UTP at EmoSPeech–IberLEF2024: Using Random Forest with FastText and Wav2Vec 2.0 for Emotion Detection

Denis Cedeño-Moreno[1], Miguel Vargas-Lombardo[1], Alan Delgado-Herrera[1], Camilo Caparrós-Láiz[2] and Tomás Bernal-Beltrán[1]

[1] *Universidad Tecnológica de Panamá, Ciudad de Panamá, Panamá*
[2] *Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Murcia, Spain*

Abstract

Automatic emotion recognition (AER) has become increasingly important in fields such as health, psychology, social sciences, and marketing. Within AER, automatic speech recognition focuses on identifying emotions expressed through speech by analyzing features such as fundamental frequency, intensity, rhythm, intonation, and phoneme duration. Multimodal approaches combine information from speech, facial expressions, body language, and text to enhance emotion identification. The goal of the EmoSPeech at IberLEF 2024 is to advance AER by addressing challenges like feature identification, scarcity of multimodal datasets, and the complexity of integrating multiple features. This shared task includes two subtasks: text-based AER and multimodal AER. The novelty of this challenge lies in its multimodal approach, analyzing language model performance on real-world datasets, a first in collaborative tasks. This paper presents the contribution of the *UTP* team to both subtasks. For Task 1, we used text embeddings from a FastText model and classified emotions with the Random Forest algorithm, achieving an M-F1 score of 0.41 and ranking 10th. For Task 2, we enhanced this approach by incorporating audio features from a pre-trained Wav2Vec 2.0 model, resulting in an M-F1 score of 0.48 and ranking 8th. Although these results did not surpass the baseline, they demonstrate that audio features complement text embeddings and improve performance.

Keywords

Speech Emotion Recognition, Automatic Emotion Recognition, Natural Language Processing, Transformers, Random Forest, FastText

1. Introduction

Automatic emotion recognition has been a significant problem for many years, and in recent years its importance has grown due to its impact on various fields such as health, psychology, social sciences, and marketing. For example, [1] shows the relationship between emotions and mental illness, as well as the importance of automatic recognition in the health field. It is a technology that uses algorithms and artificial intelligence techniques to identify and understand the emotions expressed by people through various modalities such as verbal language, body language, facial expressions, and speech prosody. Within automatic emotion recognition, automatic speech recognition refers to the identification of emotions expressed by a person through speech [2] [3]. The AER process involves analyzing acoustic and prosodic features of speech, such as fundamental frequency, intensity, rhythm, intonation, and phoneme duration, to identify patterns associated with different emotional states. These patterns are then used to classify speech into emotional categories such as happiness, sadness, anger, fear, disgust, and others. There are also multimodal approaches, which consist of combining information from different sources, such as speech, facial expression, body language, written text, and others, to identify and understand the emotions expressed by a person [4].

IberLEF 2024, September 2024, Valladolid, Spain
*Corresponding author.
†These authors contributed equally.
denis.cedeno@utp.ac.pa (D. Cedeño-Moreno); miguel.vargas@utp.ac.pa (M. Vargas-Lombardo); alan.delgado@utp.ac.pa (A. Delgado-Herrera); camilo.caparrosl@um.es (C. Caparrós-Láiz); tomas.bernalb@um.es (T. Bernal-Beltrán)
0000-0002-9640-1284 (D. Cedeño-Moreno); 0000-0002-2074-2939 (M. Vargas-Lombardo); 0000-0002-5191-7500 (C. Caparrós-Láiz); 0009-0006-6971-1435 (T. Bernal-Beltrán)

Thus, the goal of the EmoSPeech [5] at IberLEF 2024 [6] is to explore the field of Automatic Emotion Recognition (AER). Challenges associated with this classification problem are addressed, including the identification of meaningful features to distinguish between emotions, the scarcity of multimodal datasets with real-life scenarios, and the added complexity due to the combined use of multiple features. Two challenges are presented: text-based AER and multimodal AER. AER has received considerable attention in the research community, with several joint events demonstrating the growing interest in the field. The novelty of this challenge lies in its multimodal approach to AER, which analyzes the performance of language models on real-world datasets. No previous collaborative task has focused on this specific challenge.

