UKR at SatiSPeech-IberLEF 2025: Multimodal Satire Detection in Spanish with a BETO-Based Text Encoder and MFCC-Derived Audio Features

Anatoly Gladun[1], Julia Rogushina[2] and Rodrigo Martínez-Béjar[3,*]

[1]International Research and Training Center of Information Technologies and Systems of National Academy of Sciences of Ukraine and Ministry of Education and Science of Ukraine, 40, Acad. Glushkov Avenue, Kyiv, 03187, Ukraine
[2]Institute of Software Systems of National Academy of Sciences of Ukraine, 40, Acad. Glushkov Avenue, Kyiv, 03187, Ukraine
[3]Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain

Abstract

This paper presents the UKR team's participation in the SatiSpeech 2025 shared task, focused on the detection of satirical content in Spanish from both text-only and multimodal (text + audio) sources. We propose a supervised fine-tuning approach using the Spanish monolingual BERT model (BETO) for Task 1, and extend it with MFCC-based acoustic features in Task 2 to capture prosodic information. For classification, we adapt the final layers of the transformer model to integrate textual and audio inputs in a unified architecture. Our system ranked 6th in Task 1 and 9th in Task 2, with validation macro F1 scores of 0.9648 and 0.9699, respectively. The results demonstrate that while textual information carries most of the discriminative power, audio features offer complementary cues that slightly improve performance.

Keywords

Satire speech Recognition, Natural Language Processing, Transformers, BERT, MFCC

1. Introduction

Satire is a nuanced and context-dependent form of expression that presents significant challenges for traditional content classification systems. Unlike straightforward humor, it delivers critique indirectly through rhetorical strategies such as irony, parody, and exaggeration. Understanding satire often requires knowledge of cultural context and the speaker's intent, which makes automated detection more complex. The satirical meaning is influenced not only by the text itself but also by vocal features like tone, pitch, and rhythm that shape how the message is communicated and received. As a result, detecting satire in multimodal content requires models that can effectively combine both linguistic and auditory information [1].

Interest in automatic satire detection has increased, driven by its potential to counter misinformation, support content moderation, and enrich media analysis. On digital platforms, satire often closely resembles genuine news, which heightens the risk of confusion and the dissemination of false narratives. Therefore, developing reliable automated systems for satire detection is essential to promote accurate interpretation of content, particularly in multilingual and culturally diverse contexts.

Traditionally, satire detection has concentrated on textual analysis. Transformer-based language models trained on news and social media datasets have demonstrated strong performance in this area [2]. However, satire is also widespread in spoken formats such as television shows, podcasts, and video sketches, where vocal delivery is central to its expression. Despite this, benchmarks for multimodal satire classification remain scarce. Studies on sarcasm and irony detection have shown that models integrating text, audio, and visual modalities significantly outperform text-only counterparts [3]. This performance gap underscores the importance of moving beyond purely textual representations and

IberLEF 2025, September 2025, Zaragoza, Spain
*Corresponding author.
glanat@yahoo.com (A. Gladun); ladamandraka2010@gmail.com (J. Rogushina); rodrigo@um.es (R. Martínez-Béjar)
0000-0002-4133-8169 (A. Gladun); 0000-0001-7958-2557 (J. Rogushina); 0000-0002-9677-7396 (R. Martínez-Béjar)

adopting multimodal approaches—particularly those that incorporate audio—to better capture the complexities of satirical communication.

Detecting satire in spoken or audiovisual content necessitates multimodal modeling due to the inherent limitations of relying on text alone. Satirical intent is often conveyed not just by the words themselves, but by how they are spoken. Prosodic features such as intonation, pitch variation, stress, and rhythm play a crucial role in signaling irony, sarcasm, or exaggeration. These vocal cues can fundamentally alter the meaning of an utterance. For instance, a deadpan or sarcastic tone can reverse the literal interpretation of the words, which would likely be misunderstood by models that analyze text in isolation. By incorporating audio, models can capture these non-verbal signals, leading to a more accurate understanding of speaker intent and more reliable satire detection in real-world, multimodal contexts.

