UAE at SatiSpeech–IberLEF 2025: Multimodal Satire Detection Using BETO and Wav2Vec2 Embeddings with SVM

Katty Lagos-Ortiz[1,*], José Medina-Moreira[2] and Oscar Apolinario-Arzube[3]

[1]Facultad de Ciencias Agrarias, Universidad Agraria del Ecuador, Av. 25 de Julio, Guayaquil, Ecuador
[2]Universidad Bolivariana del Ecuador, Km 5.5 vía Durán-Yaguachi, Durán, Ecuador
[3]Universidad Internacional del Ecuador, Guayaquil, Av. Las Aguas y calle 15, Urbanor 2, Ecuador

Abstract

This paper presents our participation in the SatiSPeech shared task at IberLEF 2025, which focuses on detecting satirical content in Spanish using both textual and multimodal (text + audio) information. We address the two subtasks proposed: (1) satire detection based on transcribed text, and (2) multimodal satire detection integrating acoustic signals. Our approach is based on extracting deep embeddings from pretrained models—BETO for text and Wav2Vec 2.0 for audio—and using concatenated representations as input to a Support Vector Machine (SVM) classifier. Experimental results on the official test set yield macro F1-scores of 0.820 for the text-based task and 0.818 for the multimodal task. While the textual modality captures the most relevant features for satire detection, incorporating audio features using the current fusion approach does not improve classification performance.

Keywords

Satire Detection, Automatic Emotion Recognition, Natural Language Processing, Transformers, SVM

1. Introduction

Satire is a sophisticated and context-dependent form of expression that poses a unique challenge for automatic content classification systems. Unlike straightforward humor, satire often conveys implicit criticism through the use of irony, exaggeration, parody, and cultural references. These rhetorical strategies are deeply embedded in both linguistic structure and vocal delivery, requiring not only semantic understanding but also the interpretation of tone, intent, and social context. In multimodal scenarios where meaning is constructed jointly through text and speech, this complexity increases significantly, as auditory cues such as intonation, pitch, rhythm, and prosody interact with lexical and syntactic features to convey layers of meaning that are not always explicitly stated [1]. The difficulty of satire detection is further compounded by its cultural specificity, variability across speakers, and frequent use of double meanings or fictional scenarios.

Despite these challenges, satire detection has gained traction in recent years due to its relevance in a number of applied domains, including misinformation detection, media analysis, and digital content moderation. The proliferation of satirical content in online platforms, often indistinguishable from real news or commentary, raises important concerns about the misinterpretation of information and the spread of false narratives. Automatic systems capable of accurately identifying satire could play a crucial role in supporting content verification tools, improving the robustness of sentiment analysis pipelines, and enhancing the interpretability of user-generated content in multilingual and multicultural environments.

IberLEF 2025, September 2025, Zaragoza, Spain
*Corresponding author.
†These authors contributed equally.
klagos@uagraria.edu.ec (K. Lagos-Ortiz); jjmedinam@ube.edu.ec (J. Medina-Moreira); osapolinarioar@uide.edu.ec (O. Apolinario-Arzube)
0000-0002-2510-7416 (K. Lagos-Ortiz); 0000-0003-1728-1462 (J. Medina-Moreira); 0000-0003-4059-9516 (O. Apolinario-Arzube)

Most existing studies in satire detection have focused on textual data, often leveraging transformer-based language models trained on large corpora of news or social media posts [2]. However, satire frequently occurs in spoken formats—such as television monologues, radio commentaries, podcasts, and online video skits where vocal delivery contributes substantially to the intended meaning. The integration of text and audio for satire recognition remains underexplored, with few datasets capturing this multimodal dimension and limited benchmarks to evaluate models in realistic communicative settings. Notable exceptions include multimodal sarcasm and irony detection tasks, as well as studies like [3] and [4], which combine text with images or acoustic features, showing that fusion models generally outperform their unimodal counterparts.

