CogniCIC at EmoSPeech-IberLEF2024: Exploring Multimodal Emotion Recognition in Spanish: Deep Learning Approaches for Speech-Text Analysis

Miguel Soto[1,†], Cesar Macias[1,*,†], Marco Cardoso-Moreno[1,†], Tania Alcántara[1,†], Omar García[1] and Hiram Calvo[1]

[1]*Instituto Politécnico Nacional, Center for Computing Research, Cognitive Sciences Laboratory, Mexico, City, 07700, Mexico*

Abstract

Human emotion recognition, which encompasses both verbal and non-verbal signals such as body language and facial expressions, remains a complex challenge for computing systems. This recognition can be derived from various sources, including audio, text, and physiological responses, making multimodal approaches particularly effective. The importance of this task has grown significantly due to its potential to improve human-computer interaction, providing better feedback and usability in various applications such as social media, education, robotics, marketing and entertainment. Despite its potential, emotional expression is an heterogeneous phenomenon influenced by factors such as age, gender, sociocultural origin and mental health. Our study addresses these complexities and presents our findings from the recent EmoSpeech competition. Our system achieved an F1 score of 0.7256 and a precision of 0.7043, with a precision of 0.7013, in the validation task. For multimodal task 1, our CogniCIC team ranked second with an official F1 score of 0.657527 and for task 2, with an F1 score of 0.712259. These results underline the effectiveness of our approach in multimodal emotion recognition and its potential for practical applications.

Keywords

Emotion Recognition, LLM, Classification, Speech, Text

1. Introduction

Even though, human beings can automatically recognize emotions in other humans by considering verbal and non-verbal expressions (body language and facial expressions), this is still a complicated task for computer systems [1]. Human emotion recognition can be analyzed from several verbal and non-verbal sources, such as: audio, text [2], and physiological responses [3]; therefore, emotion recognition tasks are suitable for multimodal approaches [2, 4]. This task has received increasing attention in recent years [1], since human-computer interaction systems would benefit from the identification of emotions for better feedback [5, 6, 7, 8]. The range of applications is wide, ranging from: social media, education, robotics, marketing and entertainment industries in general [3, 9].

Nevertheless, it is important to understand that emotional expressions are, in general, an heterogeneous phenomenon; variations may arise from different factors such as: age, gender, sociocultural and even mental health [10, 6].

There are several approaches for human emotion recognition, from physiological signals [11, 12] to text-based approaches [13, 14] and speech-based [15, 16]. Moreover, recently there have been approaches dealing with multimodal approaches, or instance, speech-text based solutions [1, 17].

IberLEF 2024, September 2024, Valladolid, Spain

[*] Corresponding author.

[†] These authors contributed equally.

msotoh2021@cic.ipn.mx (M. Soto); cmaciass2021@cic.ipn.mx (C. Macias); mcardosom2021@cic.ipn.mx (M. Cardoso-Moreno); talcantaram2020@cic.ipn.mx (T. Alcántara); ogarciav2024@cic.ipn.mx (O. García); hcalvo@cic.ipn.mx (H. Calvo)

https://mashd3v.github.io/ (M. Soto); https://github.com/MACI-dev-96 (C. Macias); https://cardoso1994.github.io/ (M. Cardoso-Moreno); https://github.com/talcantaram (T. Alcántara); https://www.linkedin.com/in/omar-garcia-vazquez-093128219/ (O. García); http://hiramcalvo.com/ (H. Calvo)

0009-0006-4619-9352 (M. Soto); 0009-0005-1708-5359 (C. Macias); 0009-0001-1072-2985 (M. Cardoso-Moreno); 0009-0001-4391-6225 (T. Alcántara); 0009-0001-4391-6225 (O. García); 0000-0003-2836-2102 (H. Calvo)

The task EmoSPeech 24 [4] is a prominent part of the IberLEF 2024 [18] conference, which focuses on advancements in language processing and related fields. Our participation in this task was structured as follows: Section 2 shows a brief mention of the literature related to this work. Section 3 provides an overview of the dataset employed. Section 4 explains the proposed preprocessing, models and metrics used to evaluate the results. Section 5 shows the results and discussion after carefully evaluation of the models' performance. Finally, Section 6 highlights the strengths, concludes and points out future directions of this research work.