Moreover, recent research has demonstrated that audio features can enhance performance in various NLP tasks, such as emotion recognition, and across different domains [4, 5].

To address the existing gap in multimodal satire detection, the SatiSpeech shared task [6], part of IberLEF 2025 [7], introduces a new benchmark focused on Spanish-language satire. The task includes two subtasks: (1) satire classification using text only, and (2) multimodal satire classification using aligned text and audio segments.

We participated in both subtasks of the SatiSpeech challenge by adapting our prior approach from the EmoSpeech task at IberLEF 2024. For Subtask 1 (text-only classification), we fine-tuned BETO, a BERT-based model pre-trained on Spanish corpora [8]. We extracted the [CLS] token from the final hidden layer to produce fixed-size sentence embeddings, which were then used as input to a support vector machine (SVM) classifier. We optimized the SVM through grid search over kernel types, regularization parameters, and gamma values.

For Subtask 2 (multimodal classification), we extended the architecture to integrate audio features. We extracted Mel-frequency cepstral coefficients (MFCCs) from each audio sample using `librosa`, resulting in 1D acoustic embeddings. These were concatenated with BETO's contextual text embeddings to form joint multimodal vectors. A custom classification head was built to accommodate the expanded feature set, operating on top of a modified `Bert` model. Our implementation, structured as a subclass of `BertForSequenceClassification` class, combined pooled textual and acoustic representations before the final classification layer. The model was trained using binary cross-entropy loss for satire detection.

This paper is organized as follows: Section 3 provides an overview of the shared task and Section 2 a summary of related works; Section 4 describes our modeling approaches for both unimodal and multimodal configurations; Section 5 presents the experimental results and comparisons; and Section 6 concludes with insights and future directions.

2. Related Work

The automatic detection of figurative language—such as irony, sarcasm, and satire—has garnered increasing attention in natural language processing due to its relevance in combating misinformation and supporting content moderation. Early research focused primarily on text-based features, using lexical, syntactic, and semantic cues to classify satirical or sarcastic content. Traditional machine learning approaches applied handcrafted features (e.g., n-grams, sentiment polarity) with classifiers such as SVMs or random forests. However, these methods often struggled with generalization across domains due to the context-dependent nature of satire.

Recent advancements in transformer-based language models, such as BERT and its monolingual variants like BETO [8], have significantly improved performance in satire and humor detection. These models benefit from contextualized word embeddings and large-scale pretraining, allowing them to better capture subtle linguistic signals such as hyperbole, metaphor, or irony [2].

In the multimodal domain, several studies have explored the integration of speech and visual data for figurative language detection. In [3] showed that combining audio and visual features improves

sarcasm recognition compared to text-only approaches. Similar findings have been reported in emotion recognition tasks, where speech prosody—including pitch, energy, and rhythm—was found to provide complementary information to text. Mel-Frequency Cepstral Coefficients (MFCCs) [9], as used in our work, are widely adopted as a compact representation of speech signals due to their effectiveness in modeling prosodic features.

The SatiSpeech 2025 shared task [6] directly addresses this gap by introducing a standardized dataset and evaluation protocol for Spanish satire detection from both text and aligned audio. Our work builds upon these prior studies by evaluating the effectiveness of a simple early fusion strategy using BETO embeddings and MFCC features, and by highlighting current limitations and opportunities in multimodal satire classification.

3. Task Description

The SatiSpeech shared task, part of IberLEF 2025, focuses on the automatic detection of satirical content in Spanish using both text-only and multimodal (text + audio) inputs. This challenge stems from the subtle and context-dependent nature of satire and its growing relevance in areas such as media analysis, misinformation detection, and computational discourse understanding.

The task is organized into two subtasks:

- **Task 1: Text-Based Satire Detection.** Participants must develop systems that classify whether a given transcript is satirical, relying solely on textual features. These may include word choice, syntactic patterns, and rhetorical devices like irony and exaggeration.
- **Task 2: Multimodal Satire Detection.** This subtask extends the problem by incorporating audio. Participants receive aligned audio-transcript pairs and must combine linguistic and prosodic cues—such as rhythm, intonation, and stress—for binary satire classification.

