UAE at EmoSPeech–IberLEF2024: Integrating Text and Audio Features with SVM for Emotion Detection

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

[1]Facultad de Ciencias Agrarias, Universidad Agraria del Ecuador, Av. 25 de Julio, Guayaquil, Ecuador
[2]Universidad Bolivariana del Ecuador, R59R+838, Durán, Ecuador
[3]Instituto Superior Tecnológico Guayaquil, Carlos Gómez Rendón 1403, Guayaquil 090308, Ecuador

Abstract

Automatic emotion recognition (AER) has long been a significant challenge and is becoming increasingly important in various fields such as health, psychology, social sciences, and marketing. The EmoSPeech shared task at IberLEF 2024 aims to advance AER by addressing classification challenges, including feature selection for emotion discrimination, the scarcity of real-world multimodal datasets, and the complexity of combining different features. This task includes two subtasks: text-based AER and multimodal AER, emphasizing the novel aspect of multimodal AER by evaluating language models on authentic datasets. This paper presents the contributions of the *UAE* team to both subtasks. For Task 1, we used text embeddings from the pre-trained language model BETO and classified emotions using the SVM algorithm, achieving an M-F1 score of 0.51, outperforming the baseline and ranking 9th. For Task 2, we extended this approach by incorporating audio features from the Wav2Vec 2.0 model, resulting in an M-F1 score of 0.56 and a ranking of 7th. These results outperformed the baseline, demonstrating that audio features complement text features and improve the performance of the unimodal model.

Keywords

Speech Emotion Recognition, Automatic Emotion Recognition, Natural Language Processing, Transformers, SVM

1. Introduction

Automatic emotion recognition has been a significant challenge for many years and is becoming increasingly important in fields as diverse as health, psychology, social sciences, and marketing. Using algorithms and artificial intelligence, this technology aims to detect and interpret emotions expressed through various channels, including verbal language, body language, facial expressions, and speech prosody [1] [2]. For example, [3] demonstrated the relationship between emotion and mental illness and the importance of emotion recognition in health care. Specifically, within the field of automatic emotion recognition, automatic speech recognition focuses on identifying emotions conveyed through speech. This process involves analyzing acoustic and prosaic features such as fundamental frequency, intensity, rhythm, intonation, and phoneme duration to detect patterns associated with different emotional states, and then categorizing speech into emotional labels such as happiness, sadness, anger, fear, disgust, and others. In addition, multimodal approaches fuse data from multiple sources, such as speech, facial expressions, body language, and written text, to comprehensively capture the emotions conveyed by an individual [4].

The EmoSPeech shared task [5] in IberLEF 2024 [6] aims to delve into the field of Automatic Emotion Recognition (AER) and address the associated classification hurdles, including feature selection for emotion discrimination, the paucity of real-world multimodal datasets, and the complexities arising from the combination of different features. This challenge delineates two subtasks: text-based AER and multimodal AER, reflecting the burgeoning interest in this area as evidenced by numerous collaborative

IberLEF 2024, September 2024, Valladolid, Spain
*Corresponding author.
†These authors contributed equally.
klagos@uagraria.edu.ec (K. Lagos-Ortiz); jjmedinam@ube.edu.ec (J. Medina-Moreira); oapolinario@istg.edu.ec (O. Apolinario-Arzube)
0000-0002-2510-7416 (K. Lagos-Ortiz); 0000-0003-1728-1462 (J. Medina-Moreira); 0000-0003-4059-9516 (O. Apolinario-Arzube)

efforts. In particular, the novelty of this challenge lies in its focus on multimodal AER, assessing the performance of language models on authentic datasets, an aspect previously unexplored in shared tasks.

This paper describes the *UAE* team contributions to both subtasks, using conventional algorithms such as SVM in join with Wav2vec 2.0 [7] for audio feature extraction, and text embedding from a pre-trained language model such as BETO [8]. The following sections provide an overview of the task and the dataset (Section 2), outline the methodology used to address Subtask 1 and Subtask 2 (Section 3), present the results obtained (Section 4), and conclude with lessons learned and avenues for future exploration (Section 5).

2. Task description

The task at hand consists of two different subtasks, each of which presents a different way of approaching the problem of Automatic Emotion Recognition: i) the first subtask deals with the identification of emotions from textual input, ii) the second subtask addresses the more complex challenge of multimodal automatic emotion recognition. In recent years, there has been a surge of interest in AER in the research community, with collaborative events such as WASSA [9], EmoRec-Com [10], and EmoEvalES [11] demonstrating this growing fascination. What distinguishes this work apart is that it takes a multimodal approach to AER, evaluating how language models perform on real-world datasets. To facilitate this, the organizers provided the Spanish MEACorpus 2023 dataset, which includes audio segments collected from various Spanish YouTube channels, amounting to over 13.16 hours of annotated audio spanning six emotions: disgust, anger, happiness, sadness, neutral, and fear. The dataset was annotated in two phases. For this task, about 3500-4000 audio segments were selected and divided into training and test sets in an 80%-20% ratio.

