UKR at EmoSPeech–IberLEF2024: Using Fine-tuning with BERT and MFCC Features for Emotion Detection

Anatoly Gladun[1], Julia Rogushina[2] and Rodrigo Martínez-Béjar[3]

[1]International Research and Training Center of Information Technologies and Systems of National Academy of Sciences of Ukraine and Ministry of Education and Science of Ukraine, 40, Acad. Glushkov Avenue, Kyiv, 03187, Ukraine
[2]Institute of Software Systems of National Academy of Sciences of Ukraine, 40, Acad. Glushkov Avenue, Kyiv, 03187, Ukraine
[3]Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain

Abstract

Human emotion recognition is crucial for effective social interaction and the development of AI systems that can respond appropriately to users' emotional expressions. Automatic emotion recognition (AER) is a challenging task involving multiple disciplines, including health, psychology, social sciences, and marketing. Its applications range from medical diagnosis and advertising impact assessment to enhancing human-machine interactions. Despite progress, significant challenges persist, particularly in multimodal emotion recognition, which combines text, speech, and facial expressions. The IberLEF EmoSPeech 2024 shared task aims to address these challenges by employing a multimodal approach that integrates textual and auditory data to enhance emotion recognition accuracy. This paper presents the contributions of the UKR team to both subtasks of EmoSPeech 2024. For Task 1, we fine-tuned the pre-trained language model BETO on textual data, achieving the fourth-best result. For Task 2, we incorporated Mel Frequency Cepstral Coefficients (MFCCs) with text during fine-tuning, resulting in the sixth-best performance and surpassing the baseline.

Keywords

Speech Emotion Recognition, Automatic Emotion Recognition, Natural Language Processing, Transformers, Fine-tuning

1. Introduction

Human emotion recognition plays a critical role in social interaction and in the design of artificial intelligence systems that can understand and appropriately respond to users' emotional expressions [1]. From a machine learning perspective, automatic emotion recognition (AER) has emerged as a significant challenge that spans diverse disciplines such as health, psychology, social sciences, and marketing [2]. For example, in [3], you can see the connection between emotion and mental illness, and the importance of recognizing them in health care.

The growing importance of AER is evidenced by its application in contexts as diverse as medical diagnosis, advertising impact assessment in domains like the management and reuse of water (particularly for irrigation practices) and improving human-machine interactions. However, despite the progress made, significant challenges remain in the field, especially in emotion recognition in multimodal situations, where different communication channels such as text, speech, and facial expressions are combined [4].

The goal of this shared task EmoSPeech [5] at IberLEF 2024 [6] is to address these complexities through a multimodal approach that integrates both textual and auditory information to improve the accuracy of emotion recognition. By fusing data from multiple sources, we aim to improve the ability of systems to capture the richness and complexity of human emotional expressions in a variety of contexts.

This paper describes the contributions of the *UKR* team to both subtasks. For both tasks, we take a fine-tuning approach using pre-trained language models, specifically BETO [7]. In Task 1, we train

IberLEF 2024, September 2024, Valladolid, Spain
*Corresponding author.
†These authors contributed equally.
glanat@yahoo.com (A. Gladun); ladamandraka2010@gmail.com (J. Rogushina); rodrigo@um.es (R. Martínez-Béjar)
0000-0002-4133-8169 (A. Gladun); 0000-0001-7958-2557 (J. Rogushina); 0000-0002-9677-7396 (R. Martínez-Béjar)

BETO on textual data only. In Task 2, we extend this by incorporating audio features, specifically Mel Frequency Cepstral Coefficients (MFCCs), along with text during fine-tuning. Our participation in both tasks aims to explore the efficacy of fine-tuning pre-trained language models for AER, and to extend this approach to multimodal analysis by incorporating audio features. The following sections provide an overview of the task and dataset (section 2), describe 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 current task consists of two distinct subtasks, each offering a unique approach to the challenge of automatic emotion recognition: i) the first subtask focuses on emotion recognition from textual input, while ii) the subsequent subtask delves into the more complex realm of multimodal automatic emotion recognition.

Recently, there has been a notable upsurge of interest in AER within the research community, as evidenced by collaborative efforts such as WASSA [8], EmoRec-Com [9], and EmoEvalES [10], which demonstrate the burgeoning interest in the field. What sets this effort apart is the adoption of a multimodal approach to AER, which examines the performance of language models on authentic datasets. To facilitate this exploration, the organizers provided the Spanish MEACorpus 2023 dataset, which consists of audio excerpts from various Spanish YouTube channels, comprising over 13.16 hours of annotated audio spanning six emotions: disgust, anger, happiness, sadness, neutral, and fear. The dataset was annotated in two consecutive phases. For this undertaking, approximately 3500-4000 audio segments were carefully selected and partitioned into training and test sets in an 80%-20% ratio.