This paper presents the *UTP* team contribution to both subtasks, based on the use of traditional algorithms such as SVM with Wav2vec 2.0 [7] to extract audio features and text embedding of from a pre-trained language model as BETO. The rest of the paper is organized as follows. Section 2 presents the task and dataset provided. Section 3 describes the methodology of our proposed system for addressing subtask 1 and subtask 2. Section 4 shows the results obtained. Finally, Section 5 concludes the paper with some findings and possible future work.

2. Task description

The task is divided into two subtasks with two approaches to address the AER problem: i) identifying emotions across texts ii) multimodal automatic emotion recognition requires a more complex architecture to solve this classification problem. In recent years, AER has received considerable attention from the research community, with several joint events such as WASSA [8], EmoRec-Com [9], and EmoEvalES [10] highlighting the growing interest in this area. The novelty of this work lies in its multimodal approach to AER, analyzing the performance of language models on real datasets. For this purpose, the organizers provided us with the Spanish MEACorpus 2023 dataset, which consists of a set of audio segments collected from different Spanish YouTube channels. This dataset contains over 13.16 hours of audio annotated with six different emotions: disgust, anger, joy, sadness, neutral, and fear. This dataset was annotated in two phases. For this task, approximately 3500-4000 audio segments were selected and divided into training and testing in a ratio of 80%-20%.

To build the model, the training set was divided into two subsets in a ratio of 90-10: training and validation. We used the validation set to adjust the hyperparameters of a model and evaluate its performance during the process of developing and training the machine learning model. Table 1 shows the distribution of the dataset provided by the organizers.

Table 1
Distribution of the datasets

Dataset	Total	Neutral	Disgust	Anger	Joy	Sadness	Fear
Train	2,700	1,070	616	355	330	308	21
Validation	300	96	89	44	37	32	2
Test	750	291	177	100	90	86	6

3. Methodology

Figure 1 shows the general architecture of our approach for these two tasks. For Task 1, which is to identify emotions from text, we used an approach that involves using the FastText model to obtain text embeddings and then applying an Random Forest (RF) classification algorithm. FastText was chosen for its efficiency in generating word embeddings, while RF was chosen for its robust classification capability. This approach focuses on capturing the semantic meaning of texts and using it for emotion classification, considering the complexity of the model and evaluating performance using appropriate metrics.

For Task 2, which focuses on identifying emotions from audio and text, we used an approach that uses a pre-trained Transformers-based model called Wav2vec 2.0, specifically the *facebook/wav2vec2-large-xlsr-53-spanish model*, to obtain vector representations of the audio. These vectors are combined with the text embeddings and used as input to a classification RF model. The goal is to identify emotions from a combination of audio and text, taking advantage of the semantic representation capabilities of the pre-trained model and the robustness of the RF in data classification. Wav2vec 2.0 is a deep learning model developed by Facebook AI Research (FAIR) for self-monitoring in audio processing. This model is primarily used to generate high-quality vector representations (embeddings) of audio, making it particularly useful for classification tasks.

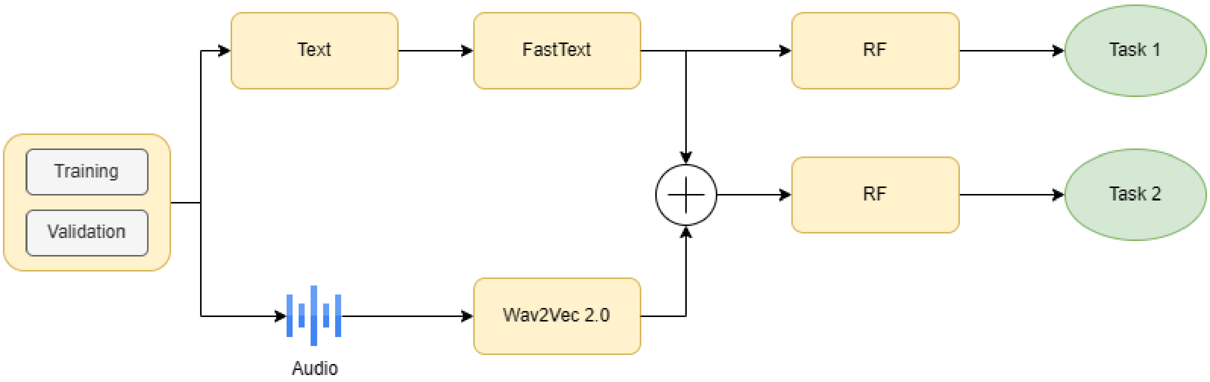

Figure 1: Overall system architecture.