2. Literature Review

In this section, a brief literature review on emotion recognition is presented, from text only proposals, through speech only, and speech and text multimodal approaches.

2.1. Text Based Emotion Recognition

The proposal by deVelasco and colleagues' [1] involved the creation of a new Spanish dataset by selecting speech fragments from a Spanish TV show, which were also transcribed to text. Their solution consists on using a paradigm known as VAD: Valence—related to polarity—, Arousal—related to calmness or excitement—, and Dominance—the degree of control over a situation—; VAD, then, encodes every emotion label as a new, three-dimensional vector, where each element is real valued and corresponds to each one of the VAD parameters. This approach allows for the emotion recognition problem to be stated as a regression task instead of a classification one. deVelasco's solution for text based emotion recognition consisted on using FastText embeddings of 300 dimensions; their proposed model was a Deep Neural Network (DNN) which yielded an MSE error of 0.1196.

In [19] emotions were analyzed by gathering tweets, each one of them with an emotional hashtag. In a first stage the authors used a ConvNet for emotion classification. During the learning process, an embedding model was extracted, which was used to further classify ROC (Receiver Operating Characteristic) curve story text emotions, based on Plutchik's emotions model. The performance of the model ranged from values of 28% to 73% in accuracy, depending on the emotion.

2.2. Speech Based Emotion Recognition

In [1], the Speech based solution consisted on extracting, from audio signals, several acoustic features such as: Pitch, Energy, Spectral Centroid, Spectral Spread, among others. The implemented model was a Long-Short Term Memory (LSTM) cell followed by a Multi-Layer Perceptron (MLP) to solve the regression task according to the VAD paradigm; their model achieved an MSE error of ranging from 0.14 to 0.16 for various subsets of acoustic features.

In [20], two models were proposed for speech emotion recognition, both a combination of ConvNets and an LSTM cell; the main difference being the dimensionality of the convolution layers, one-dimensional and two-dimensional, respectively. This proposal obtained significant results on different benchmark datasets, including IEMOCAP, where it achieved an accuracy of 89.16% for the speaker-depenent configuration, and 52.14% on the speaker-independent experiments.

Furthermore, in [21], Chatziagapi and colleagues propoposed a Generative Adversarial Network (GAN) for synthetic data generation, particularly for balancing the minority class. Once the datasets were in balance, a VGG19 model [22] was instantiated to perform the classification task. The model achieved an average UAR of 53.6%.

2.3. Multimodal Emotion Recognition

In [23], a Multimodal Dual Recurrent Encoder (MDRE) was proposed. The model consisted on two separate encoders: one for text, a Text Recurrent Encoder (TRE); and one for audio, Audio Recurrent Encoder (ARE). Both models consists on a Gated Recurrent Unit (GRU); for ARE Mel-frequency cepstral coefficients (MFFCs) are the corresponding features, whereas for TRE text was tokenized using Natural Language Toolkit (NLTK) [24] and passed through an embedding layer, yielding vectors of 300 dimensions. The dataset used was the IEMOCAP dataset [25]. The proposal achieved accuracy values from 68.8% to 71.8%, being the multimodal approach the one with better performance.

Hazarika et al. [26] proposed a method to fusion both text and audio features at early stages by means of self-attention mechanisms. Feature extraction was made with a Convolutional Neural Network (ConvNet) model, while audio feature extraction was performed with the help of the openSMILE library. Performance was also tested on the IEMOCAP dataset, yielding values around 72% in metrics such as accuracy, F1-scorre and UAR.

