 0000-0002-5305-1376 (D. Granados-Meroño); https://orcid.org/0000-0001-7703-3829 (G. Alcaraz-Mármol)

speech with audio recordings. This combination allows the exploration of prosodic and linguistic cues to better distinguish satirical from non-satirical content. Identifying satire in spoken language is particularly relevant in Spanish given its widespread use in culturally diverse regions, where satire often relies on local nuances and vocal delivery.

Multimodal satire detection faces similar challenges to those of Automatic Emotion Recognition from a similar shared task [8, 9], such as the integration of heterogeneous features and the scarcity of authentic, annotated datasets. While many previous works have focused on textual satire, incorporating prosodic features such as pitch, rhythm, or intonation offers new avenues for capturing the subtle communicative signals that characterize satirical speech.

In this edition of the task, our team participated exclusively in the multimodal track, designing a system based on ensemble learning that integrates linguistic features from UMUTextStats, sentence embeddings from various large language models (LLMs), and prosodic audio features. Our approach achieved the 8th position, with a macro F1-score of 81.4695%, showing the effectiveness of combining diverse modalities for tackling the complexity of satirical communication. Additionally, this result outperformed the provided baseline, which reached a macro F1-score of 79.9243%.

2. Dataset

According to the organizers, the SatiSPeech 2025 dataset consists of audio segments in Spanish extracted from a variety of YouTube videos containing either satirical or non-satirical content. The dataset was compiled with the aim of capturing the multimodal nature of satire, which often relies on vocal delivery, prosody, and contextual cues in addition to textual content. The satirical segments were collected from well-known Spanish-language satirical sources such as El Mundo Today and Polònia, while the non-satirical examples were taken from informative or neutral media.

Each audio segment is accompanied by its corresponding manual transcription (generated semi-automatically), and both modalities - text and audio - are provided for system development. The training set contains a total of 2,000 labeled segments, balanced between satirical and non-satirical classes. Additionally, a separate test set was provided for final evaluation.

In our experiments, we did not use the initial development set released prior to the competition. Instead, we created a custom development split by reserving 25% of the official training set for validation and hyperparameter tuning. Table 1 summarizes the dataset statistics. Although the dataset is roughly balanced in terms of class distribution, the diversity in speaker style, audio quality, and prosodic variation adds significant complexity to the classification task.

Table 1
SatiSPeech 2025 statistics

Emotion	Train	Val	Test	Total
no-satire	2534	634	-	3168
satire	2265	567	-	2832
Total	4799	1201	2000	8000

For the analysis of the dataset, we used the UMUTextStats tool [10] to extract linguistic features and compare their distribution across satirical and non-satirical texts (see Figure 1). We found that morphosyntactic features such as nouns, articles, pronouns, and adverbs are among the most informative categories. Satirical texts tend to contain a higher frequency of affix-related features (e.g., suffixes in nominative nouns), longer words, and a higher syllable-per-word ratio, possibly due to a more elaborate or playful language style. In contrast, non-satirical texts present a higher average sentence length and a more consistent use of punctuation symbols such as commas and periods. Interestingly, personal pronouns (especially first-person forms) and question marks appear more frequently in satirical documents, which may reflect a rhetorical strategy to engage the audience or simulate a conversational

tone. These observations suggest that satirical speech often relies on distinct structural and stylistic markers, making linguistic profiling a valuable component in multimodal satire detection.

3. System description

We evaluated UMUTextStats [10] for linguistic feature extraction tool. This tool is similar to LIWC, the de-facto standard for psychological and linguistic profiling. While LIWC2022 is the latest English version available [11], the last available version for Spanish dates back to 2007 [12], and lacks support for many contemporary linguistic constructs. In contrast, UMUTextStats is especifically designed for Spanish, capturing linguistic phenomena that LIWC does not address, such as grammatical gender, verb tense variations, and morphosyntactic patterns. The tool has been successfully applied in previous studies, including hate speech detection [13] and satire analysis [1], demonstrating its relevance for complex classification tasks in Spanish.

Prior to extracting the linguistic features, we applied a preprocessing pipeline to the transcriptions. This process involved cleaning the text by removing elements such as hyperlinks, hashtags, mentions, digits, and percentages, which were either eliminated or replaced by standardized tokens. In addition, we corrected expressive lengthenings and spelling errors using the ASPELL tool[1] to ensure consistency in lexical analysis. Importantly, we retained the original, unprocessed transcriptions for the extraction of features related to orthographic correctness and stylistic variation, as these aspects could be informative for satire detection.

