Team ITST at EmoSPeech-IberLEF2024: Multimodal Speech-text Emotion Recognition in Spanish Forum

Mario Andrés Paredes-Valverde[1,*,†] and María del Pilar Salas-Zárate[1,†]

[1] *Tecnológico Nacional de México/I.T.S. Teziutlán, Fracción l y ll SN, 73960 Teziutlán, Puebla, Mexico*

Abstract

This work describes the participation of the ITST team in the EmoSPeech 2024 Task - Multimodal Speech-text Emotion Recognition in Spanish. To address Task 1 Text AER (Automatic Emotion Recognition), this work proposed a fine-tuning strategy which involves adapting a set of pre-trained transformer models to a specific task. Regarding Task 2 Multimodal AER, an ensemble learning process was proposed. This approach combining multiple individual models, specifically a wac2vec model and BETO model, to improve predictive performance compared to the performance of each model separately. Furthermore, the mean and maximum probability measures were used to provide a final prediction.

Keywords

transformers, fine tuning, wav2vec

1. Introduction

With the ever-increasing use of electronic devices such as computers and smartphones, emotion recognition has become a significant are of research within the field of human-computer interaction. This area involves the identification and analysis of human emotions through various data inputs, such as facial expressions, voice intonations, textual content, and physiological signals.

In the literature there are several efforts to categorize emotions, for example Ekman (Ekman, 1992) proposed a taxonomy of six discrete emotions that are recognized across different cultures namely anger, disgust, fear, happiness, sadness, and surprise. These emotions are popular because they are easily recognizable through facial expressions and other physiological responses.

In the context of Natural Language Processing (NLP), transformers are a type of deep learning model architecture that aims to solve sequence-to-sequence tasks through a mechanism called attention (Vaswani et al., 2017). Transformers have been used to build NLP-based solutions that solve problems such as machine translation, conversational agents, question-answering, text generation as well as emotion detection. Although

IberLEF 2024, September 2024, Valladolid, Spain

[*] Corresponding author.

[†] These authors contributed equally.

mario.pv@teziutlan.tecnm.mx (M.A. Paredes-Valverde); maria.sz@teziutlan.tecnm.mx (M.P. Salas-Zárate)

0000-0001-9508-9818 (M.A. Paredes-Valverde); 0000-0003-1818-3434 (M.P. Salas-Zárate)

transformers were originally designed for NLP tasks, they have been successfully adapted to a wide range of domains such as image processing, speech processing, time series forecasting, reinforcement learning, and multimodal applications.

The scientific context surrounding AER is rich and multifaceted, encompassing various approaches to recognize emotions from different modalities. Traditional methods primarily relied on handcrafted features extracted from facial expressions, voice signals, and text, with machine learning algorithms such as support vector machines (SVMs) and hidden Markov models (HMMs) classifying these features into discrete emotion categories. With the advent of deep learning, more sophisticated approaches, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM) networks, have emerged to capture spatial and temporal dependencies in data, significantly improving emotion recognition accuracy. In recent years, multimodal approaches that combine data from multiple sources, such as text, speech, and facial expressions, have gained prominence. These models leverage complementary information from each modality, leading to more robust emotion recognition, with transformers playing a key role due to their ability to handle diverse data types through attention mechanisms.

This paper concerns our participation at EmoSPeech 2024 Task - Multimodal Speech-text Emotion Recognition in Spanish (Pan, García-Díaz, Rodríguez-García, García-Sánchez, et al., 2024) which is part of workshop IberLEF 2024 (Chiruzzo et al., 2024). This work implements a fine-tuning strategy which involves adapting a pre-trained transformer model to a specific downstream task, such as named entity recognition, sentiment analysis, or automatic emotion recognition (AER). Specifically, the fine-tuning process leverages the knowledge the transformer model has already acquired during pre-training on a large corpus, thereby requiring less task-specific data and computational resources that training a model from scratch. The corpus used for this task was a multimodal speech-text corpus for emotion analysis in Spanish from natural environments (Pan, García-Díaz, Rodríguez-García, & Valencia-García, 2024).

It is important to mention that our approach obtained first place in the AER from text task, which uses a dataset created from real-life situations, and fourth place in the multimodal AER task, which uses a dataset consisting of more than 13.16 hours of audio from audio segments annotated with five of Ekman's six emotions.

The next section outlines the fine-tuning-based strategies developed to automatically identify five of Ekman's six based emotions (anger, disgust, fear, joy, and sadness) from text as well for from a combination of text and speech cues. In the final remarks, we discuss our findings and propose directions for future research.

