LACELL at EmoSPeech-IberLEF2024: Combining Linguistic Features and Contextual Sentence Embeddings for Detecting Emotions from Audio Transcriptions

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

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

Abstract

These working notes summarize the participation of the LACELL team in the EmoSPeech 2024 shared task, focused on multimodal emotion recognition, which combines textual and intonation features to comprehensively understand human emotions. Its application in Spanish is crucial due to the language's vast global presence, enabling more accurate emotion recognition and fostering better cross-cultural communication and emotional insight in diverse Spanish-speaking communities. We participated in the textual task with a combination linguistic features from LIWC and sentence embeddings from MarIA using ensemble learning, achieving the 7th position with a macro f1-score of 52.882%. This result outperformed the baseline by 3.199 points.

Keywords

LIWC, Linguistic Features, Emotion Classification, Natural Language Processing

1. Introduction

Emotion Recognition (ER) is an essential task for building positive relationships, whether in person or through computer interactions [1]. ER is not an easy task, as there is not even scientific consensus on the definition of emotion, much less on the operationalization of this research construct. Due to the inherent difficulty of defining observable and measurable components of emotional behavior, Automatic Emotion Recognition (AER) has been a significant challenge for many years. It is gaining importance due to its impact on healthcare, psychology, social sciences, and marketing [2], as AER can provide personalized responses and recommendations, thereby increasing user engagement and satisfaction.

AER can be approached using different taxonomies, with the most popular recognizing six basic emotions: anger, disgust, fear, happiness, sadness, and surprise [3]. In this regard, it is worth noting that, even though researchers are increasingly split over the validity of Ekman's conclusions on universality and his assumptions on non-verbal expression of emotions [4], it does not affect the linguistic expression of emotions in a specific language.

The EmoSpeech 2024 shared-task [5] from IberLEF 2024 [6] aims to deepen the AER field by addressing its inherent challenges. A key issue is to identify the features that are relevant for discriminating between emotions. In order to fulfill this task, a major challenge is the scarcity of multimodal datasets that reflect real-life scenarios, as many existing datasets are derived from artificial situations that lack genuine emotional expressions. Furthermore, the complexity of the classification problem is increased

IberLEF 2024, September 2024, Valladolid, Spain
*Corresponding author.
†These authors contributed equally.
angelalm@um.es (Á. Almela); pcantos@um.es (P. Cantos-Gómez); daniel.granadosm@um.es (D. Granados-Meroño); gema.alcaraz@uclm.es (G. Alcaraz-Mármol)
https://portalinvestigacion.um.es/investigadores/331758/detalle (Á. Almela);
https://portalinvestigacion.um.es/investigadores/330963/detalle (P. Cantos-Gómez);
https://portalinvestigacion.um.es/investigadores/332724/detalle (D. Granados-Meroño);
https://www.researchgate.net/profile/Gema-Alcaraz-Marmol (G. Alcaraz-Mármol)
0000-0002-1327-8410 (Á. Almela); 0000-0001-6329-2352 (P. Cantos-Gómez); 0000-0002-5305-1376 (D. Granados-Meroño); https://orcid.org/0000-0001-7703-3829 (G. Alcaraz-Mármol)

by the combined use of multiple features, making it difficult to design advanced architectures that can integrate a wide range of features. Indeed, multimodal AER can identify, interpret and respond to emotions expressed through different modalities such as text, images, and audio. Image modalities can capture data from facial expressions and body language, while speech modalities can capture data from voice tone, intensity, duration or rhyme. The integration of these features in a multimodal paradigm, combining text and speech data, improves performance in emotion recognition tasks.

Nonetheless, instead of adopting a multimodal approach to the task, our team focused exclusively on the text task with a combination of linguistic features from LIWC and sentence embeddings from MarIA, achieving the 7th position with a macro f1-score of 52.882%. This result outperformed the baseline by 3.199 points.

2. Dataset

According to the organizers, the EmoSpeech 2024 dataset consists into audio segments from different Spanish YouTube channels. The underlying assumption is that certain topics elicit different emotional responses from content creators when they express their opinions. For example, it was observed that politicians on politics channels often conveyed disgust towards opposing parties, while interviews with athletes in sports contexts often showed anger after a loss.

The dataset is a subset of 3k audio segments of a larger corpus named Spanish MEACorpus 2023 [7]. The organizers of the task first released a development dataset but we did not use it. Besides, we selected a subset of 25% for the training annotations to build a custom development split for testing and hyperparameter optimization. Table 1 summarizes the statistics of the dataset. The dataset is unbalanced, with more documents expressing disgust and neutral emotions. Fear is the emotion with fewer examples.