Krishna and their colleagues [27] project used cross-modal attention mechanisms, so that audio features attend text features and viceversa, in addition to ConvNets. Features were extracted with two different autoencoders: for audio, thee autoencoder processed the raw signal to extract high-level features; for text, the encoder extracted high-level semantic features. The model achieved an Unweighted Accuracy of 72.82% on the IEMMOCAP dataset.

3. Dataset

The dataset for this competition is the MEACorpus 2023 dataset [28], consisting on audio segments and transcripts from YouTube videos. After the videos were downloaded, audio was extracted in segments; then, an annotation procedure was carried on taking into account the following set of five emotions: disgust, anger, sadness, joy, and fear, based on Ekman's [29] findings—surprise was not considered in the dataset due to the difficulty of extracting it from videos, and a neutral emotion was added—.

The EmoSPeech competition [4] consisted in two tasks, both of which we decided to participate in. The aim of task 1 was to perform text-based emotion recognition only, whereas task 2 was focused on multimodal (text and speech) emotion recognition.

4. Proposal

For each task, we have proposed different approaches, which are described in the following sections.

4.1. Task 1: Text AER

The dataset for Task 1, consisted of text transcripts from audio segments, those texts were written in Spanish. The approach we have carried out for this task was to use large language models (LLMs) based on transformers. Since the texts in the dataset are in Spanish, we chose to fine-tune BETO [30], a BERT model trained on a very large Spanish corpus. For this approach we have not preprocessed the data, the parameters used to fine-tune the BETO model were the following, 15 training epochs (we saved the model with the best performance from all the training epochs, to make the predictions) and the training dataset was divided into 90% for training and 10% for validation to verify the model performance during training, the average used to compute the metrics was macro average, Adam W optimizer was used, and its learning rate and epsilon were set to 3×10^{-6} and 1×10^{-9}, respectively. To make the experiments replicable, the random generation seed was set to 42.

4.2. Task 2: Multimodal AER

The dataset provided for Task 2 consisted of text transcripts and associated audio segments. To prepare the data for further analysis, the following pre-processing steps were carried out:

4.2.1. Data Upload.

The first stage of the process consisted of loading the data provided in CSV files, which contain the transcripts and labels corresponding to the training and test samples.

4.2.2. Data Pre-processing.

This focused on cleaning the textual transcripts. Specifically, the extra blanks at the beginning and end of each transcript were removed. This step was essential to ensure consistency and accuracy in the textual features subsequently extracted.

4.2.3. Feature Extraction

- **Audio feature extraction.** For audio feature extraction, the `Librosa` library was used. Audio files were loaded with a sampling rate of 22050 Hz. From these files, Mel-Frequency Cepstral Coefficients (MFCCs) were extracted. MFCCs are widely used in audio signal processing due to their ability to represent relevant acoustic features [31]. Subsequently, the mean of the MFCCs matrices was calculated to obtain a compact and efficient representation of the audio features.
- **Textual feature extraction.** For textual feature extraction, the pre-trained BERT (Bidirectional Encoder Representations from Transformers) [32] model was used. The `BertTokenizer` was used to tokenize the transcripts and the `BertModel` was used to obtain the vector representations. The BERT outputs were averaged to obtain a fixed representation of each transcript, thus effectively capturing the semantics of the texts.

4.2.4. Label Encoding.

The textual labels provided in the data were transformed into integer values using Scikit-Learn's `LabelEncoder`. This encoding is crucial to allow machine learning models to work with categorical labels efficiently.

4.2.5. Feature and Label Integration.

Once the audio and text features were extracted, they were combined into a final dataset. Each sample was represented as a pair of audio and text features. This integration was carried out for both the training and test sets. The encoded labels were associated with the training samples for use in the modelling phase.

