

UNED-UNIOVI at EmoSpeech-IberLEF2024: Emotion Identification in Spanish by Combining Multimodal Textual Analysis and Machine Learning Methods

Juan Martinez-Romo^{1,2,*}, José Farnesio Huesca Barril^{3,4}, Lourdes Araujo^{1,2} and Enrique de La Cal Marin³

¹NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), Juan del Rosal 16, Madrid 28040, Spain

²IMIENS: Instituto Mixto de Investigación, Escuela Nacional de Sanidad, Monforte de Lemos 5, Madrid 28019, Spain

³Department of Computer Science, University of Oviedo, 33003 Oviedo, Spain

⁴Faculty of Science and Technology Athabasca University, Athabasca, Canada

Abstract

This paper describes our participation in the first edition of the EmoSpeech Task (Multimodal Speech-text Emotion Recognition in Spanish) of IberLEF 2024 shared evaluation campaign, devoted to the multimodal emotion recognition in Spanish comments from YouTube channels. For the text-based component, we utilized a series of language models specifically trained on Spanish corpora to identify emotional expressions within textual data. In addressing the audio component of the task, we employed various machine learning algorithms tailored to process and analyze audio segments for emotion detection. Our best strategy consisted of integrating the outputs from both text and audio models using a voting technique. This ensemble method allowed us to harness the strengths of each individual model, thereby enhancing the robustness and accuracy of our emotion recognition system. The results of this approach proved to be promising, indicating not only the viability of combining multiple models but also highlighting the potential improvements in emotion recognition tasks through multimodal methodologies.

Keywords

Emotion Recognition, Shared Task, Large Language Model, Multimodal, Spanish

1. Introduction

Emotion recognition, a pivotal facet of human-computer interaction, has traditionally focused on analyzing singular modalities – typically, either text or audio. However, human emotion is inherently complex, communicated not just through words but through tone, tempo, and timbre of speech. Integrating multiple modalities, therefore, presents a promising avenue to enhance the accuracy and robustness of emotion recognition systems. This paper delves into the relevance and potential of multimodal emotion recognition, specifically, the integration of text and audio data as a result of our participation in the 2024 EmoSpeech shared task [1] at IberLEF 2024 [2].

Emotion recognition, an essential component of natural language processing (NLP), aims to discern human emotions from various forms of communication, enabling machines to respond to human needs more empathically. Among the diverse approaches used for emotion recognition, large language models (LLMs) have emerged as particularly potent tools, especially for processing textual data. Among the most recent approaches is the use of pre-trained language models such as Bidirectional Encoder Representations from Transformers (BERT)[3], a language model based on the transformer architecture that has revolutionized natural language processing since its introduction in 2018. In contrast to previous language models, BERT is trained on a "gap-filling" or "masking" task, where words are randomly hidden in a sentence and the model must predict them. What makes BERT especially powerful is its ability to


IberLEF 2024, September 2024, Valladolid, Spain

*Corresponding author.

†These authors contributed equally.

✉ juaner@lsi.uned.es (J. Martinez-Romo); UO285319@uniovi.es (J. F. H. Barril); lurdes@lsi.uned.es (L. Araujo); delacal@uniovi.es (E. d. L. C. Marin)

ORCID 0000-0002-6905-7051 (J. Martinez-Romo); 0000-0002-7657-4794 (L. Araujo); 0000-0001-7142-7544 (E. d. L. C. Marin)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

capture the context and relationships between words. It uses a transformer architecture that applies multiple layers of attention to process both the information before and after a given word. This allows BERT to capture the meaning and the meaning and dependency of words in a broader context. BERT is trained on large amounts of unsupervised text and learns deep, contextualized linguistic representations. Once trained, it can be used in a variety of natural language processing tasks, such as text classification, information extraction, question answering, sentiment analysis, and many more. In addition, it can be adapted to specific tasks through a process called "fine-tuning" in which the model parameters are tuned on an annotated data set.