To develop the model, the training set was divided into two subsets with a 90-10 ratio: one for training and the other for validation. The validation set was used to fine-tune the hyperparameters and to evaluate the performance of the model during training. The distribution of the dataset provided by the organizers is shown in table 1.

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 a pre-trained model like BETO to obtain text embeddings and then applying a Support Vector Machine (SVM) classification algorithm. BETO was chosen for its ability to generate highly contextualized word embeddings, capturing the semantic nuances of the text. SVM was selected for its effectiveness in handling high-dimensional data and its robust classification capabilities. This approach focuses on leveraging the advanced text representation capabilities of BETO for emotion classification, evaluating performance using appropriate metrics. The pre-trained language model used for this task is BETO [8], which is a BERT model trained on a large Spanish corpus. BETO is similar in size to the BERT base and was trained using the Whole Word Masking technique. In addition, this model has been shown to perform well especially in classification and author profiling tasks [12] [13].

For Task 2, which focuses on identifying emotions from audio and text, we used an approach that utilizes 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 from BERT and used as input to an SVM classification model. The goal is to identify emotions from a combination of audio and text, taking advantage of the rich semantic representations provided by both pre-trained models and the robust classification capabilities of SVM. Wav2Vec 2.0, developed by Facebook AI Research (FAIR), is designed for self-supervised learning in audio processing, generating high-quality vector representations of audio that are particularly useful for classification tasks. By combining audio and text embeddings, this approach aims to enhance emotion identification accuracy, leveraging the strengths of both modalities and the effectiveness of SVM in data classification.

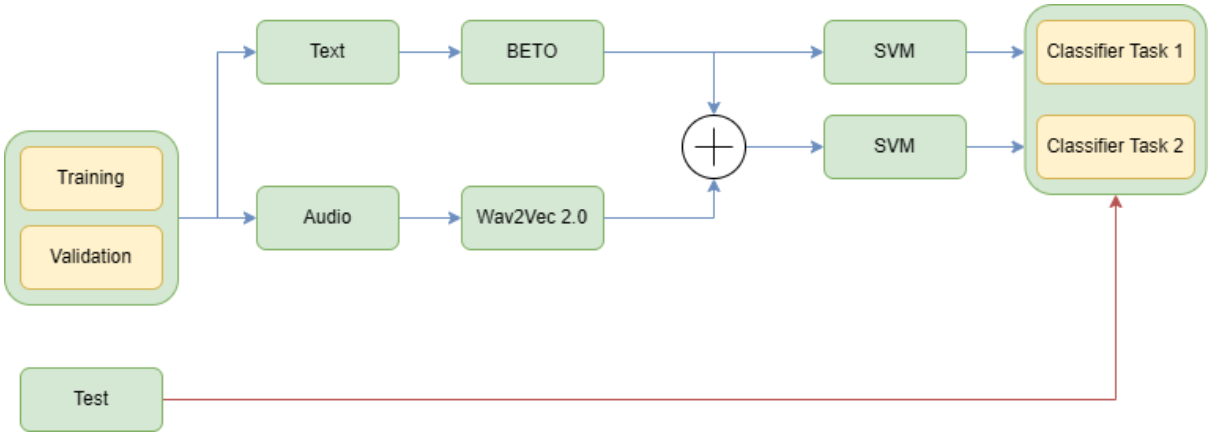

Figure 1: Overall system architecture.

4. Results

The performance of the SVM model on the test split for Task 1 is presented in Table 2. The metrics reported include macro precision (M-P), macro recall (M-R), and macro F1-score (M-F1). The SVM model achieved the following scores: 0.540 in M-P, 0.512 in M-R, and 0.518 in M-F1.

When compared to the official leaderboard for Task 1 (Table 3), the SVM model ranks 9th out of 10, with a macro F1-score of 0.518242. This places it ahead of the baseline model (0.496829) but behind the top-performing team 'TEC_TEZUITLAN" with a macro F1-score of 0.671856. The results indicate that while the SVM model performs better than the baseline, there is significant room for improvement to reach the performance levels of the leading teams.

For Task 2, the SVM model achieved a macro F1-score of 0.558898. This result places the model 7th on the official leaderboard (Table 4), surpassing the baseline score of 0.530757. However, it remains significantly behind the leading team "BSC-UPC" which achieved a macro F1-score of 0.866892.

In conclusion, we can reveal hat the SVM model, while outperforming the baseline in both tasks, does not achieve top-tier performance. The model for Task 1 ranks 9th with a macro F1-score of 0.518242, whereas for Task 2, it ranks 8th with a macro F1-score of 0.558898. These outcomes suggest that further optimization and potentially the exploration of more sophisticated models or additional feature engineering are necessary to improve performance and compete with the leading approaches on the leaderboard.