4.2.6. Model Definition.

The implemented model for the multimodal emotion recognition task is based on a neural network architecture designed to combine audio and text features. The model was designed to process and combine audio and text features as follows:

- **Audio subnet.** A linear layer (Linear) that receives audio features and transforms them into a 256-dimensional feature space.
- **Text subnet.** A linear layer that receives textual features and transforms them into a 256-dimensional feature space.
- **Feature combination.** The outputs of the audio and text subnetworks are concatenated and passed through a fully connected network with a ReLU and Dropout layer for regularization. Finally, a linear layer reduces the dimensions to the number of emotion classes, enabling classification.

4.2.7. Model Specifications.

The model was configured with several key specifications to optimize performance. Firstly, the audio feature size was set to 128, providing a detailed representation of the audio data. The dimension of the text features was 768, allowing for a comprehensive capture of textual information. The model was designed to classify into 6 distinct emotion classes, ensuring a nuanced understanding of emotional states.

In the architecture, the intermediate layer of combined features consisted of 256 neurons, enabling effective integration and processing of the audio and text features. For the loss function, cross entropy loss was chosen due to its suitability for classification tasks, particularly in handling multiple classes. The model optimization was handled by the Adam optimizer, known for its efficiency and adaptive learning rate capabilities. The learning rate was set to 0.001, a value selected to balance convergence speed and training stability.

4.2.8. Model Summary Interpretation.

The model summary reveals a detailed architecture designed to process and combine audio and text features efficiently. The first layer for audio processing, denoted as **Linear (audio)**, transforms the audio features from their original 128 dimensions to 256 dimensions. Similarly, the **Linear (text)** layer handles the text features, reducing their dimensionality from 768 to 256 dimensions.

Following the initial transformations, the model employs a **Sequential (combined)** network to integrate the concatenated audio and text features. This sequential network consists of several layers: firstly, a **Linear** layer that reduces the combined feature dimensions from 512 to 256, ensuring a more manageable and computationally efficient size. This is followed by a **ReLU** activation function, which introduces non-linearity into the model, enhancing its ability to capture complex patterns within the data. To prevent overfitting, a **Dropout** layer is included with a dropout rate of 50%, effectively regularizing the model by randomly omitting half of the neurons during training. Finally, a second **Linear** layer is used to produce the output, which matches the number of emotion classes, set to 6 in this case.

The training process was repeated for 150 epochs. This approach allowed the audio and text features to be effectively combined, achieving accurate classification of emotions in the dataset provided for the EmoSPeech 24 competition.

In addition to the basic architecture described above, variants of the model were tested to explore different configurations and improve performance:

- **Implementation of attention modules.** Attention layers were added for each modality (audio and text) in order to highlight the most relevant features before combining them.
- **Multi-head attention.** Multi-head attention layers were implemented to improve the capture of long-term dependencies in audio and text features.
- **Modification of audio vector dimensions.** Experimented with different audio feature dimensions, specifically 56 and 256, to assess their impact on model performance.
- **Use of BETO as a textual feature extractor.** BETO [30], a BERT model adjusted for Spanish, was used as a textual feature extractor instead of the original BERT model, in order to evaluate improvements in the semantic representation of the transcripts.

These variants were implemented and evaluated for their effectiveness in the multimodal emotion recognition task, providing further insight into best practices and configurations for this type of task. However, the original network performed the best.

