UAH-UVA in EmoSpeech-IberLEF2024: A Transfer Learning Approach for Emotion Recognition in Spanish Texts based on a Pre-trained DistilBERT Model

Andrea Chaves-Villota[1,*], Ana Jimenez[1] and Alfonso Bahillo[2]

[1]*Electronics Department, University of Alcala, E.P.S. Campus universitario s/n, E-28805, Alcalá de Henares (Madrid), Spain.*
[2]*University of Valladolid, Escuela Técnica Superior de Ingenieros de Telecomunicación, Campus Miguel Delibes, 47011, Valladolid, España*

Abstract

Emotion recognition is a key component in numerous domains, emphasizing its significance in understanding human behavior, enhancing communication technologies, and facilitating personalized user experiences. In this study, we present the methodology used in EmoSPeech 2024 Task to train two classification models capable of identifying five of Ekman's six basic emotions from Spanish text transcripts extracted from the Spanish MEA Corpus 2023 database (Task 1: Text Automatic Emotion Recognition). This methodology is developed from a transfer learning approach using a pre-trained model based on Distilbert's architecture. To handle the class imbalance of the dataset and to avoid a bias of the model towards the majority classes, it is proposed to use a technique based on class weighting, where a higher weight is given to the minority class (fear) and a lower weight to the majority class (neutral). Subsequently, the models' performance in classifying emotions is compared, where the weighted model outperforms the unweighted one with a f1-score of 0.63 as opposed to 0.61. Furthermore, we discuss our approach's strengths and weaknesses and share our understanding of the variables that influence its effectiveness. Our findings highlight the feasibility of developing emotion identification systems from voice transcriptions by using pre-trained models.

Keywords

Emotion Recognition, Spanish Text Classification, Distilbert, Transfer Learning

1. Introduction

Emotion recognition plays a fundamental role in the understanding of cognitive processes, human behaviors, and social dynamics faced by human beings in different facets of their lives. Nowadays, with the rise of artificial intelligence, the study of emotion recognition systems has received great attention from the scientific community [1], since they allow a better understanding of how emotions are expressed, experienced, and regulated across different cultures and contexts. This is also due to their wide range of practical applications which include developing therapies for mental health disorders, improving user experiences in areas like virtual assistants and educational tools, enhancing human-computer interaction through more intuitive and sympathetic technologies, and improving marketing strategies by understanding consumer emotions [2]. In addition, from emotion recognition systems, certain notions of collective behaviors can be discovered, bringing with them a collective benefit among groups of individuals and communities, for example, in education, these systems can enhance learning experiences by adapting the content and providing support for the learning process based on students' emotional states [3, 4]. Therefore, research in emotion recognition systems allows a better understanding of human nature and improves the design of technologies and interventions aimed at promoting mental health, and social well-being [5].

To develop more robust models for emotion recognition, the latest research focuses on training multimodal models that take into account different aspects of human behavior, including facial expressions, body language, speech patterns, and physiological responses [6, 7, 8, 9]. In particular, voice patterns

IberLEF 2024, September 2024, Valladolid, Spain

[*] Corresponding author.

andrea.chaves@uah.es (A. Chaves-Villota); ana.jimenez@uah.es (A. Jimenez); alfonso.bahillo@uva.es (A. Bahillo)
0000-0002-9272-9426 (A. Chaves-Villota); 0000-0003-0713-0054 (A. Jimenez); 0000-0003-3370-33388 (A. Bahillo)

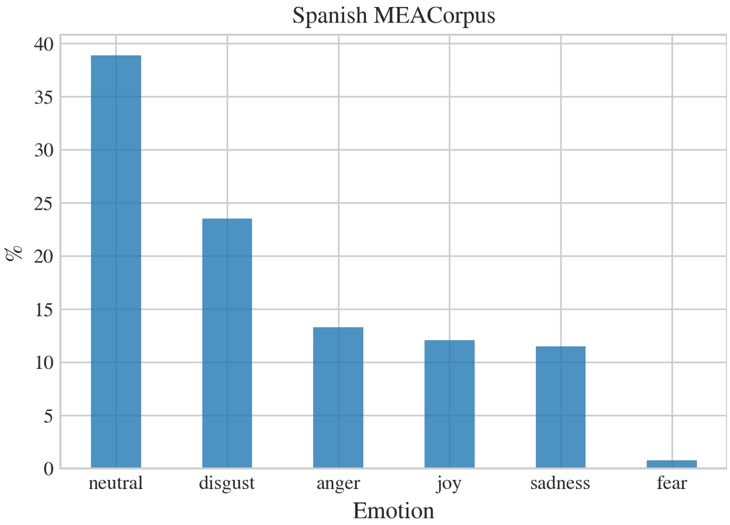


Figure 1: Distribution of Spanish MEA Corpus 2023

provide valuable information for emotion recognition, such as text and its extracted features like words or phrases, as well as certain linguistic features such as adjectives, adverbs, and intensifiers that can be associated with a particular emotion [10]. Hence, it is important to use different data sources to support a model's robust classification performance.