4. Results

Tables 2 and 3 show the results of the RF ranking models in the two defined tasks, as well as the scores obtained by different teams in the official ranking for each task.

For task 1, which consists of identifying emotions through text, it is observed that the RF model obtains a macro F1 score of 0.4102. This score indicates moderate accuracy and recall, suggesting that the model is able to identify emotions in text with some effectiveness, although there is room for improvement.

Comparing this result with the scores of the teams on the official leaderboard, our team "UTP" is ranked 10th with a macro F1 score of 0.4102, indicating that the RF model has achieved results comparable to the teams in the middle of the table. However, there are other teams that have achieved much higher scores, as in this case we have not outperformed the baseline.

For task 2, which focuses on identifying emotions from audio and text, the RF model achieves a macro F1 score of 0.4816. This score is slightly higher than that obtained in Task 1, indicating that the model performs better when audio information is included in addition to text.

Comparing this result with the team scores on the official leaderboard, our team "UTP" also ranks 8th with a macro F1 score of 0.4816. Again, this shows that the RF model has achieved results comparable to other teams on the leaderboard, but there is still room for improvement to reach the highest scores, and we have not passed the baseline.

Overall, the results show that the RF model performs acceptably in both tasks, but there is still room for improvement, especially in identifying emotions from text in Task 1.

To better understand the behavior of the model, we extracted the classification report from the test set for each task. Table 4 shows the classification report for emotion identification from text (Task 1), while Table 5 shows the report for the combined audio and text approach (Task 2).

In Task 1, the RF model shows variable accuracy and recall for different emotion classes. It stands out for its high accuracy and recall for the emotions of happiness and neutrality, but shows lower

Table 2
Results of RF model on the test split for task 1 and task 2 . In this case, the macro precision (M-P), macro recall (M-R), and macro F1-score (M-F1) are reported.

Model	M-P	M-R	M-F1
		Task 1	
RF	0.450811	0.409147	**0.410227**
		Task 2	
RF	0.538039	0.479356	**0.481559**

Table 3
Official leaderboard for task 1 and task 2

Task 1			**Task 2**		
#	**Team Name**	**M-F1**	#	**Team Name**	**M-F1**
1	TEC_TEZUITLAN	0.671856	1	BSC-UPC	0.87
2	CogniCIC	0.657527	2	THAU-UPM	0.866892
3	UNED-UNIOVI	0.655287	3	CogniCIC	0.824833
4	UKR	0.648417	4	TEC_TEZUITLAN	0.712259
-	-	-	-	-	-
10	Baseline	0.496829	9	Baseline	0.530757
10	**UTP**	**0.410227**	**8**	**UTP**	**0.481559**

performance for the emotions of anger and sadness. The weighted average accuracy is 54.03%, suggesting an overall acceptable performance, but with room for improvement.

For Task 2, the RF model also shows variable results for the different emotion classes. It stands out for its remarkable accuracy and recall for the emotion of joy, but fails to predict the emotion of fear at all (with an accuracy, recall and f1 score of 0). The weighted average accuracy is 65.08%, which is slightly better than in Task 1, but still leaves room for improvement.

Table 4
Classification report of RF model in task 1

	Precision	**Recall**	**F1-score**
anger	0.300000	0.060000	0.100000
disgust	0.472393	0.435028	0.452941
fear	0.000000	0.000000	0.000000
joy	0.704225	0.555556	0.621118
neutral	0.598109	0.869416	0.708683
sadness	0.630137	0.534884	0.578616
accuracy	0.576000	0.576000	0.576000
macro avg	0.450811	0.409147	0.410227
weighted avg	0.540314	0.576000	0.536079

5. Conclusion

This paper describes the participation of *UTP* in the IberLEF EmoSPeech 2024 shared task. This task focuses on exploring the field of Automatic Emotion Recognition (AER) from two approaches: i) a textual approach, which uses only textual content to identify the expressed emotion; and ii) a multimodal approach, which combines audios and texts to identify the emotion. Thus, this shared task is divided into two subtasks corresponding to these approaches.