5. Results

5.1. Task 1: Text AER

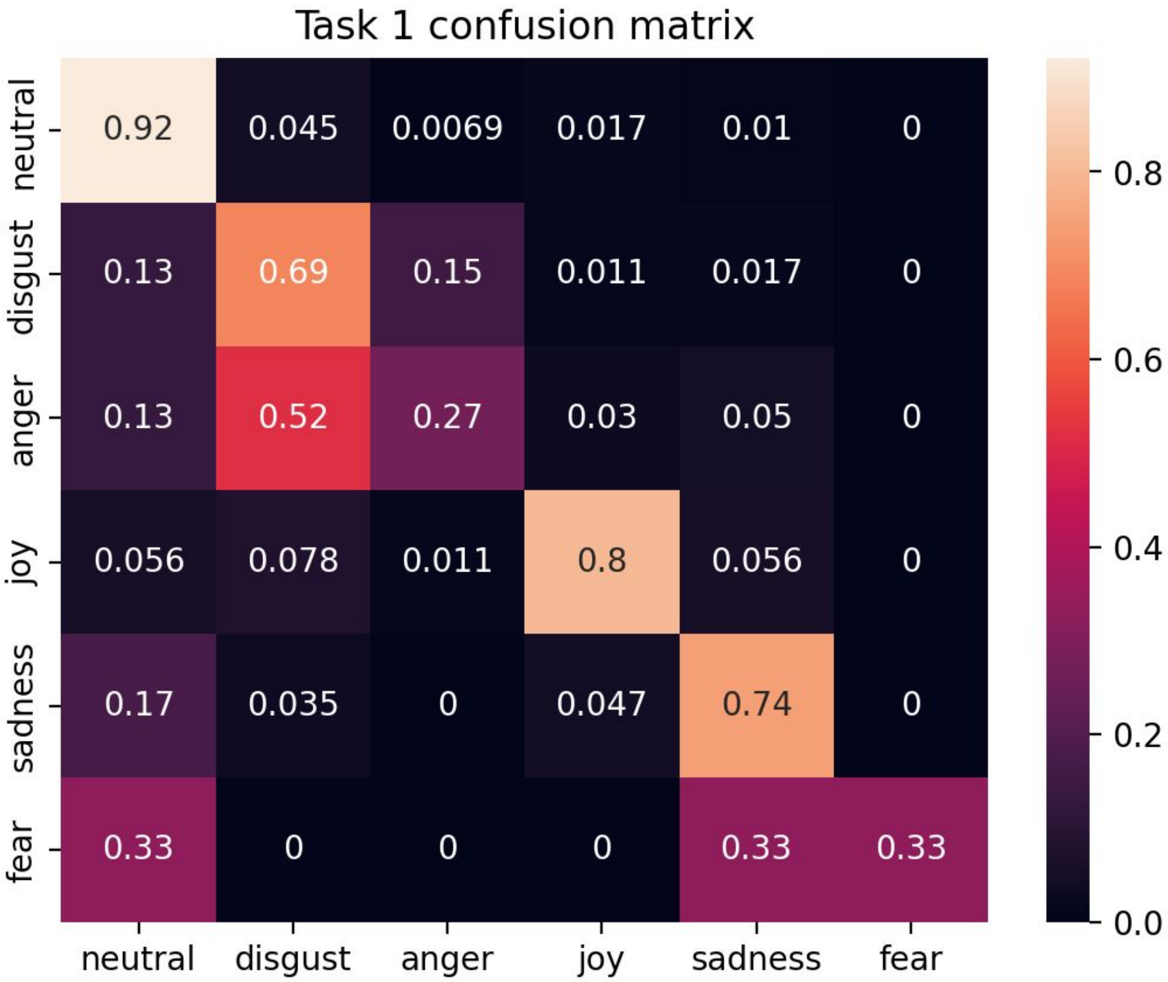
As mentioned earlier, we used the model with the best performance on the validation partition. To select this model, we focused on the one that achieved the best results in the F1-score, as this was the evaluation metric used to rank the participants. This model was obtained in the 12th training epoch. The official result over the F1-score (macro average) is shown in Table 1.

To gain a deeper understanding of our model's performance, we examined the confusion matrix for Task 1, shown in Figure 1. This analysis reveals the model's proficiency in accurately identifying each emotion category. For example, the model demonstrates a high precision of 0.92 in recognizing 'neutral' emotions. Nonetheless, there are significant misclassifications, such as the frequent confusion of ‘fear’ with ‘neutral’ and ‘sadness’.