3.1. Dataset

We participated in the task using the *SatirA* dataset, a curated collection of Spanish-language audio segments sourced from YouTube. The dataset features a diverse range of Spanish dialects and regional varieties to promote linguistic diversity and reduce potential regional bias. Audio segments were generated using automatic speaker diarization, filtering out clips longer than 25 seconds. Transcriptions were produced using Whisper. The annotation process followed a semi-supervised approach: initial automatic labels were refined by a team of three expert annotators to ensure high-quality ground truth.

The final dataset contains approximately 25 hours of labeled content and was split 80/20 into training and validation sets. All submissions and evaluations were carried out through the Codalab platform.

Table 1 summarizes the class distribution in both splits. There is a mild class imbalance, with non-satirical samples being slightly more frequent and typically longer on average. Satirical transcripts tend to show more variation in length. While these patterns may reflect real-world stylistic differences, models should avoid exploiting them directly during classification.

Set	Metric	Non-satirical	Satirical
Train	Samples	3168	2832
	Avg. Length	60.83	55.95
	Std. Dev.	12.01	17.60
Validation	Samples	633	567
	Avg. Length	61.08	55.83
	Std. Dev.	12.23	17.15

Table 1
Summary of dataset statistics by set and class.

4. Methodology

This section describes the approach we used for both subtasks of the SatiSpeech Shared Task: text-based satire detection (Task 1) and multimodal satire detection (Task 2). Instead of a traditional SVM-based pipeline, we fine-tuned neural architectures built on top of the Spanish BERT model (BETO), using custom classification heads tailored to each modality.

Our system was developed in PyTorch using the HuggingFace Transformers library. For Task 1, we relied exclusively on textual inputs. For Task 2, we extended this model by incorporating MFCC-based acoustic features extracted from the corresponding audio segment. The full architecture for the multimodal model is illustrated in Figure 1.