The SatiSpeech shared task [5] at IberLEF 2025 [6] aims to bridge a key research gap by introducing a novel benchmark for multimodal satire detection in Spanish. The task comprises two subtasks: (1) satire classification based solely on text, and (2) multimodal classification that leverages aligned text and audio segments. The dataset was curated from a broad range of Spanish-language sources, including satirical shows such as El Intermedio, Zapeando, Homo-Zapping, and El Mundo Today, alongside conventional news outlets like Antena 3 Noticias, El Mundo, and BBC News. The collected material was carefully segmented and transcribed using Whisper [7], then annotated via a semi-supervised pipeline that combined expert validation with automatic filtering. This approach ensures a high-quality, culturally rich dataset that captures diverse communicative styles and dialectal variations.

In this case, we have reused the same approaches employed for EmoSpeech 2024 at IberLEF 2024 [5, 8] based on multimodal emotion detection. In this shared task, for the text-based satire classification (Task 1), we adopted a pipeline that leverages BETO, a monolingual BERT model pretrained for Spanish [9], to obtain contextual embeddings for each sentence. We use the [CLS] token representation from BETO's final hidden layer as a fixed-size embedding of the input text. These embeddings are then used as input features for a Support Vector Machine (SVM) classifier. To optimize performance, we conducted hyperparameter tuning via a grid search over a range of kernel types, regularization parameters, and gamma values.

In the multimodal satire classification task (Task 2), we extended this approach by integrating audio features extracted from Wav2Vec 2.0, specifically the Spanish-adapted variant "facebook/wav2vec2-large-xlsr-53-spanish"[10]. The audio files were processed using `librosa` to load each segment, which was then passed through Wav2Vec 2.0 to generate high-level acoustic embeddings. We combined these audio features with the textual embeddings from BETO, concatenating both vectors into a single multimodal feature vector for each sample. These multimodal representations were then fed into an SVM classifier, with the same hyperparameter optimization strategy used in the text-only setup. This fusion approach allowed the system to capture both semantic and prosodic aspects of satirical expression.

This paper is organized as follows: Section 2 and 3 provide an overview of the shared task and the related works; Section 4 describes in detail the methodology used for each subtask; Section 5 presents the experimental results and comparisons; and Section 6 concludes with insights and future directions.

2. Related works

Satire detection has been primarily studied in the context of textual data, leveraging machine learning and deep learning approaches. Early efforts utilized handcrafted linguistic and stylistic features, such as sentiment incongruence, lexical diversity, or syntactic complexity, to capture the distinctive characteristics of satirical writing. However, the emergence of transformer-based language models has significantly improved performance in satire and irony detection tasks.

For Spanish, BETO [9], a monolingual BERT model pretrained on large Spanish corpora, has been shown to perform competitively on a variety of downstream tasks, including emotion classification and hate speech detection [11, 12]. BETO has been widely adopted in IberLEF tasks, showing its robustness in low-resource and domain-specific contexts.

In the multimodal domain, recent works have explored combining textual information with other

modalities such as images [3] and audio [4]. Particularly in sarcasm and emotion detection, studies have shown that prosodic and acoustic cues (intonation, pitch, rhythm) can enhance model performance, especially when lexical ambiguity is high. One notable approach is the integration of Wav2Vec 2.0, a self-supervised speech representation model that captures high-level acoustic features without manual annotation.

Despite these advances, satire detection in multimodal settings remains underexplored, especially for Spanish. The SatiSpeech shared task at IberLEF 2025 [5] addresses this gap by introducing a benchmark that integrates speech and text, enabling more realistic modeling of satirical expression. While some recent efforts have explored sarcasm in podcasts or online videos, there is limited work that evaluates the combined role of semantic and prosodic features in satire recognition.

3. Task description

The SatiSPeech Shared Task, organized as part of IberLEF 2025, focuses on the automatic detection of satirical content in Spanish using textual and multimodal (text + audio) data. The task is motivated by the inherent complexity of satire as a communicative phenomenon and the increasing need for systems that can distinguish satirical from non-satirical content, particularly in media analysis, misinformation detection, and computational linguistics.

