Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 359-368 recibido 04-07-2024 revisado 19-07-2024 aceptado 26-07-2024

Overview of EmoSPeech at IberLEF 2024: Multimodal Speech-text Emotion Recognition in Spanish

Resumen de EmoSPeech en IberLEF 2024: Reconocimiento de Emociones Multimodal Basado en Texto y Audio en Español

Ronghao Pan,[1] **José Antonio García-Díaz,**[1] **Miguel Ángel Rondríguez-García,**[2] **Francisco García-Sánchez**,[1] **Rafael Valencia-García**[1]
[1]Facultad de Informática, Universidad de Murcia
[2]Departamento de Ciencias de la Computación, Universidad Rey Juan Carlos
{ronghao.pan, joseantonio.garcia8, frgarcia, valencia}@um.es
miguel.rodriguez@urjc.es

Abstract: This paper presents the EmoSPeech 2024 shared task, which was organized in the IberLEF 2024 workshop within the framework of the 40th International Conference of the Spanish Society for Natural Language Processing (SEPLN 2024). The objective of this shared task is to study the field of Automatic Emotion Recognition (AER), which is becoming increasingly important due to its impact on various fields, such as healthcare, psychology, social sciences, and marketing. Specifically, two tasks are proposed and evaluated separately. The first task deals with AER from text, which focusing on feature extraction and identifying the most representative feature of each emotion in a dataset created from real-life situations. The second task deals with AER from a multimodal perspective, which requires the construction of a more complex architecture to solve this classification problem. The ranking includes the results of 13 different teams, each of which proposed a novel approach to the problem.
Keywords: Emotion Analysis, Multimodality, Audio Speech Recognition.

Resumen: Este artículo resume la tarea EmoSPeech 2024, organizada en el taller IberLEF 2024, dentro del marco de la 40ª Conferencia Internacional de la Sociedad Española de Procesamiento del Lenguaje Natural (SEPLN 2024). El objetivo de esta tarea es investigar el campo del Reconocimiento Automático de Emociones, que está adquiriendo cada vez más importancia debido a su impacto en diversos campos, como la sanidad, la psicología, las ciencias sociales y el marketing. En concreto, se proponen dos subtareas que se evalúan por separado. La primera subtarea se refiere al análisis de emociones a partir de texto, que se centra en la extracción de características y la identificación de las más representativas características de cada emoción en un conjunto de datos creado a partir de situaciones de la vida real. La segunda subtarea se centra en el análisis de emociones desde una perspectiva multimodal, lo que requiere la construcción de una arquitectura más compleja para resolver este problema de clasificación. La clasificación incluye los resultados de 13 equipos diferentes, cada uno de los cuales propuso un enfoque novedoso del problema.
Palabras clave: Análisis de Emociones, Multimodalidad, Reconocimiento de Audio.

1 Introduction

The ability to recognize human emotions is of paramount importance for establishing of positive interpersonal relationships, whether in person or through interactions with computers (Varghese, Cherian, and Kizhakkethottam, 2015). The field of Automatic Emotion Recognition (AER) has been a significant challenge for many years, and its importance is growing due to its impact on various fields such as healthcare, psychology, social sciences, and marketing (Chenchah and Lachiri, 2016; Lugovic, Dunder, and Horvat, 2016). In the context of software, AER has the potential to facilitate the

ISSN 1135-5948 DOI 10.26342/2024-73-27

delivery of more tailored responses and recommendations, which could ultimately increase user engagement and satisfaction. The process of AER can be addressed using several taxonomies. The most popular is the recognition of six basic emotional expressions: anger, disgust, fear, happiness, sadness, and surprise (Ekman, 1992). By automatically recognizing emotions, a system can identify, interpret, and respond to emotions expressed by humans in various communication channels, including text, facial expressions, voice tones, and even body language. The identification of emotions can be achieved through the use of various features, including linguistic features, voice intonation, intensity, duration, and even the conversational rhythm (Fahad et al., 2021). These features can be integrated into a multimodal paradigm that combines different types of data, such as text and speech, to achieve enhanced performance.