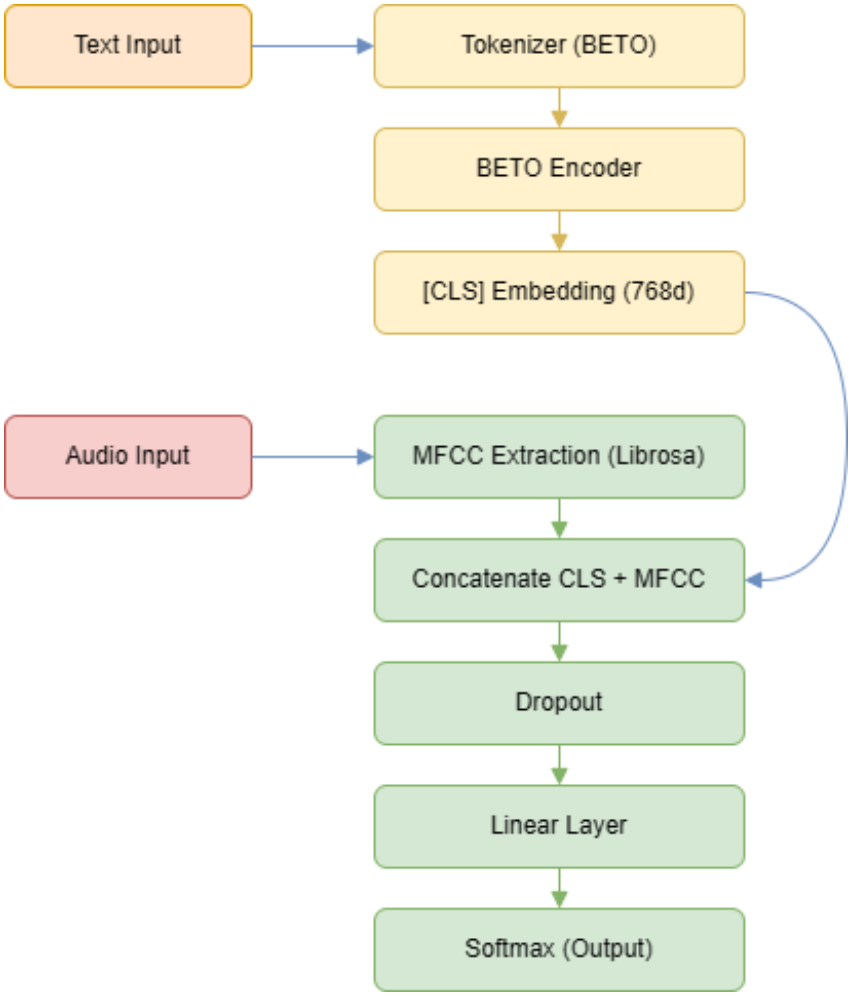

Figure 1: Multimodal classification architecture combining BETO embeddings with MFCC acoustic features.

4.1. Text-Based Satire Detection (Task 1)

For Task 1 (text-only satire detection), we fine-tuned the `dccuchile/bert-base-spanish-wwm-uncased` model, commonly known as BETO. The transcriptions were preprocessed using standard tokenization, with padding and truncation applied to ensure a consistent input length of 512 tokens. We extracted the 768-dimensional contextual embedding associated with the [CLS] token from the final hidden layer of BETO to serve as a fixed-size representation of each utterance, which was subsequently passed through a lightweight classification head for binary satire prediction.

This embedding was passed to a custom classification head consisting of a dropout layer followed by a linear projection to the number of output labels. The entire model, including BETO and the classification

head, was fine-tuned using cross-entropy loss.

Training was conducted over 10 epochs using a batch size of 16 and a learning rate of 2×10^{-5}. We selected the best model based on weighted F1-score on a validation set (10% of the training data).

4.2. Multimodal Satire Detection (Task 2)

For Task 2 (multimodal satire detection), we extended the textual model by integrating prosodic information. We extracted Mel-Frequency Cepstral Coefficients (MFCCs) from each audio segment using the librosa library. These features were averaged across the temporal axis to yield a fixed-length acoustic vector of 40 dimensions. This audio representation was then concatenated with the 768-dimensional BETO text embedding.

This MFCC vector was concatenated with the 768-dimensional BETO embedding to form a 808-dimensional multimodal input. We then adapted the classification head to project from this larger combined feature space.

The decision to combine BETO with MFCCs rather than using a full acoustic language model (e.g., Wav2Vec 2.0) was driven by both computational constraints and interpretability. MFCCs provide a compact and well-understood representation of prosodic and timbral characteristics of speech, which are highly relevant for capturing nuances of satirical delivery such as tone, rhythm, and emphasis.

The multimodal classification head was custom-designed to handle the concatenated embeddings efficiently. By treating the audio and text branches independently until the fusion point, the architecture maintains modularity, making it adaptable for other modalities or downstream tasks.

All training was conducted using PyTorch and Hugging Face Transformers, with evaluation based on weighted and macro F1-scores to account for mild class imbalance. The entire pipeline—from preprocessing to final prediction—is reproducible and compatible with GPU acceleration, ensuring scalability for larger datasets or multilingual adaptations.

The rest of the architecture remained the same. The model was trained for 20 epochs using a learning rate of 1×10^{-5} and the same batch size. As in Task 1, we used cross-entropy loss and selected the best model according to weighted F1-score on the validation split.

5. Results

For Task 1, the fine-tuned BETO model on Spanish satirical texts achieved a macro F1 score of **0.9648** on the validation split. This demonstrates that BETO, as a monolingual transformer pretrained on Spanish, is highly effective at capturing linguistic features relevant to satire, such as irony, hyperbole, and rhetorical structure.

For Task 2, the addition of prosodic features through MFCC vectors led to a modest improvement, with a macro F1 score of **0.9699**. The gain, while small, confirms that acoustic signals—particularly related to tone, rhythm, and emphasis—can reinforce textual cues in satire detection. However, the improvement is incremental rather than transformative, suggesting that simple early fusion of features (concatenation) may not fully exploit the expressive richness of the audio modality.