Table 1
EmoSpeech 2024 statistics

Emotion	Train	Val	Test	Total
Anger	299	100	100	499
Disgust	528	177	177	882
Fear	17	6	6	29
Joy	271	91	90	452
Neutral	874	292	291	1457
Sadness	258	87	86	431
Total	2247	753	750	3000

To analyze the dataset, we used the UMUTextStats tool [8] to obtain the linguistic features used by emotion (see Figure 1). We observed that features related to part-of-speech (nouns, conjunctions, articles, and pronouns) are relevant, as well as features related to spelling errors, use of title case (especially relevant for documents annotated as fear and sadness), and forms of politeness, which are not common in texts expressing disgust or sadness, but very common in documents expressing fear and joy.

3. System description

We evaluated LIWC [9] as linguistic features. On the one hand, the 2022 version of LIWC, the de-facto linguistic analysis tool that extracts a vector of psychological dimensions of language data from text documents. It is worth noting that the last version available for Spanish is from 2007 [10], as the subsequent versions of the software for English (LIWC2015 and LIWC-22) have not been translated into Spanish yet. On the other hand, UMUTextStats [8] is a linguistic extraction tool designed for Spanish language analysis, addressing specific linguistic phenomena that conventional tools like LIWC

overlook. Unlike LIWC, UMUTextStats is tailored to take into account nuances such as grammatical gender and different verb tenses inherent to the Spanish language. Furthermore, UMUTextStats has been successfully applied in various research areas, including hate speech [11] or satire [12] detection, among others.

Before extracting the LFs from LIWC, a preprocessed version of the transcriptions are generated. The second version is used to extract Part-of-Speech (PoS) features. This version lacks hyperlinks, hashtags, mentions, digits and percentages. Some of these symbols are replaced with a fixed token and others are replaced. Expressive lengthening has been removed and misspellings are fixed using ASPELL tool[1]. It is worth noting that we keep the original audio transcription to extract LFs concerning correction and style.

As for the LLMs, we focused on two Spanish Large Language Models: MarIA [13] and BETO [14], which are based, respectively, on RoBERTa and BERT architectures. We use [15] to extract sentence embeddings from the audio transcriptions.

Table 2
Hyperparameters for fine-tuning the LLMs

LLM	lr	epochs	warmup steps	weight decay
BETO	4.5e-05	4	250	0.19
MARIA	1.8e-05	5	0	0.031

As both feature sets (LFs and sentence embeddings) are encoded as vectors, we could combine them to build stronger models. Specifically, we evaluated ensemble learning, combining the output of models trained with only one feature set using different strategies. In our work, we evaluated the strategy of combining these features using the mode, different ensemble learning strategies based on obtaining the mode, the average of the probabilities, and obtaining the emotion predicted with the highest probability.

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

[1] http://aspell.net/

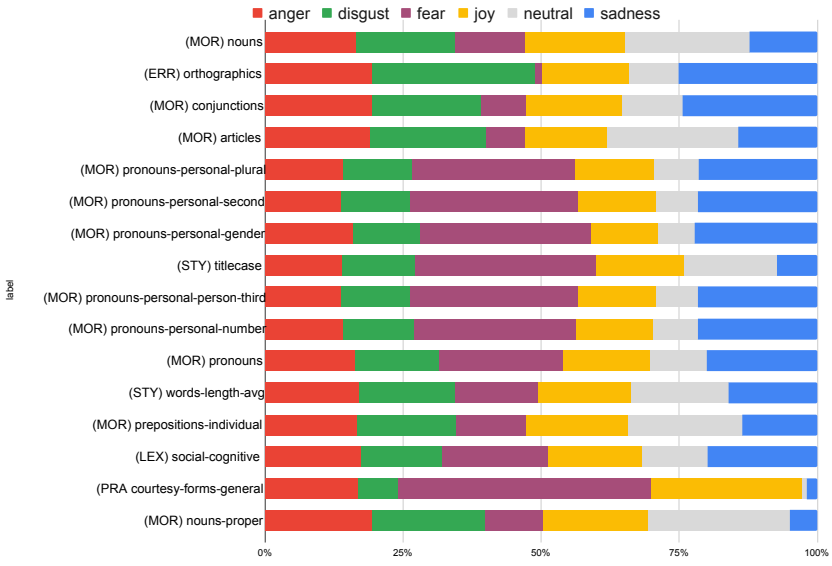

Figure 1: Information gain of the dataset with the stacked values organised by emotion

In order to combine the LLMs and LIWC, we train a traditional neural network with the inputs with another hyperparameter tuning. The results of this process is shown in Table 3. As it can be observed, all the resulting neural networks are shallow, composed by one or two layers, even in the case of the LIWC features. For the LLMs, the simplicity of the networks is expected, as the sentence embeddings were already adjusted for each emotion.