The rapid advancement of artificial intelligence (AI) has propelled LLMs to the forefront of research and application in NLP. Their ability to process human-like text has opened new avenues for automated emotion recognition systems. By analyzing text through the lens of these sophisticated models, researchers are able to capture a broad spectrum of emotional nuances, which are often embedded subtly in language usage, syntax, and semantics. This capability not only enhances interaction quality in applications such as chatbots and virtual assistants but also provides deeper insights into the emotional states conveyed through written communication.

Recent advancements in machine learning, particularly deep learning, have substantially improved the performance of models that process complex data types independently. Yet, these models often falter in scenarios where understanding context and subtleties across different forms of expression is crucial. Multimodal emotion recognition aims to bridge this gap by leveraging the complementary strengths of textual and auditory signals. By synthesizing the semantic precision of text with the expressive richness of audio, such approaches promise not only enhanced performance but also greater generalizability across diverse real-world environments.

2. Related Work

2.1. Emotion Analysis

Emotion detection in text is a natural language processing task that has been addressed from different approaches, ranging from the use of sentiment-related terminologies [4] to machine learning and deep learning [5]. One recent approach is using pre-trained language models such as BERT for emotion detection in texts [6]. RoBERTa [7] is an extension of the BERT model, which applies some modifications to various aspects of training in order to improve the performance obtained by BERT for some specific tasks, that has also been applied to emotion detection [8]. There are versions of RoBERTa trained for Spanish and tuned on specific emotion data. The XLM-roBERTa-base model has been trained on approximately 198M tweets and tuned for Spanish emotion analysis. This model participated in the EmoEvalEs competition, part of the IberLEF 2021 Conference, where the proposed task was the classification of Spanish tweets among seven different classes: anger, disgust, fear, joy, sadness, surprise, and others. It achieved first place in the competition with a macro-averaged F1 score of 71.70%.

2.2. Speech Emotion Recognition

Concerning the Speech Emotion Recognition (SER) topic, the key issue in the SER architecture design is the acoustic feature input selection. The main classification of acoustic features is the classical and modern approaches, i.e., handcrafted features vs. deep learning-based features. Classical handcrafted features employed acoustic features extracted per frame. These features are often called local features or low-level descriptors (LLDs). Besides, statistical features computed from LLDs are new ways to capture the dynamics among frames[9]. In addition, a few preset sets of handcrafted features have been published in last years in different Speech Challenges, like ComparE2016 containing 6013 features, and Geneva Minimalistic Acoustic Parameter Set (GeMAPS) with 88 features [10], which can be computed with the OpenSmile library.

On the other hand, it is reasonable to extract an acoustic representation of speech in an end-to-end manner via deep learning methods. In INTER-SPEECH 2020 ComParE challenge, two deep learning-

based features were given in the baseline system, DeepSpectrum and AuDeep. The provided DeepSpectrum features with ResNet50 network achieved the highest unweighted average recall (UAR) on the elderly emotion sub-challenge test set. Although there is a movement to use DNN-based feature extraction, the majority of SER research still relies heavily on handcrafted acoustic features[11] for interpretability and energy-saving reasons.

3. EmoSpeech Task

The objective of this shared task is to delve into the domain of Affective Emotion Recognition (AER). This task is structured to tackle the complexities inherent in this classification domain. A primary issue in AER is identifying which attributes are crucial for differentiating between distinct emotions. Additionally, the development of these recognition systems is often hindered by a shortage of multimodal datasets that depict real-life emotional scenarios, as the bulk of datasets currently available are derived from contrived settings that do not accurately capture true emotional expressions. The necessity to integrate multiple features further complicates the classification challenge, escalating the difficulty in designing sophisticated architectures and incorporating a diverse array of features. Consequently, pinpointing the most defining features for each emotion type becomes significantly more complex.