Table 2
Results of the BETO and BETO+MFCC models on the validation split for Task 1 and Task 2. The metrics include macro precision (M-P), macro recall (M-R), and macro F1-score (M-F1).

Model	M-P	M-R	M-F1
	Task 1		
BETO	0.9653	0.9645	**0.9648**
	Task 2		
BETO+MFCC	0.9701	0.9697	**0.9699**

Tables 3 and 4 summarize the official rankings for the SatiSpeech 2025 challenge. The UKR team ranked **6th** in Task 1 with an official macro F1 of **83.20**, and **9th** in Task 2 with an F1 of **80.13**. Despite using a relatively lightweight architecture and classical MFCC audio representations, the system outperformed several teams employing more complex multimodal strategies and remained competitive with top-ranking systems.

Table 3
Official leaderboard for Task 1

Task 1		
#	**Team Name**	**M-F1**
1	UPV-ELiRF	85.64
2	ITST	84.55
3	UMU-Ev	84.46
4	nguyenminhbao5032	83.27
5	Ferrara	83.21
6	**UKR**	**83.20**

Table 4
Official leaderboard for Task 2

Task 2		
#	**Team Name**	**M-F1**
1	UMU-Ev	88.34
2	UPV-ELiRF	86.44
3	Ferrara	83.70
4	nguyenminhbao5032	83.27
5	ITST	83.27
6	ngocan0987	82.78
7	UAE	81.50
8	LACELL	81.47
9	**UKR**	**80.13**

As the gold labels for the test set were not publicly released, a 20% stratified split from the training data was used as a held-out validation set to better understand model performance. Table 5 provides the macro-averaged metrics for both tasks.

Table 5
Macro-averaged classification metrics on the validation set

Task	Precision	Recall	F1-score
Task 1 (BETO)	0.9653	0.9645	0.9648
Task 2 (BETO+MFCC)	0.9701	0.9697	0.9699

A class-wise breakdown revealed that the models maintain extremely high performance across both `satire` and `no-satire` classes. In particular:

- **Task 1** showed excellent balance, with precision and recall values near 0.99 for both classes, indicating that BETO is capable of generalizing from lexical and syntactic satire cues alone.
- **Task 2** demonstrated that audio features marginally improved recall for satirical utterances. This suggests the model learned prosodic patterns such as exaggerated intonation or rhythmic anomalies common in satirical speech.

- However, the improvement was not drastic, which likely reflects the limitation of using MFCC features and early fusion. Richer acoustic representations or attention-based fusion may be needed to fully leverage the audio modality.

In sum, both models achieved results well above the challenge baseline, confirming the validity of our pipeline. The performance gap between our internal validation and leaderboard scores could be explained by slight domain shift, noise in test set audio, or differences in source domains.

5.1. Error Analysis

To better understand the limitations of our models, we conducted a qualitative analysis of several misclassified examples from the validation set. We selected representative cases from both Task 1 (text-only) and Task 2 (multimodal) that illustrate common failure patterns.

In Task 1, many errors occurred when the satirical content employed subtle rhetorical cues such as irony, sarcasm, or absurd exaggeration that were not easily distinguishable from factual reporting. These examples often relied on cultural or contextual background knowledge that the model was unlikely to capture from text alone.

In Task 2, despite the inclusion of prosodic features through MFCCs, some predictions still failed to recognize satirical tone. This may be due to the limitations of using static acoustic embeddings and early fusion strategies, which do not fully exploit temporal speech dynamics such as intonation or speech rate.

Table 6 presents a selection of misclassified samples with their full transcriptions and classification results.