Table 1
Task 1, official results over test data

Team	Metric	Score	Placement
CogniCIC	F1-score	0.657527	2nd

Figure 1: Confusion matrix for task 1



5.2. Task 2: Multimodal AER

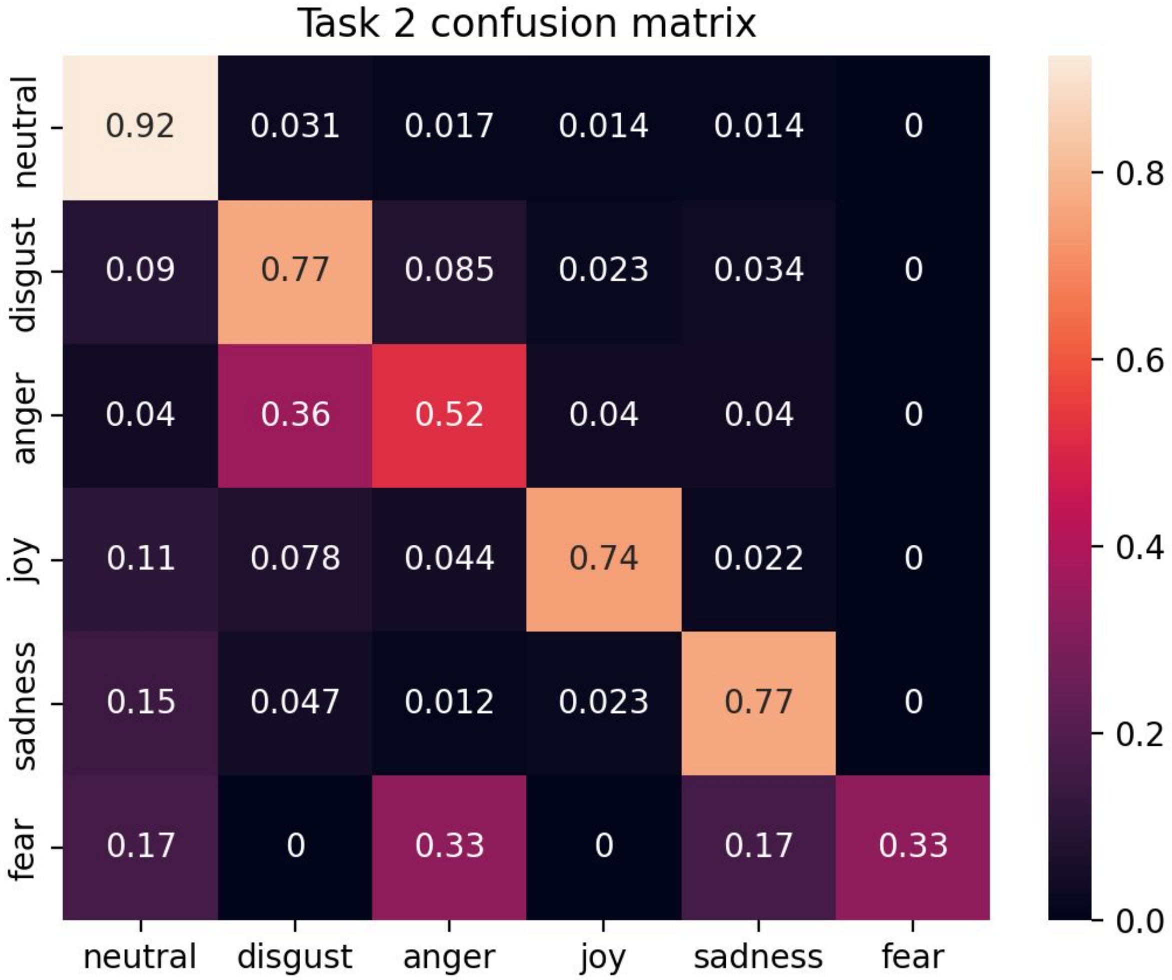
Even though several models were tested, the one achieving best performance on the validation set was the initial configuration, i.e., two plain multi-layer fully connected layers whose outputs are concatenated previous to the actual classification stage. The results yielded by such model on the testing set are shown in Table 2

To further evaluate the performance of our model, we analyzed the confusion matrix for Task 2, as depicted in Figure 2. The confusion matrix provides a detailed insight into the model’s ability to correctly classify each emotion category. For instance, the model shows high accuracy in recognizing ‘neutral’ emotions with a precision of 0.92. However, there are noticeable misclassifications, such as ‘fear’ being frequently confused with ‘anger’ and ‘sadness’. These insights are crucial for understanding the strengths and weaknesses of our model and guiding future improvements.