The EmoSPeech 2024 shared task is organized within the IberLEF 2024 (Chiruzzo, Jiménez-Zafra, and Rangel, 2024). The aim of this shared task is to explore the field of AER. The task has been designed to address the challenges associated with this classification problem. One of the major challenges in this field is to determine which features are sufficiently discriminative to distinguish between different emotions. Another limitation that affects these recognition systems is the lack of multimodal datasets with real-life scenarios. This is due to the fact that a significant portion of the available datasets have been obtained from prepared situations that do not accurately reflect real emotional expressions. Finally, the combined use of multiple features makes this classification problem more complex, thereby increasing the challenge of designing and building more advanced architectures and incorporating a greater variety of features. Together, these factors make it even more challenging to identify the most characteristic features associated with each type of emotion. To address these challenges, two subtasks are proposed. The first subtask is concerned with the extraction of AER from text, which involves the identification of the most representative feature of each emotion within a dataset derived from real-life scenarios. The second subtask deals with AER from a multimodal perspective. This requires the construction of a more complex architecture to solve the classification problem.

AER has recently attracted considerable attention from the research community. A number of shared tasks have been organized to address this challenge, including (Mohammad and Bravo-Marquez, 2017), EmoRecCom (Nguyen et al., 2021), and EmoEvalES (Plaza del Arco et al., 2021). These events have highlighted the growing interest in this area. The specificity of this task is its multimodal approach to AER, which involves analyzing of the performance of language models on real-world datasets. To the best of our knowledge, there is no existing shared task that addresses this challenge.

The rest of this paper is organized as follows. Section 2 provides a detailed description of the two tasks involved. Next, the dataset assembled for this task is delineated in Section 3. The methodologies used by by the participants to address the two tasks presented are outlined in Section 4. The official leaderboard is presented and discussed in Section 5. Finally, Section 6 provides a summary of the task and includes some further work.

2 Task description

In contrast to previous related challenge tasks, such as EmoEvalES@IberLEF2021 (Plaza del Arco et al., 2021), which focused solely on emotion recognition in tweets written in Spanish, this shared task is intended to explore solutions to the challenge of emotion recognition in a multimodal environment. The challenges inherent to this shared task are as follows:

1. Text AER. Identifying emotion in an audio transcript can be challenging due to the limited length of the text and the linguistic characteristics of spoken language.

2. Multimodal AER. Emotion analysis using multiple modalities presents a number of challenges, including heterogeneity (i.e., text and audio convey emotional information in different ways), the need for a fusion strategy, and increased computational demands.

The ultimate goal of this task is to investigate multimodal emotion recognition from

text and audio. To achieve this goal, the following subtasks have been defined

- **Task 1: AER from text**. The goal of this task is to analyze a given text and identify the emotion it conveys based on five of Ekman's six basic emotions: anger, disgust, fear, joy, and sadness, as well as one neutral emotion. The goal of this task is to extract features and identify the most representative feature of each emotion in a dataset created from real-life situations.
- **Task 2: Multimodal AER**. This task introduces a new modality to the AER challenge, namely audio. The goal is to integrate text and speech cues to determine the emotion conveyed by each given fragment, based on five of Ekman's six basic emotions: anger, disgust, fear, joy, and sadness, as well as a neutral emotion. To address the multimodal classification problem, it is necessary to construct a more sophisticated architectural framework.

The contest was organized by CodaLab and can be accessed through the following link: `https://codalab.lisn.upsaclay.fr/competitions/17647`. It was divided into three stages: Practice, Evaluation and Post-Evaluation.

In the practice phase, participants were provided with a subset of the training data in order to familiarize them with the data format. In addition, participants were provided with a notebook containing a baseline for both tasks based on Support Vector Machines and acoustic features, which served as a basis for system development. Participants were then provided with the comprehensive training set to facilitate the development of their approaches. Each participant was allowed a total of 100 submissions to CodaLab. In the evaluation phase, the test partition was made available to the participants to evaluate the developed systems. This partition was used to evaluate the teams. A total of 10 submissions could be made through CodaLab and each team had to select the most optimal one for the ranking. The ranking was determined using the macro F1-Score in both tasks, thereby allowing teams to participate independently in one or both tasks.