The shared task is divided into two subtasks:

- **Task 1: Text-Based Satire Detection.** This subtask requires participants to develop systems capable of identifying whether a given text transcription corresponds to satirical or non-satirical content. Participants must rely exclusively on textual cues such as lexical choices, sentence structure, rhetorical devices, and linguistic markers of irony or exaggeration.
- **Task 2: Multimodal Satire Detection.** This subtask extends the first by incorporating acoustic information from the original audio segment. The goal is to explore whether integrating vocal cue, such as prosody, intonation, and rhythm, alongside text improves satire detection performance. Participants are provided with audio-text pairs and must develop systems that leverage both modalities to perform binary classification.

3.1. Dataset

For the development and evaluation of the systems, we used the official dataset provided by the organizers. The corpus, named *SatirA*, comprises audio segments in Spanish extracted from YouTube videos. These include satirical programs such as *El Intermedio*, *Zapeando*, *Homo-Zapping*, and *El Mundo Today*, as well as non-satirical sources like *Antena 3 Noticias*, *El Mundo*, and *BBC News*. The dataset ensures broad linguistic and cultural diversity by incorporating content from various Spanish-speaking regions, thereby minimizing regional bias.

The video content was segmented automatically using diarization tools, discarding segments exceeding 25 seconds in length. Automatic transcriptions were generated using Whisper. A semi-supervised approach was adopted for annotation: initial automatic classifications were refined through manual validation by three expert annotators to ensure high labeling accuracy.

The final dataset contains approximately 25 hours of annotated recordings. For training and evaluation, the data were split into training and test sets with an 80/20 ratio. The evaluation was conducted through the Codalab platform, which managed the leaderboard and submission system.

Analysis of the training set reveals a slight class imbalance: approximately 52.8% of samples are labeled as non-satirical and 47.2% as satirical. Figure 1 illustrates the distribution of transcript lengths separated by class. Non-satirical transcriptions tend to be slightly longer and more concentrated around 55 words, while satirical samples exhibit higher variance. Similarly, non-satirical audio segments tend to be longer in duration. These structural differences may reflect underlying patterns in satirical versus informative discourse; however, models should be designed carefully to avoid overfitting to these surface-level characteristics.