Regarding sentence embeddings, we evaluated a diverse set of Spanish and multilingual Large Language Models (LLMs) to represent the transcribed speech content. Specifically, we employed eight Transformer-based models: BETO, BERTIN, MarIA, ALBETO, DistilBETO, mDeBERTa, Twhin, and XLM-R Twitter. These models cover a range of pretraining strategies, including formal corpora, literary texts, multilingual resources, and social media content. Sentence embeddings were extracted using the approach from [14].

Table 2 summarizes the training settings for each model. A closer look at the training configurations of the evaluated LLMs reveals that three models –MarIA, DistilBETO, and mDeBERTa– were trained without any warm-up steps, suggesting that in these cases the models adapted to the task without requiring a gradual increase in learning rate. This may indicate their robustness or compatibility with

[1] http://aspell.net/

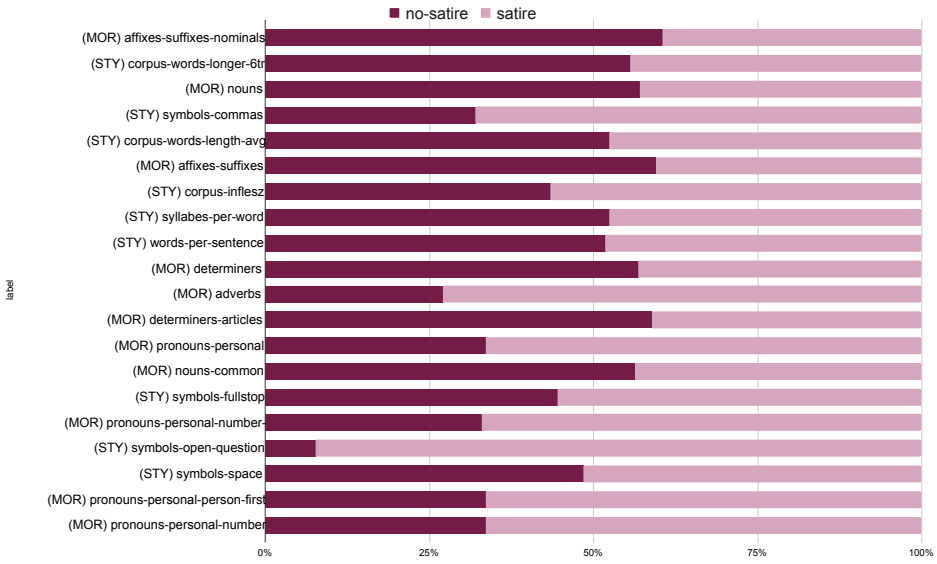

Figure 1: Information gain of the dataset with the stacked values organized by label

the goal of satire detection. Regarding the batch size, some models were tuned with a batch size of 8, while others achieved their better results with 16, reflecting different trade-offs in terms of memory consumption and training stability. In addition, most models were trained on a relatively high number of epochs (3 to 5), with 5 being the upper limit, suggesting that satire detection requires extended exposure to training data to capture nuanced patterns in both structure and meaning.

Table 2
Hyperparameters for fine-tuning the LLMs

LLM	lr	epochs	batch size	warmup steps	weight decay
ALBETO	1.7e-05	5	8	250	0.28
BERTIN	3.3e-05	5	16	1000	0.036
BETO (base)	3.7e-05	5	8	250	0.28
DistilBETO	3.1e-05	4	16	0	0.007
MarIA (base)	3e-05	3	8	0	0.087
mDeBERTa (base)	3.2e-05	4	8	0	0.16
Twhin	2.3e-05	4	8	1000	0.28
XLM-R Twitter	4.2e-05	4	16	1000	0.088

Among the different types of acoustic features available, we opted for the use of MFCC (Mel-Frequency Cepstral Coefficients), following the baseline system provided by the organizers. MFCCs are widely used in speech processing tasks because they approximate the way humans perceive sound by emphasizing perceptually relevant frequency bands. They capture essential aspects of speech prosody and articulation that are often indicative of expressive intentions, such as emphasis, exaggeration, or irony—elements frequently present in satirical speech. The MFCC vectors were extracted from each audio segment and used as input for a dedicated model in our ensemble.

Since all feature sets -—linguistic features (LFs), sentence embeddings, and acoustic features-— are represented as vectors, we explored various strategies to combine them and improve classification performance. In particular, we adopted an ensemble learning approach, where individual models were trained using each feature set independently. For the acoustic modality, we included MFCC features, following the strategy used in the official baseline system.