3 Dataset

The dataset for this task was compiled by collecting audio segments from a variety of Spanish YouTube channels. The hypothesis is that certain topics evoke certain emotions in their authors when they express their opinions. For example, our findings showed that politicians in political channels showed a sense of disgust towards the opposing party. Similarly, we observed that anger was a prevalent emotion in interviews with athletes expressing their frustration after losing a game in a sports context. A systematic analysis was conducted to identify and select a number of channels for each emotional category. This corpus is an extension of the one presented in (Pan et al., 2024).

Emotion	Train	Test	Total
anger	399	100	499
disgust	705	177	882
fear	23	6	29
joy	362	90	452
neutral	1166	291	1457
sadness	345	86	431
total	3000	750	3750

Table 1: Corpus statistics per emotion.

The annotation process was carried out in two phases. First, the YouTube channels that elicited the desired emotions were selected, the videos were downloaded, and the audio segments were extracted. Second, three members of the research group manually annotated of each sample, categorizing the emotions as disgust, anger, joy, sadness, and fear (note that surprise was not included in this analysis, because it not could be extracted from the YouTube data). A neutral emotion was also included. As a result, the Spanish MEACorpus 2023 was assembled, comprising over 13.16 hours of audio from audio segments annotated with five of Ekman's six emotions. As can be seen in Table 1, a total of 3,750 audio segments were selected for this contest, with 80% used for training and 20% for testing. Table 2 provides examples of the dataset, including the ID, which serves as a unique identifier for each sample; the transcription, which is the text contained in the sample; and the label, which classifies the emotion.

id	label	transcription
55ea70ba-9ce28715	anger	"¿Me puede decir usted, señor Marlaska, dónde lleva metido estos tres días que hay disturbios en la calle? Le exigí al señor Torra que condenara la violencia de manera firme, rotunda, sin adjetivos ni medias tintas. Palabras que yo comparto al cien por cien con usted, señor Marlaska."
6b1a5914-1d9ed49b	fear	"Cerró la puerta ahí, no me dejó entrar. Es horrible."
fa83510c-46074ae9	joy	"por aquí. Y deciros, mi querido hermano del alma, tesoro, te quiero, que no te preocupes, que yo lo comparto contigo y, Juanita, si no te lo dan ya tenemos otro en casa, así que lo compartamos."

Table 2: Examples of the dataset.

4 Participant approaches

A total of 38 users have registered for the EmoSPeech 2024 shared task. Of these, 13 teams have submitted results, and a total of 11 have submitted working notes describing their systems. 12 of the 13 teams participated in Task 1 (AER from Text), while a total of 9 teams out of the 13 submitted their results in Task 2 (Multimodal AER).

Next, we present a brief summary of the participants' systems:

- **adri28**. This team placed the 11th position in Task 1 and the 9th position in Task 2. This team did not submit the working notes describing their proposed approach.
- **BSC–UPC** (Casals-Salvador et al., 2024). This team placed 1st in Task 2. To achieve this, the participants propose a multimodal emotion classifier that combines a pre-trained model based on text transformers, namely RoBERTa (Liu et al., 2019), with a pre-trained model based on speech transformers, namely XLSR-wav2vec 2.0 (Conneau et al., 2021). The embeddings returned by each model are concatenated, and the dimension of the resulting vector is reduced by attention pooling. A classifier module is then used to process the above vector with the goal of selecting the predicted class. The proposed system ranked first position in this challenge with a Macro F1-score of 86.69%.
- **CogniCIC** (Soto et al., 2024). This team placed 2nd position in Task 1 and 3rd in Task 2. The participants presented two different approaches, one for each task. In Task 1, the authors use BETO (Cañete et al., 2023), fine-tuning it with the training instances provided. In Task 2, BERT is used to extract of textual features, while Mel-Frequency Cepstral Coefficients (MFCCs) (Zheng, Zhang, and Song, 2001) are used to represent audio features. The audio and text features are then integrated to generate the predictions.
- **Iris5**. This team placed 12th position in Task 1. This team did not submit the working notes describing their proposed approach.
- **ITST** (Paredes-Valverde and Salas-Zárate, 2024). This team placed 1st position in Task 1 and 4th position in Task 2. They proposed two different pipelines, one for each subtask. For Task 1, which involves identifying individual emotions from text in real-life situations, a five-step pipeline was defined. The proposed pipeline was primarily based on the following pre-trained Spanish language models: BETO, BERTIN (de la Rosa et al., 2022), and MarIA (Villegas, 2023). A different strategy was proposed for Task 2, which is a multimodal classification problem. This task was approached using an ensemble learning method that integrated the most effective fine-tuned text model, BETO, with another fine-tuned model derived from Wav2Vec 2.0. Two methods were used to combine the results: Mean and Max.
- **LACELL** (Almela et al., 2024). This team placed 7th position in Task 1. In this context, the participants evaluate a