2. Developed Strategies

2.1. Task 1 Automatic Emotion Recognition

The aim of the first challenge of the EmoSpeech 2024 Task is to explore the field of AER from text, i.e., extracting features and identifying the most representative feature of each emotion. The dataset used for this task was created from real-life situations; specifically, it consists of 3000 comments for the training phase and 750 for the testing phase. Figure 1 shows the flowchart of the fine-tuning process for automatic emotion recognition proposed in this work. As can be seen, this process consists of five main phases. A brief description of this process is provided below.

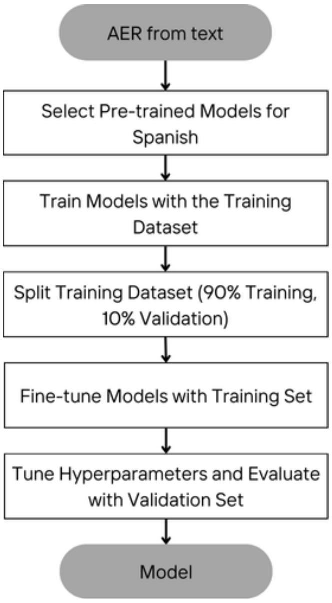
AER from text
Select Pre-trained Models for Spanish
Train Models with the Training Dataset
Split Training Dataset (90% Training, 10% Validation)
Fine-tune Models with Training Set
Tune Hyperparameters and Evaluate with Validation Set
Model

Figure 1: Flow diagram of the fine-tuning process for automatic emotion recognition from text.

1. Select pre-trained models for Spanish. In this case, the models selected are BETO (Cañete et al., 2023), BERTIN (De La Rosa et al., 2022), and MarIA (Gutiérrez-Fandiño et al., 2021). BETO is a BERT-based language model pre-trained exclusively on Spanish data. BERTIN is a model pre-trained using perplexity sampling. Meanwhile, MarIA is a family of Spanish language models thar includes ROBERTa-base, RoBERTa-large, GPT2, and GPT-2large Spanish language models.
2. Train the three pre-trained models with the dataset provided by the challenge.
3. Split the training dataset to perform the fine-tuning process in a 90-10 ratio.
4. Use the training dataset to perform the fine-tuning process and adapt the pre-trained models to the automatic emotion recognition task.
5. Use the validation dataset to tune the model's hyperparameters and to evaluate the model's performance on unseen data. In this case, the hyperparameters were configured as follows: learning rate of 2e-5, a training batch size of 16, and an evaluation strategy based on epoch.

The BETO model achieves an overall accuracy of 73.7%, indicating a good level performance for the AER task from text. Table 1 shows the classification report obtained by this model. As can be seen, BETO performs reasonably well across emotion classes, with particularly strong performance for the "neutral" emotion. However, performance varies across different emotions, with lower precision and recall for classes such as "fear" and "sadness".

Table 1
Classification report of the BETO model for AER task from text.

Emotion	Precision	Recall	F1-score
anger	0.512195	0.525000	0.518519
disgust	0.656716	0.619718	0.637681
fear	0.666667	1.000000	0.800000
joy	0.789474	0.833333	0.810811
neutral	0.832000	0.888889	0.859504
sadness	0.769231	0.588235	0.666667
accuracy			0.736667
macro avg	0.704380	0.742529	0.715530
weighted avg.	0.734556	0.736667	0.733446

The MarIA model demonstrates solid performance for the proposed challenge, achieving an overall accuracy of 74.7%. Table 2 shows the results obtained by this model, where it can be shown that the model performs particularly well for the “neutral” emotion class, with high precision, recall, and F1-score. The “joy” and “disgust” classes also shown strong performance. However, MarIA’s performance on the “anger” and “sadness” classes is somewhat lower. Is should be noted that the “fear” class shows perfect precision but lower recall due to its small support. Overall, the model shows good generalizability with reasonably balanced precision and recall for most emotion classes.

Table 2
Classification report of the MarIA model for AER task from text.

Emotion	Precision	Recall	F1-score
anger	0.460000	0.575000	0.511111
disgust	0.656716	0.619718	0.637681
fear	1.000000	0.500000	0.666667
joy	0.870968	0.750000	0.805970
neutral	0.877049	0.914530	0.895397
sadness	0.758621	0.647059	0.698413
accuracy			0.746667
macro avg	0.770559	0.667718	0.702540
weighted avg	0.755965	0.746667	0.748585

After fine-tuning from AER from text, The BERTIN model achieves an overall accuracy of 71.3%. Table 3 shows the classification report obtained by the BERTIN model. Notably, this model performs well in recognizing “fear” instances. However, there is a room for improvement in recognizing “anger” and “sadness” instances, as indicated by lower precision and recall values.