We then evaluated multiple fusion strategies to combine the predictions from the unimodal models. These included majority voting (mode), probability averaging, and max-confidence selection, where the predicted label corresponds to the category with the highest probability across models. This multimodal ensemble setup allowed us to leverage the complementary strengths of textual, linguistic, and prosodic cues in the satire detection task.

In order to adjust the LLMs for this task, we first fine-tuned the models with the training dataset using hyperparameter tuning. For each LLM, we evaluate 10 configurations that include variations on the learning rate, the warm-up steps, the weight decay, the number of epochs, and the batch size. Table 2 depicts the results for both models resulting in a larger number of epochs (4 for BETO, 5 for Maria) and little or no warm-up steps.

To integrate the outputs of the LLMs, MFCC features and the linguistic features extracted with UMUTextStats, we train a deep neural network on the concatenated feature vectors. After performing hyperparameter tuning, the resulting architecture consisted of a light funnel-shaped network with seven layers, starting from an initial layer with 90 neurons, and progressively reducing the dimensionality. The network used ELU (Exponential Linear Unit) as the activation function and did not include dropout, as regularization was not required during training. This simple architecture proved effective, likely because the input features—particularly the sentence embeddings—already captured rich semantic information, while the linguistic features complemented them with structural and stylistic cues relevant to satire.

First, we present the experiments conducted using the custom validation split in Table 4.1. The results are grouped into three main blocks: linguistic and acoustic features, sentence embeddings from various LLMs, and feature integration strategies, including ensemble learning (EL) and Knowledge Integration

(KI).

The first block shows the performance of models trained with individual feature sets. The linguistic features extracted with UMUTextStats achieved strong results on their own, with an F1-score of 89.30%, highlighting the relevance of surface-level and structural features such as part-of-speech distributions, punctuation, and stylistic cues for detecting satirical content. In contrast, the MFCC-based audio model obtained significantly lower results with an F1-score of only 48.86%. This suggests that, although prosodic cues are informative, they are less reliable when used in isolation, possibly due to variability in speaker style or recording conditions. This confirms that acoustic features alone are not sufficient for robust satire detection and need to be complemented with textual information.

In the second group, we evaluated several LLMs by extracting sentence embeddings from the transcribed speech. All models outperformed the linguistic and acoustic baselines. Among them, MarIA achieved the highest individual F1-score (94.90%), followed closely by Twhin and BERTIN, with 94.30% and 94.22%, respectively. These results suggest that domain-specific pretraining can provide embeddings well-suited for capturing the subtle semantics and discourse patterns characteristic of satirical texts. While models like DistilBETO and mDeBERTa performed slightly worse, they still exceeded 93% F1-score, confirming that even compact or multilingual architectures contribute valuable representations to the task.

The last block reports results from the integration strategies. As expected, combining different sources of information consistently improved performance. The weighted ensemble (EL WEIGHTED) achieved the highest scores across all metrics, with a macro F1-score of 95.82%, indicating that giving more weight to stronger models benefits the ensemble. Interestingly, the KI approach achieved the exact same performance, demonstrating that even a simple fusion architecture can compete with more complex ensemble methods. Other strategies such as mean probability and mode voting also performed well, but slightly below the best. These results highlight that multimodal integration, when carefully designed, leads to robust and high-performing satire detection systems.

4. Results

In this section, we report the results with our custom validation split (see Section 4.1), the official leader board (see Section 4.2, and an error analysis of the custom validation split (see Section 4.3).

4.1. Validation

Next, we present the detailed classification report of the Knowledge Integration (KI) strategy, where all features—sentence embeddings, MFCC audio features, and linguistic features—were combined into a single shallow neural network. Table 4 shows the results on the custom validation split, including precision, recall, and F1-score for both classes, as well as macro and weighted averages. The model achieved very balanced performance across classes, with macro and weighted F1-scores around 96%, indicating that it generalizes well to both satirical and non-satirical content. Notably, the F1-score for satire reached 95.73%, showing that the integrated feature representation effectively captures the nuanced characteristics of satirical speech.

4.2. Official results

Table 5 shows the official leaderboard for Task 2 (multimodal satire detection) of the SatiSPeech 2025 shared task. Our team, submitted under the name LACELL, ranked **8th out of 12** participating teams, achieving a macro 1-score of 81.4695%. This result outperformed the official baseline (79.9243%) and demonstrates the effectiveness of our multimodal integration approach combining linguistic, acoustic, and semantic features. The top-ranked system reached an F1-score of 88.3403%, highlighting the competitive level of the task.