Table 2
Task 2, official results over test data

Team	Metric	Score	Placement
CogniCIC	F1-score	0.712259	3rd

Figure 2: Confusion matrix for task 2



6. Conclusions and future work

Automatic Emotion Recognition has proven to be a difficult task for Machine Learning algorithms. Under this context is that the EmoSpeech competition presents a crucial environment in which to test hypotheses and models. With this regard, we have presented two different proposals for the task in hand: a fine-tuned BETO LLM for pure text emotion recognition with no preprocessing, leveraging LLMs context capabilities; for multimodal recognition, a model consisting on two multi-layer fully connected networks, each one dedicated to analyze a different *mode* of data and just before the classification stage, the transformed values of each input are concatenated.

We believe our proposal to be a significant contribution to the field of single modal and multimodal emotion recognition, since both implementations need minimal configuration and preprocessing stages, allowing for easy development and understanding of the process. Additionally, our second proposal stands out for its simplicity both preprocessing architecture wise. Nevertheless, the importance of our contribution is by the positions obtained in the contest: second and third place for task 1 and task 2, respectively.

Given the promising results of our proposals, with second and third place finishes in tasks 1 and 2 respectively, we see several avenues for future work to build upon these findings. Firstly, exploring other pretrained models beyond BETO could provide additional insights and potentially improve performance. Enhanced multimodal integration methods, such as attention mechanisms or transformer-based architectures, could capture deeper interactions between audio and text features. Incorporating advanced data augmentation techniques could generate more robust models by increasing the diversity and size of the training dataset.

Expanding the scope of emotion recognition to include multiple languages and cultural contexts could enhance the generalizability of our models. This involves training and evaluating models on diverse datasets to ensure they can accurately recognize emotions across different populations. Developing and optimizing models for real-time emotion recognition applications, such as virtual assistants or customer service bots, is another significant direction, requiring low-latency and efficient processing of multimodal data.

Incorporating user-specific adaptations and personalization mechanisms could improve the accuracy of emotion recognition systems by accounting for individual differences in emotional expression. Additionally, interdisciplinary collaborations with experts in psychology and cognitive science could better inform the development of more sophisticated and accurate emotion recognition models.

By addressing these areas, we aim to further advance the field of emotion recognition, making models more robust, versatile, and applicable to a wider range of real-world scenarios.

Acknowledgments

The authors wish to thank the support of the Instituto Politécnico Nacional (COFAA, SIP-IPN, Grant SIP 20240610) and the Mexican Government (CONAHCyT, SNI).

References

[1] M. de Velasco, R. Justo, J. Antón, M. Carrilero, M. I. Torres, Emotion detection from speech and text., in: IberSPEECH, 2018, pp. 68–71.

[2] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, X. Zhao, Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects, Expert Systems with Applications 237 (2024) 121692. URL: https://www.sciencedirect.com/science/article/pii/S0957417423021942. doi:https://doi.org/10.1016/j.eswa.2023.121692.

[3] A. Dzedzickis, A. Kaklauskas, V. Bucinskas, Human emotion recognition: Review of sensors and methods, Sensors 20 (2020). URL: https://www.mdpi.com/1424-8220/20/3/592. doi:10.3390/s20030592.

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

[5] 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.

[6] S. K. Pandey, H. S. Shekhawat, S. R. M. Prasanna, Deep learning techniques for speech emotion

recognition: A review, in: 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), 2019, pp. 1–6. doi:10.1109/RADIOELEK.2019.8733432.