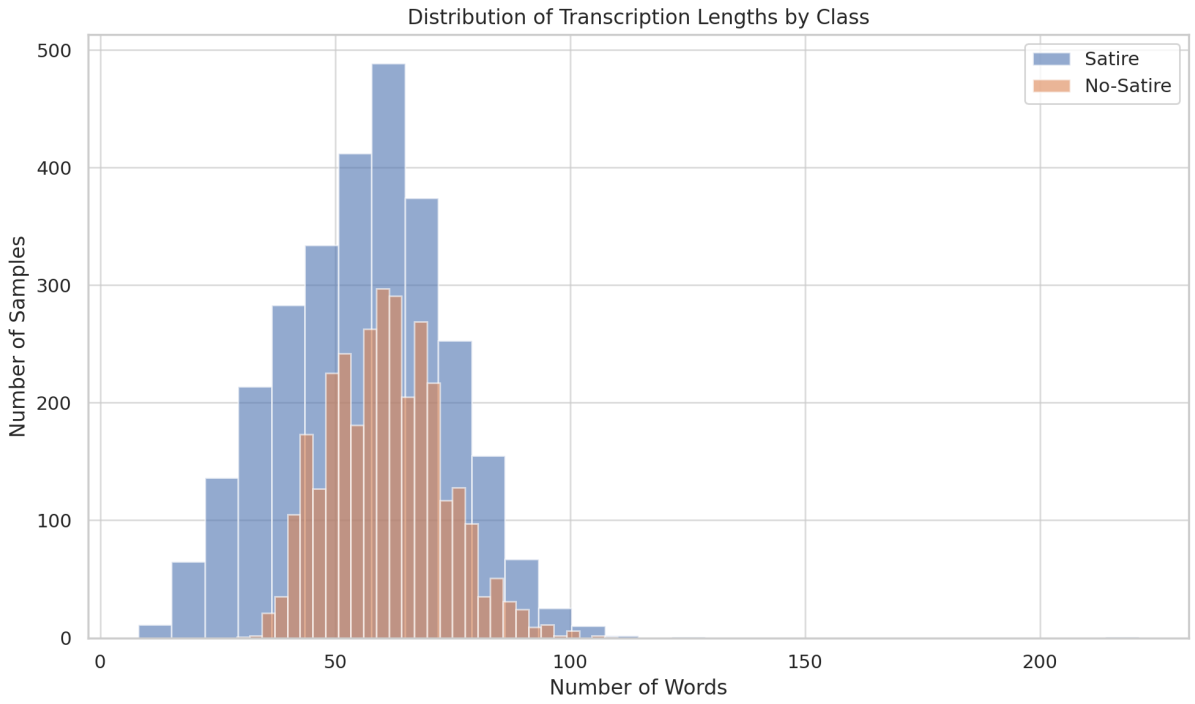


Figure 1: The distribution of transcript lengths separated by class.

4. Methodology

This section describes the methodology used to address both subtasks of the SatirA Shared Task: textual satire detection (Task 1) and multimodal satire detection (Task 2). For both tasks, we adopted a classical machine learning pipeline based on Support Vector Machines (SVM), using feature vectors extracted from state-of-the-art pretrained models.

Our approach to both subtasks of the SatirA Shared Task is structured around a modular pipeline that combines feature extraction from pretrained models and classical classification with Support Vector Machines (SVM). Figure 2 provides a visual summary of the process. The architecture includes independent branches for processing textual and acoustic information, followed by feature concatenation and classification.

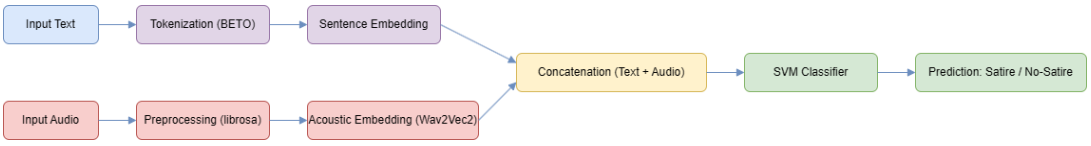


Figure 2: Multimodal Satire Detection Pipeline.

Briefly, the system operates in the following main steps:

- Preprocess each transcription using a tokenizer based on the BETO model.
- Extract sentence embeddings from the `[CLS]` token of BETO.
- Load and preprocess each audio file using `librosa` and extract acoustic embeddings with Wav2Vec 2.0.
- Concatenate the text and audio feature vectors to create a joint multimodal representation.
- Train an SVM classifier using these features and perform hyperparameter optimization via grid search.

4.1. Text-Based Satire Detection (Task 1)

For the text-only subtask, we utilized the Spanish version of BERT, known as BETO [9], using the model checkpoint `dccuchile/bert-base-spanish-wwm-uncased`. Each input transcription was tokenized using the associated tokenizer with truncation and padding to a maximum sequence length of 512 tokens. We extracted the hidden state corresponding to the `[CLS]` token from the last hidden layer as a fixed-size representation of the sentence.

These 768-dimensional vectors were used as features for training a Support Vector Machine (SVM) classifier. We performed a hyperparameter grid search using 5-fold cross-validation over a range of values for the regularization parameter C, the kernel coefficient γ, and two kernel types (`rbf` and `poly`). The best model was selected based on macro F1-score on a held-out validation set (10% of the training data).

4.2. Multimodal Satire Detection (Task 2)

In the multimodal setup, we extended the feature representation by including acoustic information from the corresponding audio segment. For this purpose, we used the Wav2Vec 2.0 model pre-trained for Spanish, specifically the `facebook/wav2vec2-large-xlsr-53-spanish`[1]. Each audio file was loaded using `librosa` and resampled to 16 kHz before feature extraction.

The feature extractor provided by HuggingFace was used to preprocess the audio, and we extracted the embedding corresponding to the first frame (i.e., the vector at position `[0,0,:]`) from the final hidden layer. This vector was then concatenated with the textual embedding from BETO to form a single multimodal feature vector of size 1536 (768 from text + 768 from audio).