In this study, we use a particular machine learning technique known as transfer learning, in which a pre-trained model is adapted to perform tasks related to text classification. Thus, the main contribution of the paper refers to the methodology for training the model in emotion classification from texts using Spanish MEA Corpus 2023 (Task 1: Text AER). This methodology is based on a transfer learning approach using two models, a weighted model that weights classes in the training stage to deal with the problem of unbalanced data, and an unweighted one, which takes an equal weighting for all classes. The paper is organized as follows: section II gives a brief description of the corpus, models, and techniques used, section III describes the main results found, and finally, we discuss the conclusions in Section IV.

2. Materials and Methods

2.1. Dataset

To fine-tune the pre-trained model, we make use of the transcriptions provided by the multimodal Spanish MEA Corpus 2023 dataset [11], made available by the EmoSPeech 2024 Task [12, 13] which consists of approximately 13 hours of audio segments extracted from different Spanish YouTube channels, including political, sports and entertainment topics. Thus, the dataset provided comprises audio segments labeled with five of Ekman's basic emotions: disgust, anger, joy, sadness, fear, and neutral one [14]. Table 1 and Figure 1 show the distribution of the training data categorized by each emotion, it can be seen that the target classes are not balanced. Especially, there is a significantly different proportion of the emotion fear represented by only 00.76% compared to the rest of the classes. Likewise, the samples of the neutral state are presented in greater quantity with 38.86% of the total dataset, followed by 23.50% for the emotion disgust. The amount of samples for anger, joy and sadness are approximately balanced between them.

Unbalanced datasets are a common problem faced by emotion recognition systems [15]. To address this challenge, this paper proposes a model that weights the classes during the training stage. This approach aims to prevent the model from being biased towards the majority class (neutral) while ensuring adequate performance in classifying the minority class (fear) as well.

Table 1
Dataset distribution

Emotion	No.	%
neutral	1166	38.86
disgust	705	23.50
anger	399	13.30
joy	362	12.06
sadness	345	11.50
fear	23	00.76
Total	3000	100%

2.2. Training based on Pre-trained Distilbert Architecture and Class Weighting Technique

In this paper, we propose the use of a technique based on a transfer learning approach for emotion classification from Spanish texts. As shown in Fig. 2, first, it is necessary to define a proper pre-trained model that solves a classification problem (Task 1: tweet classification). This pre-trained model will serve as a starting point in the new classification task to be developed (Task 2: emotion classification), thus the selected pre-trained model must develop a classification task related to the new problem. In addition, a class weighting technique is proposed to deal with the unbalanced dataset problem. These phases are explained in the following subsections.

2.2.1. Pre-trained Distilbert Architecture

We selected a Distilbert pre-trained model to obtain better results in the emotion classification task (Task 1 in Fig.2) [16]. Its architecture is based on the compressed version of BERT (Bidirectional Encoder Representations from Transformers) known as *Distilbert* that presents similar performances in the development of NLP tasks such as text classification, question answering, and named entity recognition. Its main advantage is that it can be used with reduced computational resources, its inference time is faster and it uses fewer parameters (up to 40% less), making it a more practical model for developing real-world applications [17]. This pre-trained model was selected since it solutions a classification problem with certain patterns and features similar to the emotion classification task to be achieved with the new target model (Task 2 in Fig.2), in this way, an appropriate knowledge transfer between both architectures could be developed. Specifically, we selected the *Distilbert base finetuned with Spanish tweets*, whose objective is to classify tweets in Spanish into three categories, positive, negative, and neutral. Taking advantage of the model's classification task that classifies Spanish text data, they are additionally categorized into discrete classes.