[7] M. A. Cardoso-Moreno, C. Macias, T. Alcantara, M. Soto, H. Calvo, C. Yañez-Marquez, Convolving emotions: A compact cnn for eeg-based emotion recognition, in: 2023 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2023, pp. 1472–1476.

[8] E. Lieskovská, M. Jakubec, R. Jarina, M. Chmulík, A review on speech emotion recognition using deep learning and attention mechanism, Electronics 10 (2021). URL: https://www.mdpi.com/2079-9292/10/10/1163. doi:10.3390/electronics10101163.

[9] D. Wang, X. Zhao, Affective video recommender systems: A survey, Frontiers in Neuroscience 16 (2022). URL: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2022.984404. doi:10.3389/fnins.2022.984404.

[10] S. Narayanan, P. G. Georgiou, Behavioral signal processing: Deriving human behavioral informatics from speech and language, Proceedings of the IEEE 101 (2013) 1203–1233. doi:10.1109/JPROC.2012.2236291.

[11] S. Jerritta, M. Murugappan, R. Nagarajan, K. Wan, Physiological signals based human emotion recognition: a review, in: 2011 IEEE 7th International Colloquium on Signal Processing and its Applications, 2011, pp. 410–415. doi:10.1109/CSPA.2011.5759912.

[12] A. Dzedzickis, A. Kaklauskas, V. Bucinskas, Human emotion recognition: Review of sensors and methods, Sensors 20 (2020). URL: https://www.mdpi.com/1424-8220/20/3/592. doi:10.3390/s20030592.

[13] N. Alswaidan, M. E. B. Menai, A survey of state-of-the-art approaches for emotion recognition in text, Knowledge and Information Systems 62 (2020) 2937–2987.

[14] P. Thakur, D. R. Shrivastava, A. DR, A review on text based emotion recognition system, International Journal of Advanced Trends in Computer Science and Engineering 7 (2018).

[15] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, E. Ambikairajah, A comprehensive review of speech emotion recognition systems, IEEE Access 9 (2021) 47795–47814. doi:10.1109/ACCESS.2021.3068045.

[16] M. Swain, A. Routray, P. Kabisatpathy, Databases, features and classifiers for speech emotion recognition: a review, International Journal of Speech Technology 21 (2018) 93–120.

[17] K. Sailunaz, M. Dhaliwal, J. Rokne, R. Alhajj, Emotion detection from text and speech: a survey, Social Network Analysis and Mining 8 (2018) 28.

[18] 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.

[19] S.-H. Park, B.-C. Bae, Y.-G. Cheong, Emotion recognition from text stories using an emotion embedding model, in: 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), 2020, pp. 579–583. doi:10.1109/BigComp48618.2020.00014.

[20] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1d & 2d cnn lstm networks, Biomedical signal processing and control 47 (2019) 312–323.

[21] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, S. Narayanan, Data augmentation using gans for speech emotion recognition., in: Interspeech, 2019, pp. 171–175.

[22] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[23] S. Yoon, S. Byun, K. Jung, Multimodal speech emotion recognition using audio and text, in: 2018 IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 112–118. doi:10.1109/SLT.2018.8639583.

[24] S. Bird, Nltk: the natural language toolkit, in: Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, 2006, pp. 69–72.

[25] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Language resources and evaluation

42 (2008) 335–359.

[26] D. Hazarika, S. Gorantla, S. Poria, R. Zimmermann, Self-attentive feature-level fusion for multimodal emotion detection, in: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2018, pp. 196–201. doi:10.1109/MIPR.2018.00043.

[27] D. Krishna, A. Patil, Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks., in: Interspeech, 2020, pp. 4243–4247.

[28] R. Pan, J. A. García-Díaz, M. 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.

[29] P. Ekman, Facial expressions of emotion: New findings, new questions, 1992.

[30] 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.

[31] Z. K. Abdul, A. K. Al-Talabani, Mel frequency cepstral coefficient and its applications: A review, IEEE Access 10 (2022) 122136–122158.

[32] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).