Table 5
Classification report of RF model in task 2

	Precision	Recall	F1-score
anger	0.476190	0.100000	0.165289
disgust	0.516807	0.694915	0.592771
fear	0.000000	0.000000	0.000000
joy	0.833333	0.611111	0.705128
neutral	0.748571	0.900344	0.817473
sadness	0.653333	0.569767	0.608696
accuracy	0.665333	0.665333	0.665333
macro avg	0.538039	0.479356	0.481559
weighted avg	0.650820	0.665333	0.633524

For task 1, we used an approach based on classifying emotions through text embeddings obtained with a FastText model and the Random Forest algorithm, obtaining a score of 0.41 in M-F1, reaching the 10th position in the classification table. On the other hand, for task 2, we have modified the approach used for task 1, adding audio features through a pre-trained audio model based on Wav2Vec 2.0. With this approach, we obtained a score of 0.48 on M-F1, ranking 8th in the leaderboard. Although the results of both tasks have not exceeded the baseline, we can see that the audio features obtained with Wav2Vec 2.0 complement the text embeddings and improve their performance.

As a future line, we plan to improve the approach using fine-tuning techniques and test other classification algorithms, such as recurrent neural networks (RNN), support vector machines (SVM) and convolutional neural networks (CNN). We also propose to test different pre-trained linguistic models such as BETO, MarIA, among others, since in [11], [12], and[13] the good performance of these models in the classification task of different domains has been demonstrated. We also suggest exploring whether sentiment features can enhance emotion detection, given their complementary nature. As demonstrated in [14], this approach has proven effective in various domains, including politics, marketing, healthcare, and others.

References

[1] A. Salmerón-Ríos, J. A. García-Díaz, R. Pan, R. Valencia-García, Fine grain emotion analysis in Spanish using linguistic features and transformers, PeerJ Computer Science 10 (2024) e1992. doi:10.7717/peerj-cs.1992.

[2] 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), 2015, pp. 1–5. doi:10.1109/ICSNS.2015.7292443.

[3] F. Chenchah, Z. Lachiri, Speech emotion recognition in noisy environment, in: 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), 2016, pp. 788–792. doi:10.1109/ATSIP.2016.7523189.

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

[5] R. Pan, J. A. García-Díaz, M. Á. Rodríguez-García, F. García-Sanchez, R. Valencia-García, Overview of EmoSPeech at IberLEF 2024: 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] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[8] S. Mohammad, F. Bravo-Marquez, WASSA-2017 shared task on emotion intensity, in: A. Balahur, S. M. Mohammad, E. van der Goot (Eds.), Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 34–49. URL: https://aclanthology.org/W17-5205. doi:10.18653/v1/W17-5205.

[9] N.-V. Nguyen, X.-S. Vu, C. Rigaud, L. Jiang, J.-C. Burie, ICDAR 2021 competition on multimodal emotion recognition on comics scenes, in: International Conference on Document Analysis and Recognition, Springer, 2021, pp. 767–782.

[10] F. M. Plaza-del Arco, S. M. Jiménez-Zafra, A. Montejo-Ráez, M. D. Molina-González, L. A. Ureña-López, M. T. Martín-Valdivia, Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021, Procesamiento del Lenguaje Natural 67 (2021) 155–161. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6385.

[11] R. Pan, J. 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. doi:10.7717/peerj-cs.1377.

[12] J. A. García-Díaz, S. M. J. Zafra, M. T. M. Valdivia, F. García-Sánchez, L. A. U. López, R. Valencia-García, Overview of PoliticEs 2022: Spanish Author Profiling for Political Ideology, Proces. del Leng. Natural 69 (2022) 265–272. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6446.

[13] J. A. García-Díaz, G. Beydoun, R. Valencia-García, Evaluating Transformers and Linguistic Features integration for Author Profiling tasks in Spanish, Data & Knowledge Engineering 151 (2024) 102307. URL: https://www.sciencedirect.com/science/article/pii/S0169023X24000314. doi:https://doi.org/10.1016/j.datak.2024.102307.

[14] F. Ramírez-Tinoco, G. Alor-Hernández, J. Sánchez-Cervantes, M. Salas Zarate, R. Valencia-García, Use of Sentiment Analysis Techniques in Healthcare Domain, 2019, pp. 189–212. doi:10.1007/978-3-030-06149-4_8.