This shared task tackles the challenges associated with AER through two distinct approaches:

1. **AER from Text:** This aspect concentrates on the extraction and identification of key features, pinpointing the most indicative feature of each emotion within a dataset derived from genuine real-life circumstances. The objective of this task is to evaluate textual data to discern the emotions expressed within. It focuses on five of Ekman’s six fundamental emotions—anger, disgust, fear, joy, and sadness plus an additional category for neutral emotions. The effectiveness of the systems is assessed using Precision, Recall, and the F1-score. Rankings will be determined based on the macro F1-score for emotion classification.
2. **Multimodal AER:** This approach necessitates the development of a more intricate architecture to address the classification intricacies. The unique aspect of this task is its focus on a multimodal methodology, evaluating how language models perform with datasets that mirror real-world scenarios. This task extends AER to include an additional modality—audio. It aims to analyze both text and audio cues to accurately classify the emotions expressed in each sample, again covering the same set of Ekman’s emotions: anger, disgust, fear, joy, sadness, and neutral. Systems are evaluated and ranked based on their Precision, Recall, and F1-score, with a focus on the macro F1-score for emotion classification. Participants have the option to engage in either or both tasks independently.

3.1. Dataset

The dataset [12] for this shared task was curated by gathering audio clips from various Spanish-language YouTube channels. The starting hypothesis is that specific topics trigger distinct emotions in speakers as they share their views. For instance, the organizers’ analysis revealed that politicians on political channels often exhibited disgust when discussing the opposition. Likewise, a prevalent emotion of anger was noted in sports interviews, particularly from players voicing their frustrations following a defeat. The process of annotation was conducted in two stages. Initially, they identified and downloaded videos from YouTube channels that reliably provoked specific emotions, from which they then extracted the audio segments. Subsequently, these segments were manually labeled by three researchers with the emotions of disgust, anger, joy, sadness, and fear, excluding surprise due to its absence in the content analyzed. A category for neutral emotion was also included.

The dataset used for this shared task is only a part of the Spanish MEACorpus 2023, featuring over 13.16 hours of annotated audio covering five of Ekman’s six basic emotions. These audios are proportionally divided into training and test sets, adhering to an 80%-20% split, respectively. For this

EmoSpeech2024 challenge, there were 3000 utterances distributed across six categorical emotions as in figure 1.

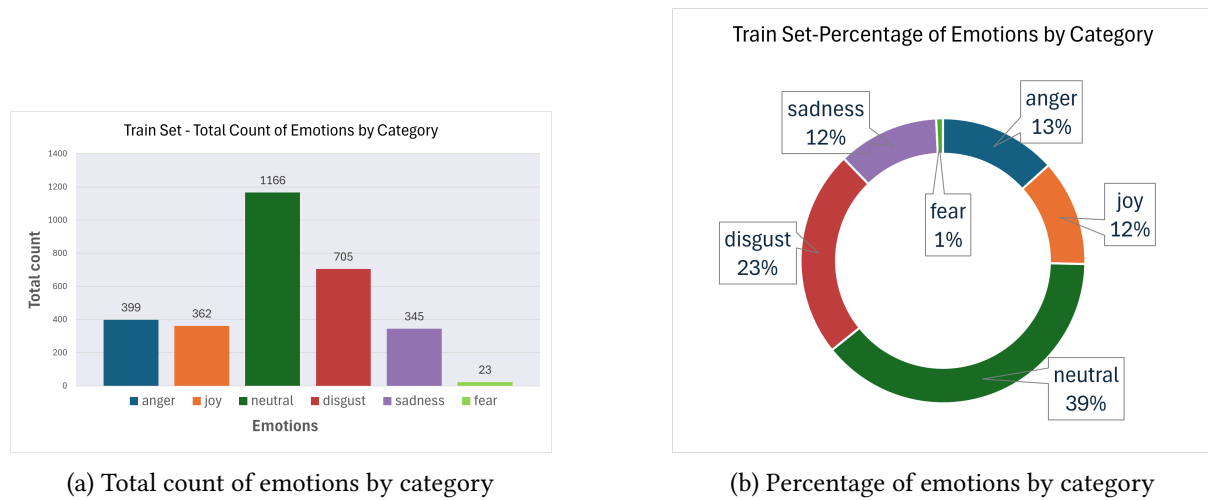


Figure 1: Statistics of the Evaluation Phase of the EmoSpeech2024 dataset

4. Proposed System

4.1. UNED System (Text)

The text-based portion of our model employs two advanced language models. These models were specifically trained on a dataset of messages extracted from the social network X (formerly Twitter). The choice of this dataset was strategic; the brevity and directness of messages on this platform make it an ideal source for training models to capture nuanced emotional cues in textual content. Both models underwent a fine-tuning process with the data provided by the organizers for training.