feature integration scheme based on ensemble learning that combines textual features derived from the Linguistic Inquiry and Word Count (LIWC) (Boyd et al., 2022) with sentence embeddings extracted from two pre-trained large Spanish language models, namely, MarIA and BETO. To predict the final label, three probability combinations between classes are evaluated: mode, average, and highest.

- **SINAI** (García-Baena, García-Cumbreras, and Jiménez-Zafra, 2024). This team placed 8th in Task 1. For this task, the participants evaluated a number of the most popular Transformer-based models and specific open-source transformers for emotion analysis that are publicly available on Hugging Face. In total, the authors used 14 systems, including the baseline provided by the organizers.
- **THAU–UPM** (Esteban-Romero et al., 2024). This team placed 6th position in Task 1 and 2nd in Task 2. To achieve this, the participants used several Multimodal Large Language Models (MM-LLMs) (Zhang et al., 2024) for the tasks. The authors addressed the text-only task by finetuning the multilingual Gemma-2B model using the Low-Rank Adaptation (LoRA) method (Hu et al., 2022), which optimizes low-rank adaptation parameters. In the context of multimodal classification, two primary approaches were investigated: the Qwen-Audio-Chat model, which was fine-tuned using LoRA, and a combination of Gemma with two different audio encoders, Whisper and Emo2vec.
- **UAE** (Lagos-Ortiz, Medina-Moreira, and Apolinario-Arzube, 2024). This team placed 9th position in Task 1 and 7th in Task 2. Participants explored the potential of combining different text and audio models to accurately detect emotion. In the text-based analysis of Task 1, the authors conducted experiments using text embeddings from the pre-trained language model BETO and classified emotions using the SVM algorithm. For the audio component in Task 2, they extended the approach from Task 1 by incorporating audio features from the Wav2Vec 2.0 model.
- **UAH–UVA** (Chaves-Villota, Jimenez, and Bahillo, 2024). This team placed 5th position in Task 1. They used a transfer learning approach with a pre-trained DistilBERT model (Sanh et al., 2019), which addresses class imbalance by applying a class weighting technique to avoid majority class bias. The weighted model, which gives more weight to the minority class (i.e., fear) and less to the majority class (i.e., neutral), outperforms the unweighted model in emotion classification, as evidenced by a higher Macro F1-Score.
- **UKR** (Gladun, Rogushina, and Martínez-Béjar, 2024). This team placed 4th position in Task 1 and 6th in Task 2. In Task 1, the UKR team focused on fine-tuning the pre-trained language model BETO using textual data, and achieved the fourth best performance. For Task 2, the team improved their approach by incorporating MFCCs along with text during fine-tuning, resulting in the sixth best performance, outperforming the baseline. In contrast to the other teams participating in both tasks, the BETO-MFCC model used in Task 2 demonstrated inferior performance relative to Task 1. This suggests that the incorporation of MFCC audio features does not significantly enhance the model's performance in multimodal emotion classification.
- **UNED–UNIOVI** (Martinez-Romo et al., 2024). This team placed 3rd position in Task 1 and 5th in Task 2. Participants explore the potential of combining different text and audio models to accurately detect emotion. In the text-based analysis of Task 1, the authors conducted experiments to fine-tune two models on the data provided by the organizers. One model was built on top of RoBERTuito (Pérez et al., 2022) using the TASS 2020 corpus, while the other was built on top of the XML-RoBERTa-base (Conneau et al., 2020) architecture trained on tweets from the EmoEval 2021 shared task. For the audio component in Task 2, 64 acoustic, prosodic and spectral features are extracted, and different clas-

sifiers are tested. The Support Vector Classifier (SVC) turns out to be the most effective, outperforming the other classifiers. A strategic ensemble approach using a hard voting system is used to integrate text and audio features.