The BETO, MarIA, and BERTIN models achieved varying levels of performance in AER from text. Overall, all models have strengths and areas for improvement, suggesting the need for ongoing refinement to improve their performance, specifically for less common emotion classes.

Table 3
Classification report of the BERTIN model for AER task from text.

Emotion	Precision	Recall	F1-score
anger	0.400000	0.400000	0.400000
disgust	0.626667	0.661972	0.643836
fear	1.000000	0.500000	0.666667
joy	0.800000	0.777778	0.788732
neutral	0.842975	0.871795	0.857143
sadness	0.714286	0.588235	0.645161
accuracy			0.713333
macro avg	0.730655	0.633297	0.666923
weighted avg	0.714024	0.713333	0.712204

2.2. Task 2: Multimodal Automatic Emotion Recognition

The second challenge of the EmoSpeech 2024 Task was multimodal AER, which aims to analyze the performance of language models in solving this classification problem. The dataset used for this task was the Spanish MEACorpus 2023, consisting of more than 13.16 hours of audio from segments annotated with five of Ekman's six emotions. This dataset comprises about 3500-4000 audio segments, divided into training and test sets in an 80%-20% split. The ensemble learning process followed for automatic emotion recognition from text and speech is shown in Figure 2. This process is described next.

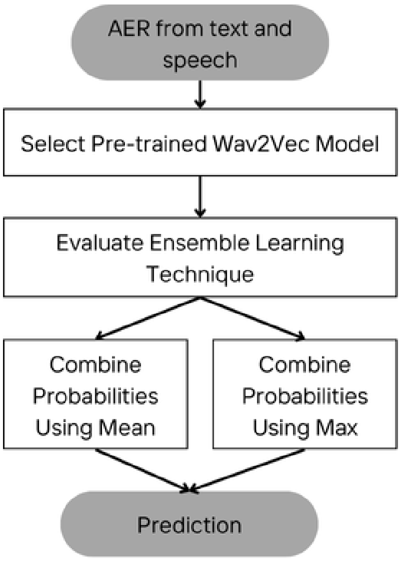
AER from text and speech
Select Pre-trained Wav2Vec Model
Evaluate Ensemble Learning Technique
Combine Probabilities Using Mean
Combine Probabilities Using Max
Prediction

Figure 2: Flow diagram of the ensemble learning process for automatic emotion recognition from text and speech.

1. Select a Wav2Vec pre-trained model for Spanish. A Wav2Vec model is trained in large amounts of unlabeled audio data aiming to improve acoustic model training (Schneider et al., 2019). In this work, the wav2vec2-large-xlsr-53-spanish model (Grosman, 2021) was selected.
2. Evaluate the technique called Ensemble Learning that consists of combining multiple individual models to improve predictive performance compared to the performance of each model separately. In this work, an ensemble that combines the best fine-tuned text model (BETO) and another fine-tuned model from Wav2Vec 2.0 has been used for emotion classification.
    a. Combine the probabilities obtained for each emotion through the two models using Mean. Mean is the mean of the classification probabilities of both sources (text and audio) for each emotion class.
    b. Combine the probabilities obtained for each emotion through the two models using Max. Max is the maximum probability of each emotion class of the two models.
3. Prediction. For both previous cases, the class with the maximum probability is considered as the final prediction.

The ensemble learning process that combines the Wav2Vec pre-trained model with the BETO model achieves an overall accuracy of 75.0% for this challenge using mean of the classification probabilities of both sources. As can be seen in Table 4, this approach performs well across most emotion classes, particularly for "neutral," "joy," and "disgust." It demonstrates the ability to recognize instances of "fear" with high precision and recall. However, the recall is relatively lower in recognizing "sadness" instances. Overall, the ensemble model shows promise in accurately identifying emotions from both text and speech inputs.

Table 4
Classification report of the ensemble learning process based on a Wav2Vec and BETO using mean of the classification probabilities of both sources.