- Model A: The model was developed using the TASS 2020 corpus, which consists of approximately 5,000 tweets spanning various Spanish dialects. The base model is RoBERTuito [13], a variant of the RoBERTa model that has been specifically trained on Spanish-language tweets.
- Model B: This model is based on the XLM-RoBERTa-base architecture [14] and has been trained on approximately 198 million tweets from the EmoEval 2021 shared task. It has been further fine-tuned specifically for emotion analysis in the Spanish language.

4.2. UNIOVI System (Audio)

Akay et al. [9] have provided a great summary of SER acoustic features from the literature, mainly these four categories: prosodic, spectral, Teager Energy Operator, and Voice Quality. In our SER proposal, because of interpretability and simplicity reasons, a reduced handcrafted acoustic features set has been selected and compared against the well-known ComparE2016[15] preset set of features for EmoSpeech2024 dataset.

Thus, our SER proposal is composed of the following steps: i) audio preprocessing, ii) features engineering, iii) features normalization, iv) ML training and v) Feature sets comparison.

Audio preprocessing: All the audio files have been converted to mono-channel and resampled to 44,1kHz sampling rate, and emphasized the signal applying the following high-pass filter, with α coefficient 0.97:

$$y(t) = x(t) - \alpha * x(t - 1) \quad (1)$$

Features engineering We have decided to select a reduced set of 64 acoustic (see table 1), prosodic and spectral features compared to the bigger preset features like ComparE2016[15] or GeMAPS:

Table 1

List of proposed features and number of coefficients per each

| Features | mfcc | mfcc delta | mfcc delta2 | Spectral Contrast | Centroid | Rolloff | ZRC | RMS |
|----------|------|------------|-------------|-------------------|----------|---------|-----|-----|
| Total | 13 | 13 | 13 | 12 | 1 | 1 | 1 | 1 |

The features have been computed on the whole audio file obtaining 64 features per file, and normalized deploying a $[0, 1]$ scaling.

Training models: Inspired by the work of Pratama [16] on Support Vector Classifier (SVC) for speech emotion recognition, for the audio-based portion of our model we have opted to leverage simple and more interpretable machine learning models that can easily be replicated using hyper-parameters tuning through Grid Search for the audio-based portion of our model deployment. Thus, five shallow machine learning algorithms have been selected: Support Vector Classifier (SVC), Random Forest (RF), K-Nearest Neighbors (KNN), Decision Tree (DT) and CatBoost. The algorithms have been validated using a 10KFold cross validation strategy, and a fine-tuning of the hyperparameters of each algorithm with at least 3^3 combinations has been run for each fold.

Sets of features comparison After training and optimizing the five selected ML techniques (see table 3 with our set of features, the winner (SVC), has been compared against the best model trained with the ComparE2016 set of features (see2). It can be stated that the precision of our set of features (just 64 features) outperformed the ComparE2016 (6336 features) set of features precision in three out of the six classes.

Table 2

Precision per class/emotion in test set.

| Emotion | Anger | Disgust | Fear | Joy | Neutral | Sadness |
|-------------|-------|---------|-------|------|---------|---------|
| Ours | 0,63 | 0,67 | 0,50 | 0,67 | 0,81 | 0,75 |
| ComparE2016 | 0,26 | 0,73 | 1,00 | 0,66 | 0,91 | 0,61 |
| Diff | 59% | -9% | -100% | 1% | -12% | 19% |

Table 3

Task 1. Audio Analysis SER. (*) Not considered

| Ranking | Model | Macro F1 |
|----------|-------------------------|-------------|
| 1 | SVC GridSearchVC | 73.3 |
| 2 | Random Forest | 70.80 |
| 3 | SVC | 68.00 |
| 4 | KNN | 67.18 |
| * | CatBoost | 56.00 |
| * | DT | 52.00 |

4.3. Multimodal System

Our multimodal emotion recognition system was designed to integrate and leverage both textual and auditory data to enhance the accuracy of emotion detection. The system architecture employed a strategic ensemble approach, using a hard voting system to combine outputs from multiple specialized models. This method aimed to capitalize on the unique strengths and insights provided by each individual model.