The main difference between the model architectures is centered on the *head*, which for the source model corresponds to an output layer capable of categorizing Spanish tweets into three different labels (Positive, Negative, Neutral), while the *new head* of the target model is randomly initialized and trained to categorize into the 6 new classes (Disgust, Anger, Joy, Sadness, Fear, Neutral). The remaining hyperparameters are kept constant to preserve the learned features and prevent them from being updated during training. This approach maximizes the use of the model's existing knowledge. These hyperparameters are detailed in Table 2.

2.2.2. Class Weighting Technique

Furthermore, to deal with the class imbalance problem presented by the dataset distribution, avoid biasing the model to the majority class and thus obtain a possible more robust classification. We propose the evaluation of a simple and common technique during training that performs class weighting, penalizing the minority class by setting a higher weight and at the same time reducing the weighting for the majority classes. The class weighting vector $\boldsymbol{w}$ was adjusted in a range of $\boldsymbol{w_1} \in [0.5, 2]$ according

Table 2
Hyperparameters of model

Hyperparameter	Value
No. Transformers Layers	6
Attention heads	12
Transformer Activation	Gaussian Error Linear Unit (GELU)
Dropout	0.1
Optimizer	AdamW
Loss	Cross Entropy
Epochs	12
Batch size	8
Learning rate	4e-5

Table 3
Class weighting

Emotion	$\%$	w_i
neutral	38.86	0.5
disgust	23.50	1
anger	13.30	1.5
joy	12.06	1.5
sadness	11.50	1.5
fear	00.76	2

to the percentage of data belonging to each class i under the convention given by (1). It is important to highlight that the selected weights $\boldsymbol{w}$ were approximated depending on the data per label. However, it is important to highlight that these values can be adjusted using optimization techniques that could improve the model performance.

$$w_i = \begin{cases} 2, & \% < 5 \\ 1.5, & 5 < \% < 20 \\ 1, & 20 < \% < 35 \\ 0.5, & \% > 35 \end{cases} \tag{1}$$

The training of both weighted and unweighted models was run on a Tesla T4 GPU with 15 GBs RAM provided by the Google Colaboratory cloud service, using the Simple Transformers package based on the Transformers library by HuggingFace.

3. Results and discussion

3.1. Convergence analysis

Figure 3 shows the behavior of the loss in the training phase of the two models for each step, i.e. each time the batch training is completed. It is possible to appreciate that the two models manage to converge in approximately 3000 steps to a loss of approximately 0. The two models present a high variability during the training, more evident for the weighted model in comparison with training achieved by the unweighted model. It can also be noted that these variances decrease at approximately 5000 steps, resulting in values close to zero.

3.2. Model performances

Figure 4a and Figure 4b show the confusion matrixes accomplished in the test stage for the unweighted and weighted models, respectively. In both cases, it can be seen that it is achieved a better classification

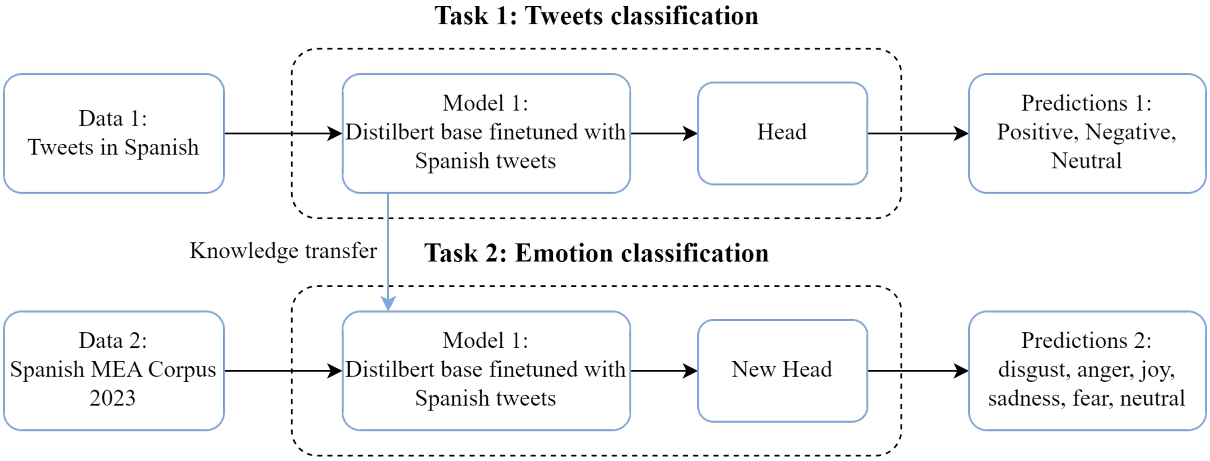

Task 1: Tweets classification
Data 1:
Tweets in Spanish
Model 1:
Distilbert base finetuned with Spanish tweets
Head
Predictions 1:
Positive, Negative, Neutral
Knowledge transfer
Task 2: Emotion classification
Data 2:
Spanish MEA Corpus 2023
Model 1:
Distilbert base finetuned with Spanish tweets
New Head
Predictions 2:
disgust, anger, joy, sadness, fear, neutral