- **UTP** (Cedeño-Moreno et al., 2024). This team placed 10th position in Task 1 and 8th in Task 2. In Task 1, UTP used text embeddings from a FastText model and Random Forest algorithm for emotion classification, achieving a Macro F1-Score of 41.02%. In Task 2, the team improved their approach by incorporating audio features extracted from a pretrained Wav2Vec 2.0 model. This adaptation resulted in an improved Macro F1-Score of 48.15%. Although these results did not outperform the baseline, they demonstrate the synergistic benefit of combining audio features with text embeddings to improve overall performance.

5 Results and discussion

The official leaderboards for the EmoSPeech 2024 shared task are shown in Table 3 for Task 1 and in Table 4 for Task 2.

Ranking	Team	Macro F1-Score
01	ITST	67.186
02	CogniCIC	65.753
03	UNED–UNIOVI	65.529
04	UKR	64.842
05	UAH–UVA	61.750
06	THAU–UPM	58.314
07	LACELL	52.882
08	SINAI	52.000
09	UAE	51.824
–	BASELINE	49.683
10	UTP	41.023
11	adri28	37.852
12	Iris5	33.459
	Mean	*59.310*

Table 3: EmoSPeech official leaderboard for Task 1, including the official ranking and result.

As can be observed, not all participants engaged with the multimodal aspect of AER, as only nine teams were included in the official ranking of Task 2. It would have been informative to observe the participants utilizing the proposed baseline to construct multimodal AER, leveraging straightforward methodologies such as ensemble learning. By employing this strategy, it would have been possible to ascertain whether the incorporation of audio features enhances the performance of text-only systems.

Ranking	Team	Macro F1-Score
01	BSC–UPC	86.689
02	THAU–UPM	82.483
03	CogniCIC	71.226
04	ITST	68.758
05	UNED–UNIOVI	67.093
06	UKR	57.797
07	UAE	55.889
–	BASELINE	53.076
08	UTP	48.156
09	adri28	9.417
	Mean	*68.439*

Table 4: EmoSPeech official leaderboard for Task 2, including the official ranking and result.

In this context, it can be observed that the most effective multimodal AER system outperforms the most effective text-based AER system by a margin of 19.503%. This is also evidenced by the baseline, where the results improve from 49.683% to 53.076%. Moreover, the majority of participants who submitted results to both tasks demonstrated an improvement in their results from text-based to multimodal AER systems. For instance, THAU–UPM achieved a 24.169% improvement. However, there were also participants whose results were lower in multimodal AER systems. This was the case for UKR, which exhibited a 7.045% decline in Task 2.

A review of the mean results reveals that the text-based AER system attained an average Macro F1-score of 59.310%, while the multimodal AER system achieved an average of 68.439%. It is noteworthy that the latter result includes the output of one participant in Task 2, which fell below 10%.

6 Conclusions

This paper presents the first edition of the EmoSPeech shared task in IberLEF 2024, which consists of two subtasks. The first subtask concerns to the extraction of AER from text, which involves the identification of salient features and the determination of

the most representative ones of each emotion within a dataset derived from real-life scenarios. The second subtask deals with AER from a multimodal perspective. This requires the construction of a more complex architecture to solve the classification problem. The dataset for this task was compiled by collecting audio segments from different Spanish YouTube channels and was annotated with the emotions of disgust, anger, joy, sadness, and fear, as well as a neutral emotion.