As shown in Table 5, we achieved the **8th position** with a macro F1-score of 81.4695% using a Knowledge Integration strategy that combined linguistic, acoustic, and semantic features within a

unified neural network. This result outperformed the official baseline provided by the organizers (79.9243%) and placed us within a competitive range, only 6.8708 points below the top-ranked team, which achieved a macro F1-score of 88.3403%. These results highlight the challenge of multimodal satire detection and the potential of integrated architectures even in highly competitive settings.

Table 3
Results with the validation split

Strategy	precision	recall	f1-score
UMUTEXTSTATS	89.329	89.281	89.303
MFCC (AUDIO)	51.919	51.603	48.865
ALBETO	93.803	93.847	93.823
BERTIN	94.425	94.111	94.219
BETO-BASE	94.088	93.777	93.884
DILSTILBETO	93.699	93.345	93.463
MARIA-BASE	94.949	94.863	94.900
MDEBERTA-BASE	93.628	92.918	93.107
TWHIN	94.519	94.190	94.302
XLM-TWITTER	92.930	92.868	92.896
EL (HIGHEST)	77.861	71.362	68.487
EL (MEAN)	94.770	94.705	94.734
EL (MODE)	93.183	91.980	92.236
EL (WEIGHTED)	**95.875**	**95.777**	**95.820**
KI	**95.875**	**95.777**	**95.820**

Table 4
Classification report of the ensemble learning strategy based on the mode with the custom validation split.

	precision	recall	f1-score
no-satire	95.497	97.003	96.244
satire	96.589	94.885	95.730
macro avg	96.043	95.944	95.987
weighted avg	96.012	96.003	96.001

Table 5
Official leader-board for Task 2 (SatiSPeech 2025)

#	Team	MACRO F1-SCORE
1	edu_valero	88.340300
2	mcastro	86.444200
3	MarcoBortolotti	83.704100
4	nguyenminhbao5032	83.273900
5	ITST	83.271400
6	ngocan0987	82.776400
7	klagos1875	81.496100
8	**LACELL**	81.469500
9	AnGladun	80.130500
-	BASELINE	**79.924300**
10	cespinr	79.478700
11	deniscedeno	76.475800

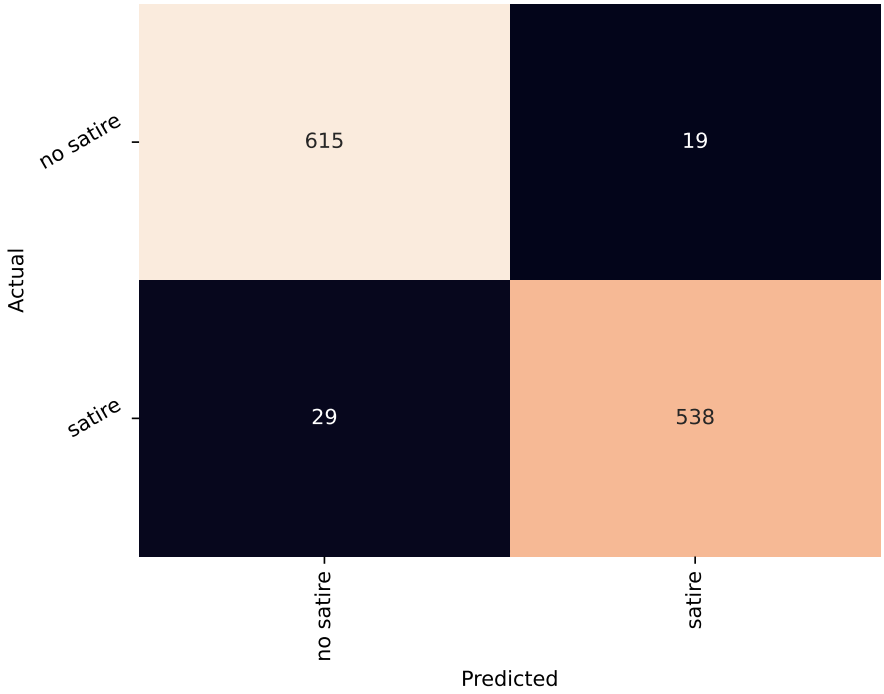

Figure 2: Confusion matrix of the ensemble model based on the mode

4.3. Error Analysis

To conduct the error analysis, we examined the confusion matrix of our best system, which used a Knowledge Integration strategy with the custom validation split (see Figure 2). The results show a strong performance across both classes, with most instances being correctly classified. However, we observed an asymmetry in the errors: the system misclassified 29 satirical segments as non-satirical, compared to only 19 non-satirical instances misclassified as satire. This suggests a slight bias toward the "no satire" class, possibly due to its more regular prosodic or linguistic patterns. Despite these misclassifications, the overall balance in predictions and the low error rates confirm the robustness of the integrated approach.