Table 3
Best hyperparameters per model

features	shape	# of layers	neurons	dropout	lr	batch size	activation
LIWC	brick	2	8	0	0.01	64	linear
BETO	brick	1	16	False	0.01	32	linear
MARIA	brick	2	128	0.3	0.01	64	sigmoid

First, we present the experiments with the custom validation split in Table 4. The results are organized by the LIWC linguistic features in the first subsets of rows, the sentence embeddings of the LLMs in the second set of rows and the feature integration strategies in the last set of rows. From the results, it can be observed that LIWC-22 achieved limited results compared with the sentence embeddings. With the sentence embeddings, the performance of BETO and MarIA are similar with better macro f1-score of MarIA and better precision but a slightly more limited recall. Concerning the feature integration strategies, the best results are achieved using an ensemble based on highest probability. However, our previous background yielded bad results when passing from custom validation to official test sets and we decided to submit the ensemble based on the mode as our final submission.

4. Results

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

4.1. Validation

Table 4
Results with the validation split

Strategy	precision	recall	f1-score
LIWC	41.643	40.814	39.997
BETO	70.959	72.856	71.520
MarIA	76.348	71.117	73.117
Ensemble Learning / HIGHEST	**76.240**	**68.364**	**70.855**
Ensemble Learning / MEAN	75.715	67.623	70.211
Ensemble Learning / MODE	64.923	59.946	60.431

Next, we show the detailed classification report of the ensemble learning based on the mode with the custom validation split in Table 5. This report includes the precision, recall, and f1-score of all emotions as well as the macro and weighted values. The model achieved similar weighted and macro f1-scores, which indicates that it performs well regardless the emotion, including fear, that was the most underrepresented one. However, the precision of some emotions is not very high, as it is the case of anger and joy.

4.2. Official results

Table 6 depicts the official leaderboard for the competition. Our team ranked 7th from a total number of 12 participants and improved the baseline (52.882% vs 49.683% of macro F1-score). It is worth noting that CIPIN team outperformed our best result, 84.993%, but the team was not in consideration for the official leaderboard as they submitted their task a few hours later according to the organizers.

As it can be observed from Table 6, we achieved 7th position with a macro f1-score of 52.88210% with a combination of LIWC features and MarIA using an ensemble based on the mode. This results outperformed the proposed baseline based on statistical features based on TF–IDF by 3.1992 points, but it was 14.3035% lower than the 1st team, that achieved a macro f1-score of 67.18560%. It is worth noting that we would have achieved the 8th position if the CICIPN team had submitted their runs on time, as they achieved slightly better results than our approach.

4.3. Error Analysis

To conduct the error analysis, we obtained the confusion matrix of MarIA and LIWC ensemble learning based on the mode with the custom validation split (see Figure 2).

As expected, documents considered neutral are hard to classify. When our model output is neutral, there were 8 documents tagged as anger, 13 as disgust, 1 as fear, 4 as joy, and 12 as sadness, but there was a major number of missclassifications for the actual neutral documents, as 71 of them were identified as disgust, 32 as joy, and 18 as anger. We observed that our model tends to confuse anger and disgust.

Table 5
Classification report of the ensemble learning strategy based on the mode with the custom validation split.

	precision	recall	f1-score
anger	44.531	57.000	50.000
disgust	46.457	66.667	54.756
fear	80.000	66.667	72.727
joy	54.386	68.132	60.488
neutral	81.553	57.534	67.470
sadness	82.609	43.678	59.363
macro avg	64.923	59.946	60.431
weighted avg	65.213	59.363	60.166

Table 6
Official leader-board for Task 1

#	Team	MACRO F1-SCORE
1	TEC_TEZUITLAN	67.186
2	mashd3v	65.753
3	UNED-UNIOVI	65.529
4	UKR	64.842
5	AndreaJohanaCV	61.751
6	jaime	58.314
7	**LACELL**	**52.882**
8	SINAI	52.000
9	UAE	51.824
-	Baseline	49.683
10	UTP	41.023
11	adri28	37.852
12	Iris5	33.459
-	CICIPN	54.993