As this is the first time we have organized this event, we are very pleased with the response, with the registration of 38 users and the participation of 13 teams, who sent promising approaches for solving both tasks, most of them based on Transformers. The shared task is still accessible in the post-evaluation phase `https://codalab.lisn.upsaclay.fr/competitions/17647#learn_the_details`. The Codalab page contains two notebooks for the preparation of the baseline and the submission file to be sent to the competition, as well as the full dataset including the golden labels. It is our hope that these resources will prove beneficial to the Spanish NLP community.

It is our intention to organize a second edition of this task. In this context, the dataset will be expanded to include additional channels and modalities, such as video.

Acknowledgments

This work is part of the research project LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MCIN/AEI/10.13039/501100011033 and the European Fund for Regional Development (FEDER)-a way to make Europe, and 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. Mr. Ronghao Pan is supported by the Programa Investigo grant, funded by the Region of Murcia, the Spanish Ministry of Labour and Social Economy and the European Union - NextGenerationEU under the "Plan de Recuperación, Transformación y Resiliencia (PRTR)".

References

Almela, A., P. Cantos-Gómez, D. G.-M. no, and G. Alcaraz-Mármol. 2024. LACELL at EmoSPeech-IberLEF2024: Combining Linguistic Features and Contextual Sentence Embeddings for Detecting Emotions from Audio Transcriptions. 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.*

Boyd, R. L., A. Ashokkumar, S. Seraj, and J. W. Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, 10:1–47.

Cañete, J., G. Chaperon, R. Fuentes, J. Ho, H. Kang, and J. Pérez. 2023. Spanish Pre-trained BERT Model and Evaluation Data. *CoRR*, abs/2308.02976.

Casals-Salvador, M., F. Costa, M. India, and J. Hernando. 2024. BSC-UPC at EmoSPeech-IberLEF2024: Attention Pooling for Emotion Recognition. 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.*

Cedeño-Moreno, D., M. Vargas-Lombardo, A. Delgado-Herrera, C. Caparrós-Láiz, and T. Bernal-Beltrán. 2024. UTP at EmoSPeech–IberLEF2024: Using Random Forest with FastText and Wav2Vec 2.0 for Emotion Detection. 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.*

Chaves-Villota, A., A. Jimenez, and A. Bahillo. 2024. UAH-UVA at EmoSPeech-IberLEF2024: A Transfer Learning Approach for Emotion Recognition in Spanish Texts based on a Pre-trained DistilBERT Model. In *Proceedings of the Iberian Languages Evalu-*

ation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org.

Chenchah, F. and Z. Lachiri. 2016. Speech emotion recognition in noisy environment. In *2nd International Conference on Advanced Technologies for Signal and Image Processing, ATSIP 2016, Monastir, Tunisia, March 21-23, 2016*, pages 788–792. IEEE.

Chiruzzo, L., S. M. Jiménez-Zafra, and F. Rangel. 2024. 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.*

Conneau, A., A. Baevski, R. Collobert, A. Mohamed, and M. Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, editors, *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 2426–2430. ISCA.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

de la Rosa, J., E. G. Ponferrada, M. Romero, P. Villegas, P. G. de Prado Salas, and M. Grandury. 2022. BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling. *Proces. del Leng. Natural*, 68:13–23.

Ekman, P. 1992. Facial expressions of emotion: New findings, new questions. *Psychological Science*, 3(1):34–38.

Esteban-Romero, S., J. Bellver-Soler, I. Martín-Fernández, M. Gil-Martín, L. F. D'Haro, and F. Fernández-Martínez. 2024. THAU-UPM at EmoSPeech-IberLEF2024: Efficient Adaptation of Mono-modal and Multi-modal Large Language Models for Automatic Speech Emotion Recognition. 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.*

Fahad, M. S., A. Ranjan, J. Yadav, and A. Deepak. 2021. A survey of speech emotion recognition in natural environment. *Digit. Signal Process.*, 110:102951.