4.4. Conclusions and further work

In this working notes, we have described the participation of the LACELL team in Task 2 of the SatiSPeech 2025 competition, focused on multimodal satire detection in Spanish. Our system integrated sentence embeddings from several LLMs, linguistic features extracted with UMUTextStats, and acoustic features based on MFCCs, combined through a shallow neural network following a Knowledge Integration strategy. We achieved the 8th position in the official ranking with a macro F1-score of 81.4695%, outperforming the baseline by 1.5452 points. While our approach did not reach the top-tier performance, it demonstrated the viability of lightweight fusion strategies and confirmed the contribution of each modality individually.

As further work, we plan to explore the incorporation of pretrained acoustic language models such as Wav2Vec 2.0 or UniSpeech, which could enhance the representation of prosodic and phonetic cues in satire detection, as suggested in [15]. Additionally, we aim to investigate alternative fusion strategies between modalities, including early, late, and hybrid approaches, to more effectively exploit the complementary information of text and audio. We are also considering incorporating feature selection or attention-based mechanisms to identify the most relevant acoustic or lexical cues contributing to the satirical tone.

Acknowledgments

This work is part of the research projects LaTe4PoliticES (PID2022- 138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF)-a way to make Europe and LT-SWM (TED2021-131167B-I00) funded by MICIU/AEI/10.13039/ 501100011033 and by the European Union Next Generation EU/PRTR. This work is also part of the research project "Services based on language technologies for political microtargeting" (22252/PDC/23) funded by the Autonomous Community of the Region of Murcia through the Regional Support Program for the Transfer and Valorization of Knowledge and Scientific Entrepreneurship of the Seneca Foundation, Science and Technology Agency of the Region of Murcia.

Declaration on Generative AI

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

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.

[2] T. Jiang, H. Li, Y. Hou, Cultural Differences in Humor Perception, Usage, and Implications, Frontiers in psychology 10 (2019) 123.

[3] M. del Pilar Salas-Zárate, G. Alor-Hernández, J. L. Sánchez-Cervantes, M. A. Paredes-Valverde, J. L. García-Alcaraz, R. Valencia-García, Review of English Literature on Figurative Language applied to Social Networks, Knowledge and Information Systems 62 (2020) 2105–2137.

[4] L. Li, O. Levi, P. Hosseini, D. Broniatowski, A Multi-Modal Method for Satire Detection using Textual and Visual Cues, in: Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, 2020, pp. 33–38.

[5] 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.

[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] R. Pan, J. A. García-Díaz, M. Á. Rodríguez-García, F. García-Sanchez, R. Valencia-García, Overview of EmoSPeech 2024IberLEF: Multimodal Speech-text Emotion Recognition in Spanish, Procesamiento del Lenguaje Natural 73 (2024).

[9] 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 (2024) 103856.

[10] J. A. García-Díaz, P. J. Vivancos-Vicente, A. Almela, R. Valencia-García, Umutextstats: A linguistic feature extraction tool for Spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 6035–6044.

[11] R. L. Boyd, A. Ashokkumar, S. Seraj, J. W. Pennebaker, The development and psychometric properties of LIWC-22, Austin, TX: University of Texas at Austin (2022) 1–47.

[12] N. Ramírez-Esparza, J. W. Pennebaker, F. A. García, R. Suriá, La psicología del uso de las palabras: Un programa de computadora que analiza textos en español, Revista mexicana de psicología (2007) 85–99.

[13] J. A. García-Díaz, S. M. Jiménez-Zafra, M. A. García-Cumbreras, R. Valencia-García, Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers, Complex & Intelligent Systems 9 (2023) 2893–2914.
[14] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992.
[15] L. Pepino, P. Riera, L. Ferrer, Emotion Recognition from Speech Using wav2vec 2.0 Embeddings, Proc. Interspeech 2021 (2021) 3400–3404.