In the development of the model, the training set was further divided into two subsets in a 90-10 ratio: one for training purposes and the other for validation. The validation subset was used to fine-tune the hyperparameters and evaluate the performance of the model throughout the training phase. The distribution of the dataset provided by the organizers is shown in table 1. From the table, we can see that there is a significant imbalance between the emotional classes in the training, validation and test split. The emotion *neutral* and *disgust* are the most represented, while *fear* and *sadness* are significantly underrepresented. This imbalance may affect the model's ability to effectively detect underrepresented emotions and bias the results towards the most common classes.

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 architecture of the approaches used for both tasks. For Task 1, we used the approach of fine-tuning a pre-trained language model, such as BETO [7], for emotion classification. Fine-tuning involves taking a model that has been trained on a large-scale task, such as predicting the next word in a text, and tuning it with a specific dataset so that it can perform a classification task. In this way, we can leverage the linguistic knowledge and semantic representation learned during pre-training. The pre-trained language model used for both Task 1 and Task 2 is BETO, a BERT model trained on a large corpus of Spanish text, similar in size to the BERT-Base model, and trained using the Whole Word Masking technique. In addition, this model has demonstrated high performance in various classification tasks in different domains, such as finances [11], author profiling [12] [13], and others.

For Task 2, which focuses on multimodal emotion detection with text and audio, we adapted the previous approach by adding MFCC audio features in the fine-tuning process. This adaptation consists in concatenating the pooled output obtained with BETO with the MFCC vector of the audio. This allows the model to consider both linguistic and acoustic information when making predictions about the emotions expressed in text and audio. The pooled output of BETO refers to a summarized representation of the model output, typically obtained from a pooling process (such as the average or maximum operation) over all token representations. MFCCs, on the other hand, are a feature representation commonly used in audio signal processing to capture the spectral characteristics of the signal. They are computed by a feature extraction process that involves transforming the audio signal into the Mel frequency domain, followed by applying the Discrete Cosine Transform (DCT) to obtain the cepstral coefficients.

Since we did not have sufficient hardware resources, we did not perform hyperparameter optimization, since all training was done on the CPU. Therefore, we used the same hyperparameters for both tasks: a training batch size of 16, 10 epochs, a weight decay of 0.01, and a learning rate of 1e-5.

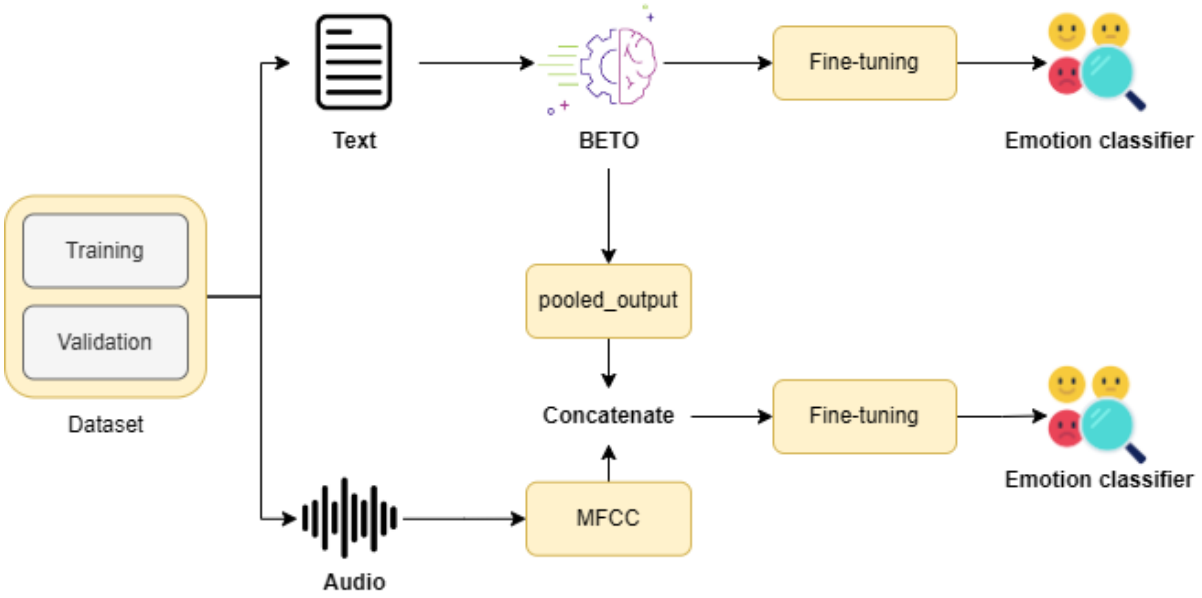

Figure 1: Overall system architecture.