The multimodal system incorporated five core models: two text-based models and three audio-based models. Each of these models was chosen based on its performance metrics and relevance to our task objectives.

Voting systems in machine learning are ensemble methods used to improve the predictive performance of models by combining the decisions from multiple models. This approach leverages the diversity among a set of algorithms to create a more robust and accurate ensemble model. The fundamental principle behind voting systems is that by pooling the outputs of various models, one can capitalize on the strengths and compensate for the weaknesses of individual models, thus achieving better performance than any single model alone.

Empirical studies and practical applications in the literature have consistently demonstrated the efficacy of voting systems in enhancing model performance across various domains and problems. These systems are particularly effective in scenarios involving high variability or complexity, where single models might struggle with generalization. For instance, in fields such as bioinformatics, speech recognition, and emotion recognition, ensemble methods like voting have shown to significantly reduce error rates and improve the stability of predictions.

Voting systems are not only beneficial for achieving higher accuracy but also for increasing the fault tolerance of the application. By distributing the risk among multiple models, the ensemble is less likely to fail catastrophically if one model makes a poor prediction. This is crucial in applications where reliability is as important as accuracy, such as medical diagnostics or financial forecasting.

5. Results

In this section we will discuss the main results obtained by the proposed system on the test dataset of the EmoSPeech task, as well as a comparative among the different systems participating in the task.

Our system integrated two distinct language models for task 1, Model A and Model B, both tailored for emotion recognition from text. In the competition, Model B outperformed Model A in several key metrics, including precision, and F1-score. Table 4 shows the ranking of the systems participating in task 1, where the third place of our team can be observed.

Table 4

Task 1. Monomodal SER. (*) This team submit their results a few hours past the limit

| Ranking | Team | Macro F1 |
|---------|--------------------|-----------------|
| 1 | TEC_TEZUITLAN | 67.18560 |
| 2 | CogniCIC | 65.75270 |
| 3 | UNED-UNIOVI | 65.52870 |
| 4 | UKR | 64.84170 |
| 5 | N/A | 61.75050 |
| 6 | THAU-UPM | 58.31430 |
| 7 | LACELL | 52.88210 |
| 8 | SINAI | 52.00010 |
| 9 | UAE | 51.82420 |
| - | Baseline | 49.68290 |
| 10 | UTP | 41.02270 |
| 11 | N/A | 37.85170 |
| 12 | N/A | 33.45860 |
| * | CICIPN | 54.9929 |

In the case of task 2, the results of which are shown in Table 5, the best performing system, being the fifth best system, was a hard voting system integrating the two text-based emotion recognition models presented in section 4.1, and three audio-based models presented in section 4.2. As in the case of task 1 where text-based model B performed better than model A, in the case of audio the best performing system was the SVC-based model. As for the audio models, the best-performing systems in descending

order were SVC GridSearchVC , RF, SVC. For this reason, they were selected for the voting system. In the case of the different systems, SVC GridSearchVC performed much better individually than the other models. SVC and KNN performed similarly and less than RF.

Table 5

Task 2. Multimodal SER. (*) This team submit their results a few hours past the limit

| Ranking | Team | Macro F1 |
|---------|--------------------|-----------------|
| 1 | BSC-UPC | 86.68920 |
| 2 | THAU-UPM | 82.48330 |
| 3 | CogniCIC | 71.22590 |
| 4 | TEC_TEZUITLAN | 68.75820 |
| 5 | UNED-UNIOVI | 67.09290 |
| 6 | UKR | 57.79690 |
| 7 | UAE | 55.88980 |
| - | Baseline | 53.07570 |
| 8 | UTP | 48.15590 |
| 9 | N/A | 9.41740 |
| * | CICIPN | 54.8168 |