As with the text-only system, these vectors were used to train an SVM classifier. The same grid search procedure was applied to tune hyperparameters and select the best performing model.

5. Results

Table 1 presents the official rankings for Task 1 (text-based satire detection) and Task 2 (multimodal satire detection), respectively, based on the macro F1 score.

In Task 1, our team **UAE** achieved a macro F1 score of **81.63**, securing the **8th position** among twelve participating teams. Our system surpassed the official baseline (79.37) by over 2 points, demonstrating the effectiveness of our textual pipeline, which was based on fine-tuning transformer-based models for satire detection in Spanish.

In Task 2, which incorporated both textual and audio information, **UAE** ranked **7th**, with a macro F1 score of **81.50**. Again, our system outperformed the multimodal baseline (79.92), validating our strategy of combining text embeddings with audio features such as MFCCs, despite using a relatively simple fusion mechanism.

These results reflect the robustness of our approach across both tasks, but contrary to expectations, the addition of audio features slightly decreased the macro F1-score. This suggests that either prosodic features are not informative enough for satire, or that the simple feature concatenation approach is insufficient. While validation scores approached 99%, the drop to 82% on the test set indicates overfitting. This discrepancy may result from dataset shift or excessive confidence on the training data.

5.1. Analysis

The models were evaluated using a 10% validation split from the training set (`random_state=11`), consisting of 600 samples. On this split, the text-based classifier (Task 1) achieved a macro F1-score of 0.9950, while the multimodal system (Task 2) obtained a similar macro F1-score of 0.9933. Overall accuracy exceeded 99% in both tasks.

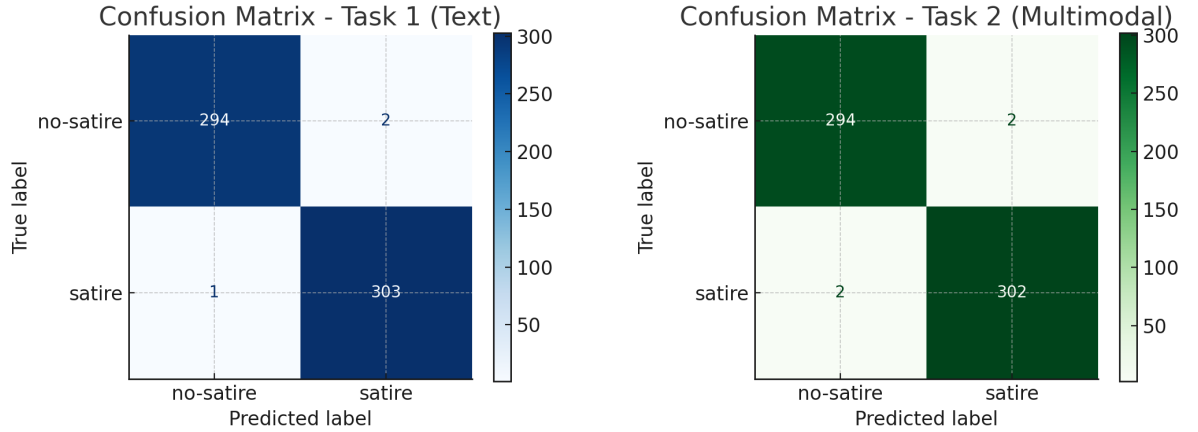
[1] https://huggingface.co/facebook/wav2vec2-large-xlsr-53-spanish

Team	F1 (Task 1)	Rank	F1 (Task 2)	Rank
UPV-ELiRF	85.63	1	86.44	2
ITST	84.54	2	83.27	5
UMU-Ev	84.45	3	**88.34**	1
nguyenminhbao5032	83.27	4	83.27	4
Ferrara	83.21	5	83.70	3
UKR	83.20	6	80.13	9
ngocan0987	83.20	7	82.78	6
UAE	**81.63**	**8**	**81.50**	**7**
LACELL	81.46	9	81.47	8
EcuPLN	79.48	10	79.48	10
Baseline	**79.37**	–	**79.92**	–
UTP	77.39	11	76.48	11

Table 1
Combined leaderboard: Macro F1 scores and rankings for Task 1 (Text-based) and Task 2 (Multimodal).