Emotion	Precision	Recall	F1-score
anger	0.50000	0.50000	0.50000
disgust	0.64789	0.64789	0.64789
fear	0.66667	1.00000	0.80000
joy	0.80556	0.80556	0.80556
neutral	0.85600	0.91453	0.88430
sadness	0.84000	0.61765	0.71186
accuracy			0.75000
macro avg	0.71935	0.74760	0.72493
weighted avg	0.75015	0.75000	0.74755

As can be seen in Table 5, the ensemble learning process combining the Wav2Vec pre-trained model with the BETO model using the maximum probability of each emotion class of the two models demonstrates decent precision and recall for some emotion classes like "disgust" and "neutral," it struggles with others such as "anger," "joy," and "sadness," where either precision or recall or both are relatively lower. Notably, it fails to predict any instances of "fear," indicating significant room for improvement. Furthermore, it should be mentioned that this approach achieves an overall accuracy of 70.3%.

Table 5
Classification report of the ensemble learning process based on a Wav2Vec and BETO using the maximum probability of each emotion class of the two models.

Emotion	Precision	Recall	F1-score
anger	0.64286	0.22500	0.33333
disgust	0.57426	0.81690	0.67442
fear	0.00000	0.00000	0.00000
joy	0.64286	0.50000	0.56250
neutral	0.81343	0.93162	0.86853
sadness	0.73913	0.50000	0.59649
accuracy			0.70333
macro avg	0.56876	0.49559	0.50588
weighted avg	0.69977	0.70333	0.67788

The ensemble learning process combining the Wav2Vec pre-trained model with the BETO model exhibits promising performance in automatic emotion recognition from both text and speech. Despite showing promise in certain areas, such as recognizing specific emotions accurately, the ensemble model has clear scope for refinement to enhance its performance across a broader range of emotions. There's a need for further development and optimization to maximize the ensemble's effectiveness in accurately capturing and understanding diverse emotional expressions from both text and speech inputs.

3. Final remarks

This paper presents a fine-tuning approach for EmoSPeech 2024 Task - Multimodal Speech-text Emotion Recognition in Spanish. For AER from text, BETO, MarIA, and BERTIN pre-trained models were used achieving varying levels of performance. Some of the models used in this work obtained lower precision and recall values for less common emotion classes such as “anger” and “sadness”. This fact suggests the need for ongoing refinement to improve their performance. Regarding, the ensemble learning approach described in this work, while the ensemble approach based on Mean achieves an impressive overall accuracy of 75.0%, demonstrating strong performance across multiple emotion classes, another the ensemble approach based on the maximum probability achieves a slightly lower accuracy of 70.3%, indicating areas for improvement, particularly in recognizing certain emotions like "anger," "joy," and "sadness." The obtained results emphasize the potential of ensemble learning approaches in advancing automatic emotion recognition technology, while also emphasizing the ongoing need for optimization and development to achieve more robust and comprehensive emotional understanding from diverse textual and audio inputs.

Acknowledgements

We are grateful to the Tecnológico Nacional de Mexico (TecNM, by its Spanish acronym) for supporting this work. This research was also sponsored by Mexico's National Council of Humanities, Sciences and Technologies (CONAHCYT).

References

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2023). *Spanish Pre-trained BERT Model and Evaluation Data*. https://arxiv.org/abs/2308.02976v1

Chiruzzo, L., Jiménez-Zafra, S. M., & Rangel, F. (2024). Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages. *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*.

De La Rosa, J., Ponferrada, E. G., Villegas, P., González De Prado Salas, P., Romero, M., Grandury, M., & Project, B. (2022). BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling. *Procesamiento Del Lenguaje Natural*, *68*(0), 13–23. https://doi.org/10.26342/2022-68-1

Ekman, P. (1992). Facial Expressions of Emotion: New Findings, New Questions. *Psychological Science*, *3*(1), 34–38. https://doi.org/10.1111/J.1467-9280.1992.TB00253.X

Grosman, J. (2021). *Fine-tuned XLSR-53 large model for speech recognition in Spanish*.

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Gonzalez-Agirre, A., Armentano-Oller, C., Rodriguez-Penagos, C., & Villegas, M. (2021). MarIA: Spanish Language Models. *Procesamiento Del Lenguaje Natural*, *68*, 39–60. https://doi.org/10.26342/2022-68-3

Pan, R., García-Díaz, J. A., Rodríguez-García, M. Á., García-Sánchez, F., & Valencia-García, R. (2024). Overview of EmoSPeech at IberLEF 2024: Multimodal Speech-text Emotion Recognition in Spanish. *Procesamiento Del Lenguaje Natural*, *73*(0).

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

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised Pre-training for Speech Recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2019-September*, 3465–3469. https://doi.org/10.21437/Interspeech.2019-1873

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*.