6. Conclusions and Future Work

This paper describes our participation in the EmoSpeech task of the IberLEF 2024 shared evaluation campaign, devoted to the multimodal emotion recognition in Spanish comments from Youtube channels. The main contribution of our work consists in the combination of different text and audio models to detect emotions accurately. Our participation in the multimodal emotion recognition task has culminated in significant insights and outcomes. For the text component of our task, we trained models using X messages, achieving optimal performance. This success underscores the effectiveness of our natural language processing techniques in identifying and analyzing emotional expressions from text, which ultimately led to our third-place finish in the text-based emotion recognition category of the competition. In terms of audio analysis, we experimented with several classification algorithms, identifying the Support Vector Classifier (SVC) as the most effective. This choice proved crucial in optimizing our ability to discern emotions from audio data accurately, reflecting the importance of selecting appropriate methodologies for processing distinct data types. The overall results of the competition were enlightening; our team secured a fifth-place ranking in the multimodal emotion recognition task. Although not at the top of the leaderboard, these results demonstrate strong performance and provide a solid foundation for future enhancements. These results highlight the potential of multimodal approaches in emotion recognition tasks. Looking forward, there are clear opportunities to improve the integration of our text and audio models, perhaps by exploring more sophisticated fusion techniques or algorithms that can more effectively handle inconsistencies between different types of data.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the DOTT-HEALTH Project (MCI/AEI/FEDER, UE) under Grant PID2019-106942RB-C32, OBSER-MENH Project (MCIN/AEI/10.13039/501100011033 and NextGenerationEU^{PRTR}) under Grant TED2021-130398B-C21, and EDHER-MED under grant PID2022-136522OB-C21 as well as project SICAMESP (UNED, 2023-VICE-0029).

References

- [1] R. Pan, J. A. García-Díaz, M. Á. Rodríguez-García, F. García-Sánchez, R. Valencia-García, Overview of EmoSpeech 2024 at IberLEF: Multimodal Speech-text Emotion Recognition in Spanish, *Procesamiento del Lenguaje Natural* 73 (2024).
- [2] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024)*, co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.
- [3] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of naacL-HLT*, volume 1, 2019, p. 2.
- [4] S. Mohammad, P. Turney, Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon, in: *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, 2010, pp. 26–34.
- [5] S. Zad, M. Heidari, J. H. J. Jones, O. Uzuner, Emotion detection of textual data: An interdisciplinary survey, in: *2021 IEEE World AI IoT Congress (AIIoT)*, 2021, pp. 0255–0261.
- [6] F. A. Acheampong, H. Nunoo-Mensah, W. Chen, Transformer models for text-based emotion detection: a review of bert-based approaches, *Artificial Intelligence Review* 54 (2021) 5789–5829.
- [7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).
- [8] R. Kamath, A. Ghoshal, S. Eswaran, P. Honnavalli, An enhanced context-based emotion detection model using roberta, in: *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 2022, pp. 1–6.
- [9] M. B. Akçay, K. Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, *Speech Communication* 116 (2020) 56–76. doi:10.1016/j.specom.2019.12.001.
- [10] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, K. P. Truong, The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing, *IEEE Transactions on Affective Computing* 7 (2016) 190–202.
- [11] B. T. Atmaja, A. Sasou, M. Akagi, Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion, *Speech Communication* 140 (2022) 11–28. URL: <https://doi.org/10.1016/j.specom.2022.03.002>. doi:10.1016/j.specom.2022.03.002.
- [12] 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.
- [13] J. M. Pérez, D. A. Furman, L. A. Alemany, F. Luque, Robertuito: a pre-trained language model for social media text in spanish, *arXiv preprint arXiv:2111.09453* (2021).
- [14] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, *arXiv preprint arXiv:1911.02116* (2019).
- [15] The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity native language, *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 08-12-September-2016* (2016) 2001–2005. doi:10.21437/Interspeech.2016-129.
- [16] A. Pratama, S. W. Sihwi, Speech emotion recognition model using support vector machine through mfcc audio feature, 2022.