For Task 1, the *no-satire* class showed very high precision (99.66%) and strong recall (99.32%). The *satire* class achieved similarly robust metrics, with a recall of 99.67% and an F1-score of 99.51%. Task 2 followed the same trend, though with slightly lower values, indicating that the addition of audio features did not significantly enhance model performance within this split.

Figures 3a and 3b display the confusion matrices derived from the validation predictions. Both matrices reflect highly balanced classifications, with minimal misclassifications and nearly symmetric performance across the two classes.



(a) Task 1 – Text-Based

(b) Task 2 – Multimodal

Figure 3: Confusion matrices on the 10% validation split (600 samples).

These results confirm the high capacity of pretrained language models such as BETO for satire classification in Spanish. The slight performance drop in the multimodal setting suggests that our current fusion approach (simple feature concatenation) may not fully leverage the prosodic or acoustic cues captured by Wav2Vec 2.0.

Common misclassifications included neutral news reports labeled as satire, and ironic satire misclassified as non-satirical. For instance, a segment beginning with "Claramente, este gobierno lo ha hecho todo perfecto..." was misclassified due to subtle sarcasm not clearly marked prosodically. Despite including prosodic information, the multimodal system underperformed compared to text-only. This may point to limitations in acoustic modeling or redundancy in the information conveyed via both modalities.

6. Conclusion

In this paper, we presented our system for the SatirA Shared Task at IberLEF 2025, which aims to detect satirical content in Spanish through text-only and multimodal (text + audio) inputs. Our approach was based on combining deep embeddings from pretrained models BETO for text and Wav2Vec 2.0 for audio with a classical SVM classifier trained on top of the fused representations.

Results on the official test set show that both models achieved competitive performance, with macro F1-scores of 0.820 (text) and 0.818 (multimodal). The small difference in performance suggests that textual information alone provides a strong signal for satire detection, and that our current fusion method may not fully exploit the additional prosodic cues offered by the audio modality.

Our analysis indicates that the system achieves high recall for the *satire* class and high precision for the *no-satire* class, which reflects a desirable tradeoff for many real-world applications. Qualitative analysis of misclassified samples suggests that stylistic ambiguity—particularly irony expressed subtly—was a primary challenge. However, this requires further empirical study.

As future work, we plan to explore more advanced fusion strategies such as attention-based cross-modal transformers or hierarchical models to better capture the nuanced interactions between speech and language in satirical expression. In addition, incorporating linguistic features such as syntactic complexity or rhetorical devices may further improve interpretability and accuracy in satire detection.

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] G. Wick-Pedro, C. F. da Silva, M. L. Inácio, O. A. Vale, H. de Medeiros Caseli, Using large language models for identifying satirical news in Brazilian Portuguese, in: P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, R. Amaro (Eds.), Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1, Association for Computational Linguistics, Santiago de Compostela, Galicia/Spain, 2024, pp. 156–167. URL: https://aclanthology.org/2024.propor-1.16/.

[5] 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).

[6] 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.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato,

J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 28492–28518. URL: https://proceedings.mlr.press/v202/radford23a.html.

[8] 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.

[9] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: Proceedings of the Practical ML for Developing Countries Workshop at ICLR 2020 (PML4DC), 2020. URL: https://arxiv.org/abs/2002.12191.

[10] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL: https://arxiv.org/abs/2006.11477. arXiv:2006.11477.

[11] J. A. García-Díaz, F. García-Sánchez, R. Valencia-García, Smart analysis of economics sentiment in spanish based on linguistic features and transformers, IEEE Access 11 (2023) 14211–14224.

[12] R. Pan, J. A. García-Díaz, F. Garcia-Sanchez, R. Valencia-García, Evaluation of transformer models for financial targeted sentiment analysis in spanish, PeerJ Computer Science 9 (2023) e1377.