Task	Prediction	Gold Label	Transcription
Task 1	no-satire	satire	Este parque es súper importante para el movimiento del barrio, que además ahora se usa para los rodajes, que es otra fuente de ingresos. Pero claro, como lo usa la tele, ya nos lo van a cerrar.
Task 1	no-satire	satire	El 5G no deja de sorprender. Ayer se pudo ver a una vecina del barrio que, tras vacunarse, sintonizaba los canales rusos con sólo tocarse la frente.
Task 1	no-satire	satire	No nos responden. Luego tuve ocasión de, con los servicios secretos franceses, hacerme con una grabación de Macron escuchando la COPE mientras se duchaba.
Task 2	no-satire	satire	Bueno, pues me temo que vamos a tener que dejar las fronteras abiertas otra vez. Parece que el virus ha pedido vacaciones.
Task 2	satire	no-satire	Por su altura de 1 metro y 52 centímetros, nunca pensaron que sería capaz de trepar esa valla de seguridad. Pero lo logró. Y con elegancia.

Table 6
Examples of misclassified validation samples from both tasks. Transcriptions illustrate the subtlety of satire that models failed to capture.

This qualitative review reveals that improving satire detection may require enhanced modeling of pragmatic and cultural cues, and more sophisticated multimodal fusion strategies that account for the temporal structure of speech. Incorporating prosody-aware embeddings, pitch contours, or pre-trained acoustic transformers such as Wav2Vec2 could help address these limitations in future work.

6. Conclusion

In this paper, we presented the UKR team's system for the SatiSpeech 2025 shared task, addressing both textual and multimodal satire detection in Spanish. Our approach relied on fine-tuning the BETO language model, with and without the integration of acoustic features derived from MFCCs. Despite

using a simple early fusion strategy for the multimodal task, our system achieved strong results, placing 6th in Task 1 and 9th in Task 2 on the official leaderboard.

The validation results showed that BETO alone is highly capable of capturing the linguistic and rhetorical markers of satire, with a macro F1 score of 0.9648. The addition of MFCC features led to a modest gain (macro F1 of 0.9699), suggesting that prosodic cues—although helpful—were not fully leveraged with the current feature representation. These findings underscore the need for more effective fusion strategies to harness the full potential of multimodal features.

Our system outperformed the baseline by a large margin, and remained competitive among teams that adopted more complex multimodal architectures. As future work, we plan to explore richer audio embeddings (e.g., Wav2Vec 2.0 fine-tuning), attention-based fusion layers, and the use of prosodic-specific features like pitch contours and speech rate, which may reveal subtler satire signals.

Overall, our results confirm the effectiveness of Transformer-based models in satire detection and highlight the potential of integrating speech features to enhance detection in multimodal communication settings. Additionally, we aim to experiment with LLMs, given their demonstrated effectiveness in various classification tasks within domains such as hate speech and satire detection [10, 11]. Their capacity for contextual understanding and generalization could further enhance performance in nuanced tasks like SatiSpeech.

Declaration on Generative AI

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

References

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

[2] 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.

[3] L. Li, O. Levi, P. Hosseini, D. Broniatowski, A multi-modal method for satire detection using textual and visual cues, in: G. Da San Martino, C. Brew, G. L. Ciampaglia, A. Feldman, C. Leberknight, P. Nakov (Eds.), Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, International Committee on Computational Linguistics (ICCL), Barcelona, Spain (Online), 2020, pp. 33–38. URL: https://aclanthology.org/2020.nlp4if-1.4/.

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

[5] 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.

[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] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[9] V. Tiwari, Mfcc and its applications in speaker recognition, International journal on emerging technologies 1 (2010) 19–22.

[10] A. Salmerón-Ríos, J. A. García-Díaz, R. Pan, R. Valencia-García, Fine grain emotion analysis in spanish using linguistic features and transformers, PeerJ Computer Science 10 (2024). doi:10.7717/PEERJ-CS.1992.

[11] R. Pan, J. A. García-Díaz, R. Valencia-García, Spanish mtlhatecorpus 2023: Multi-task learning for hate speech detection to identify speech type, target, target group and intensity, Computer Standards & Interfaces 94 (2025) 103990. URL: https://www.sciencedirect.com/science/article/pii/S0920548925000194. doi:https://doi.org/10.1016/j.csi.2025.103990.