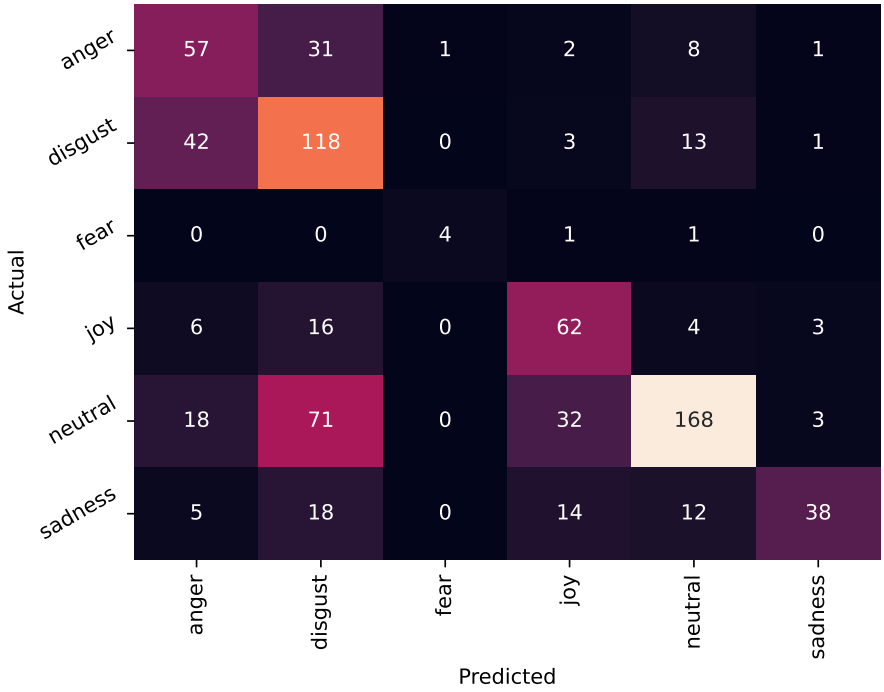

Figure 2: Confusion matrix of the ensemble model based on the mode

4.4. Conclusions and further work

In this working notes, we have described the participation of the LACELL team in the first task of the EmoSpeech 2024 competition, based on textual emotion analysis. Our proposal is grounded on the feature integration of features based on sentence embeddings from MarIA, a Spanish LLM, and linguistic features from LIWC. We reached the 7th position in the official ranking with a macro f1-score of 52.882%, outperforming the baseline by 3.199 points.

As further work, we plan to include features from novel acoustic LLMs in order to participate in multimodal tasks. Specifically, we will evaluate models such as Wav2Vec 2.0, as suggested in [16].

Acknowledgments

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

References

[1] A. A. Varghese, J. P. Cherian, J. J. Kizhakkethottam, Overview on emotion recognition system, in: 2015 international conference on soft-computing and networks security (ICSNS), IEEE, 2015, pp. 1–5.

[2] F. Chenchah, Z. Lachiri, Speech emotion recognition in noisy environment, in: 2016 2nd Interna-

tional Conference on Advanced Technologies for Signal and Image Processing (ATSIP), IEEE, 2016, pp. 788–792.
[3] P. Ekman, Lie catching and microexpressions, The philosophy of deception 1 (2009) 5.
[4] C. Crivelli, J. A. Russell, S. Jarillo, J. M. Fernández-Dols, Recognizing spontaneous facial expressions of emotion in a small-scale society of papua new guinea, Emotion 17 (2017) 337.
[5] R. Pan, J. A. García-Díaz, M. Á. Rodríguez-García, F. García-Sanchez, R. Valencia-García, Overview of EmoSPeech 2024@IberLEF: Multimodal Speech-text Emotion Recognition in Spanish, Procesamiento del Lenguaje Natural 73 (2024).
[6] 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.
[7] 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.
[8] J. A. García-Díaz, P. J. Vivancos-Vicente, A. Almela, R. Valencia-García, Umutextstats: A linguistic feature extraction tool for spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 6035–6044.
[9] R. L. Boyd, A. Ashokkumar, S. Seraj, J. W. Pennebaker, The development and psychometric properties of liwc-22, Austin, TX: University of Texas at Austin (2022) 1–47.
[10] N. Ramírez-Esparza, J. W. Pennebaker, F. A. García, R. Suriá, La psicología del uso de las palabras: Un programa de computadora que analiza textos en español, Revista mexicana de psicología (2007) 85–99.
[11] J. A. García-Díaz, S. M. Jiménez-Zafra, M. A. García-Cumbreras, R. Valencia-García, Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers, Complex & Intelligent Systems 9 (2023) 2893–2914.
[12] 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.
[13] A. Gutiérrez Fandiño, J. Armengol Estapé, M. Pàmies, J. Llop Palao, J. Silveira Ocampo, C. Pio Carrino, C. Armentano Oller, C. Rodriguez Penagos, A. Gonzalez Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022).
[14] 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.
[15] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992.
[16] L. Pepino, P. Riera, L. Ferrer, Emotion recognition from speech using wav2vec 2.0 embeddings, Proc. Interspeech 2021 (2021) 3400–3404.