4. Results

For Task 1, the BETO model obtained a macro F1 score of 0.6484, indicating a moderately good performance in text-based emotion classification. In contrast, for Task 2, where MFCC audio features were included in the fine-tuning process, the BETO+MFCC model obtained a macro F1 score of 0.5780. Although this score is lower than that obtained in Task 1, it suggests that the inclusion of audio features can provide valuable additional information to improve emotion recognition in a multimodal context.

Tables 3 and 4 show the results of the ranking table for Task 1 and Task 2. We can see that we ranked fourth in Task 1 with an F1 macro score of 0.6484. This result reflects a solid performance in text-based emotion classification compared to other participating teams. However, in Task 2, we ranked sixth with an F1 macro score of 0.5589. Although this result indicates a decent performance, there is potential for improvement in multimodal emotion classification with the addition of audio features.

To better understand the behavior of the models, we have used the classification report, as shown in Figures 5 and 6. In Task 1, which focuses on text-based emotion classification, the BETO model shows strong overall performance. High accuracy and recall are observed for the *neutral* and *joy* classes, indicating that the model is effective in identifying these emotions. However, the classes *anger* and *fear* have lower accuracies and recalls, suggesting that the model has difficulty distinguishing these

Table 2
Results of the BETO model and MFCC features 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			
BETO	0.6862011	0.634429	**0.648417**
Task 2			
BETO+MFCC	0.626187	0.560649	**0.577969**

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**
-	-	-
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
-	-	-
6	**UKR**	0.558898
8	Baseline	0.530757

emotions in text. Overall, the average F1 score (macro avg) is decent, indicating acceptable performance on the text-based emotion classification task.

On the other hand, for Task 2, which involves multimodal emotion classification with text and audio, the BETO-MFCC model shows somewhat lower performance compared to Task 1. Accuracies and recalls are lower for several classes, including *anger*, *fear*, and *sadness*. This suggests that the incorporation of MFCC audio features does not significantly improve the model's performance in multimodal emotion classification. Although the model achieves reasonable accuracy and recall in the *neutral* and *joy* classes, the average F1 score (macro avg) is lower compared to Task 1.

5. Conclusion

This paper describes UKR's participation in the IberLEF EmoSPeech 2024 shared task. This task is focused on exploring the field of Automatic Emotion Recognition (AER) from two perspectives: i) from a textual point of view, that is, using only textual content to identify the expressed emotion; and ii) from a multimodal point of view, which consists of combining audios and texts to identify the emotion. Thus, this shared task is divided into two subtasks with these two perspectives.

Table 5
Classification report of BETO model in Task 1

	precision	recall	f1-score
anger	0.435897	0.510000	0.470046
disgust	0.632184	0.621469	0.626781
fear	0.666667	0.333333	0.444444
joy	0.719626	0.855556	0.781726
neutral	0.874101	0.835052	0.854130
sadness	0.788732	0.651163	0.713376
accuracy	0.718667	0.718667	0.718667
macro avg	0.686201	0.634429	0.648417
weighted avg	0.728596	0.718667	0.721159

Table 6
Classification report of BETO-MFCC model in Task 2

	precision	recall	f1-score
anger	0.348837	0.300000	0.322581
disgust	0.547619	0.649718	0.594315
fear	0.500000	0.166667	0.250000
joy	0.784810	0.688889	0.733728
neutral	0.822917	0.814433	0.818653
sadness	0.752941	0.744186	0.748538
accuracy	0.678667	0.678667	0.678667
macro avg	0.626187	0.560649	0.577969
weighted avg	0.679556	0.678667	0.676786

For task 1, we used a fine-tuning approach of a pre-trained language model called BETO and obtained consistent results, achieving the fourth-best result in the league table. On the other hand, for task 2, we have modified the fine-tuning approach used for task 1 by adding MFCC audio features in the pooled output of BETO. With this approach, we obtained the sixth-best result, outperforming the baseline.

As a future line, we plan to improve the fine-tuning approach with MFCC features by modifying the classification layer and testing other pre-trained language models and audio features. In addition, we plan to apply the research results to increase social awareness on the reuse and management of water resources for irrigation practices in farming and agriculture settings. We also propose to analyze whether sentiment features can improve emotion detection, as they are complementary to each other. While sentiment analysis focuses on determining the polarity of a text (positive, negative, or neutral), emotions refer to more specific affective states, such as happiness, sadness, fear, etc. Integrating both approaches can provide a more complete and deeper understanding of textual content. In [14], this analysis has proven to be efficient in different fields such as politics, marketing, healthcare, among others.

Acknowledgments

This study formed part of the AGROALNEXT programme and was supported by MCIN with funding from European Union NextGenerationEU (PRTR-C17.I1) and by Fundación Séneca with funding from Comunidad Autónoma Región de Murcia (CARM)

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

[2] 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.

[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. 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] 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.

[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. P. del-Arco y Salud María Jiménez-Zafra y Arturo Montejo-Ráez y M. Dolores Molina-González y L. Alfonso Ureña-López y M. Teresa 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.