The classification report for Task 1, as shown in Table 5, provides detailed performance metrics for each emotion class. Our approach achieves an accuracy of 0.682667. The macro averages for precision, recall, and F1-score are 0.540791, 0.512783, and 0.518242 respectively, indicating a balanced but not exceptional performance across different classes. The weighted averages are higher, reflecting the model's better performance on more prevalent classes.

The classification report for Task 2, as shown in Table 6, provides similar performance metrics for each emotion class when both audio and text data are considered. The overall accuracy for Task 2 is

Table 2
Results of the SVM model on the test split for Task 1 and Task 2 are reported. The metrics include macro precision (M-P), macro recall (M-R), and macro F1-score (M-F1).

Model	M-P	M-R	M-F1
Task 1			
SVM	0.540791	0.512783	**0.518242**
Task 2			
SVM	0.571862	0.553863	**0.558898**

Table 3
Official leaderboard for task 1

Task 1		
#	**Team Name**	**M-F1**
1	TEC_TEZUITLAN	0.671856
2	CogniCIC	0.657527
3	UNED-UNIOVI	0.655287
4	UKR	0.648417
-	-	-
9	**UAE**	**0.518242**
10	Baseline	0.496829

Table 4
Official leaderboard for task 2

Task 2		
#	**Team Name**	**M-F1**
1	BSC-UPC	0.866892
2	THAU-UPM	0.824833
3	CogniCIC	0.712259
4	TEC_TEZUITLAN	0.712259
-	-	-
7	**UAE**	**0.558898**
8	Baseline	0.530757

0.717333. The macro averages for precision, recall, and F1-score are 0.571862, 0.553863, and 0.558898 respectively, reflecting a more balanced and slightly improved performance compared to Task 1. The weighted averages show that the model handles the more frequent classes better, similar to Task 1.

Thus, our approach performance varies across different emotion classes, with some classes like *joy* and *neutral* showing high precision and recall, while others like *fear* show very poor performance. The addition of audio data in Task 2 generally improves the model's performance, as evidenced by the higher accuracy and macro metrics. However, the SVM model struggles with less frequent and potentially more ambiguous emotions like *anger* and *fear*.

5. Conclusion

This paper describes the participation of *UAE* in the IberLEF EmoSPeech 2024 shared task. This task focuses on exploring the field of Automatic Emotion Recognition (AER) through two subtasks: i) a textual approach, which uses only textual content to identify the expressed emotion; and ii) a multimodal approach, which combines audio and text to identify the emotion.

Table 5
Classification report of SVM model in task 1

	precision	recall	f1-score
anger	0.408163	0.200000	0.268456
disgust	0.565421	0.683616	0.618926
fear	0.000000	0.000000	0.000000
joy	0.794872	0.688889	0.738095
neutral	0.765766	0.876289	0.817308
sadness	0.710526	0.627907	0.666667
accuracy	0.682667	0.682667	0.682667
macro avg	0.540791	0.512783	0.518242
weighted avg	0.661836	0.682667	0.663992

Table 6
Classification report of SVM model in task 2

	precision	recall	f1-score
anger	0.485294	0.330000	0.392857
disgust	0.586538	0.689266	0.633766
fear	0.000000	0.000000	0.000000
joy	0.786517	0.777778	0.782123
neutral	0.829582	0.886598	0.857143
sadness	0.743243	0.639535	0.687500
accuracy	0.717333	0.717333	0.717333
macro avg	0.571862	0.553863	0.558898
weighted avg	0.704614	0.717333	0.707209

For Task 1, we used an approach based on classifying emotions through text embeddings obtained with a pre-trained language model called BETO and the SVM algorithm, obtaining a score of 0.51 in M-F1, beating the baseline and reaching rank 9th in the classification table. On the other hand, for Task 2, we modified the approach used for Task 1 by adding audio features through a pre-trained audio model based on Wav2Vec 2.0. With this approach, we obtained a score of 0.56 in M-F1, ranking 7th in the table. The results of both tasks have exceeded the baseline proposed by the organizers, and through the results obtained in Task 2, we can conclude that the audio features complement the text features and improve the performance of the unimodal model.

As a future line, we propose to improve the approach using fine-tuning techniques and to test other classification algorithms, such as Recurrent Neural Networks (RNN), Random Forest (RF), and Convolutional Neural Networks (CNN). We also plan to test other pre-trained models based on Transformers. In addition, we propose to add a sentiment feature to the model to enrich its ability to understand and analyze text in different contexts, since sentiment indicates the polarity of sentences and is complementary to emotion. In [14], the analysis is very useful in different domains such as politics, marketing, healthcare, among others.

References

[1] 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.

[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] 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.

[4] R. Pan, J. A. García-Díaz, M. A. 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] 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.

[9] 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.

[10] 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.

[11] 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.

[12] J. A. García-Díaz, S. M. Jiménez-Zafra, M. T. Martín-Valdivia, F. García-Sánchez, L. A. Ureña-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.