García-Baena, D., M. A. García-Cumbreras, and S. M. Jiménez-Zafra. 2024. SINAI at EmoSPeech-IberLEF2024: Evaluating Popular Tools and Transformers Models for Multimodal Speech-Text Emotion Recognition in Spanish. 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.*

Gladun, A., J. Rogushina, and R. Martínez-Béjar. 2024. UKR at EmoSPeech–IberLEF2024: Using Fine-tuning with BERT and MFCC Features for Emotion Detection. 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.*

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Lagos-Ortiz, K., J. Medina-Moreira, and O. Apolinario-Arzube. 2024. UAE at EmoSPeech–IberLEF2024: Integrating Text and Audio Features with SVM for Emotion Detection. 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.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Lugovic, S., I. Dunder, and M. Horvat. 2016. Techniques and applications of emotion recognition in speech. In P. Biljanovic, Z. Butkovic, K. Skala, T. G. Grbac, M. Cicin-Sain, V. Sruk, S. Ribaric, S. Gros, B. Vrdoljak, M. Mauher, E. Tijan, and D. Lukman, editors, *39th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2016, Opatija, Croatia, May 30 - June 3, 2016*, pages 1278–1283. IEEE.

Martinez-Romo, J., J. F. Huesca-Barril, L. Araujo, and E. de La Cal Marin. 2024. UNED-UNIOVI at EmoSPeech-IberLEF2024: Emotion Identification in Spanish by Combining Multimodal Textual Analysis and Machine Learning Methods. 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*.

Mohammad, S. M. and F. Bravo-Marquez. 2017. WASSA-2017 Shared Task on Emotion Intensity. In A. Balahur, S. M. Mohammad, and E. van der Goot, editors, *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSAEMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 34–49. Association for Computational Linguistics.

Nguyen, N., X. Vu, C. Rigaud, L. Jiang, and J. Burie. 2021. ICDAR 2021 Competition on Multimodal Emotion Recognition on Comics Scenes. In J. Lladós, D. Lopresti, and S. Uchida, editors, *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part IV*, volume 12824 of *Lecture Notes in Computer Science*, pages 767–782. Springer.

Pan, R., J. A. García-Díaz, M. Á. Rodríguez-García, and R. Valencia-García. 2024. Spanish MEACorpus 2023: A multimodal speech-text corpus for emotion analysis in Spanish from natural environments. *Computer Standards & Interfaces*, page 103856.

Paredes-Valverde, M. A. and M. d. P. Salas-Zárate. 2024. Team ITST at EmoSPeech-IberLEF2024: Multimodal Speech-text Emotion Recognition in Spanish Forum. 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*.

Pérez, J. M., D. A. Furman, L. A. Alemany, and F. M. Luque. 2022. RoBERTuito: a pre-trained language model for social media text in Spanish. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 7235–7243. European Language Resources Association.

Plaza del Arco, F. M., S. M. Jiménez-Zafra, A. Montejo-Ráez, M. D. Molina-González, L. A. Ureña López, and M. T. Martín-Valdivia. 2021. Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021. *Proces. del Leng. Natural*, 67:155–161.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Soto, M., C. Macias, M. Cardoso-Moreno, T. Alcántara, O. García, and H. Calvo. 2024. CogniCIC at EmoSPeech-IberLEF2024: Exploring Multimodal Emotion Recognition in Spanish: Deep Learning Approaches for Speech-Text Analysis. 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.

Varghese, A. A., J. P. Cherian, and J. J. Kizhakkethottam. 2015. Overview on emotion recognition system. In *2015 international conference on soft-computing and networks security (ICSNS)*, pages 1–5. IEEE.

Villegas, M. 2023. MarIA: Spanish Language Models. In A. P. Rocha, L. Steels, and H. J. van den Herik, editors, *Proceedings of the 15th International Conference on Agents and Artificial Intelligence, ICAART 2023, Volume 1, Lisbon, Portugal, February 22-24, 2023*, page 9. SCITEPRESS.

Zhang, D., Y. Yu, C. Li, J. Dong, D. Su, C. Chu, and D. Yu. 2024. MM-LLMs: Recent Advances in MultiModal Large Language Models. *CoRR*, abs/2401.13601.

Zheng, T. F., G. Zhang, and Z. Song. 2001. Comparison of Different Implementations of MFCC. *J. Comput. Sci. Technol.*, 16(6):582–589.