Figure 2: Methodology based on transfer learning for emotion classification.

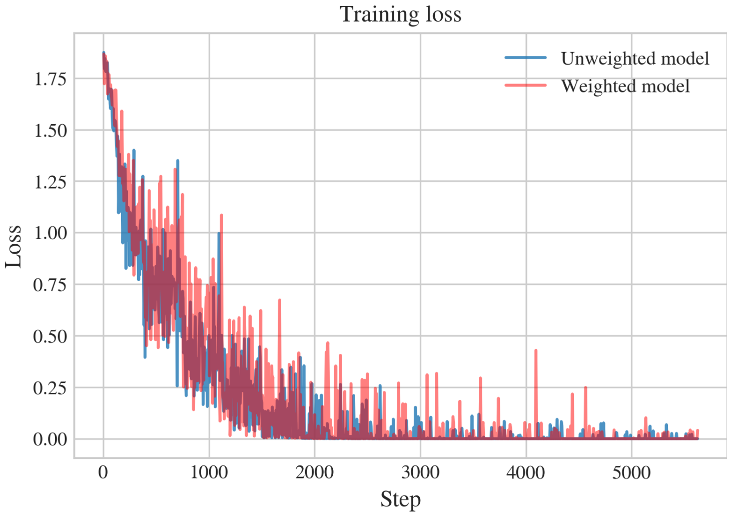
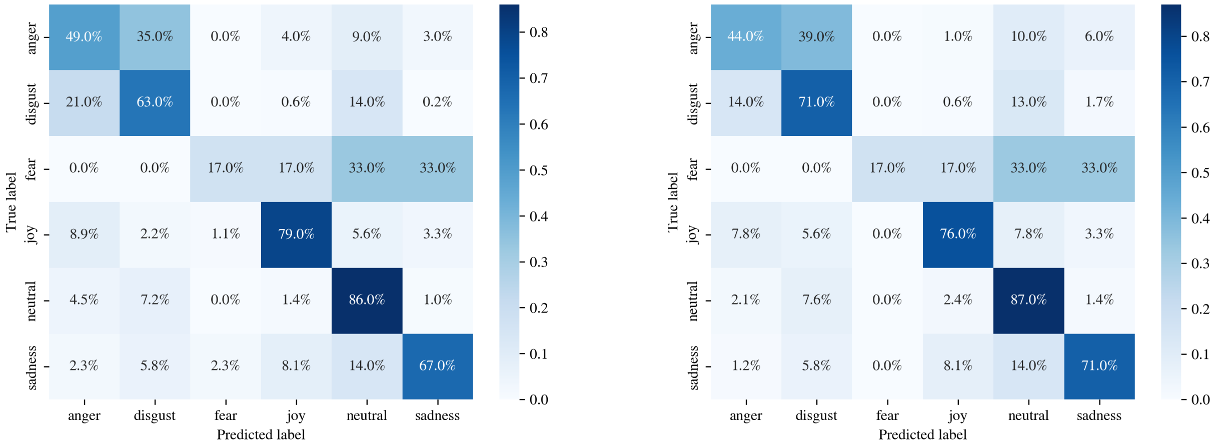


Figure 3: Loss convergence in model training

of the neutral class concerning the others and that the poorest performance is for the fear label. This performance is expected due to the unbalanced nature of the data. The weighted model achieves better classification performance for the neutral, sadness, and disgust classes compared to the unweighted one. However, the unweighted model obtains better metrics for the anger and joy classes. It can also be verified that both models mispredict a high percentage of the anger emotion with disgust, 39%, and 35% for weighted and unweighted models respectively. Additionally, they present the same behavior in the classification of fear, misclassifying it with the emotions joy, neutral, and sadness.

Overall, both models score high on the main diagonal, except for the fear label, as is desirable in the evaluation of the classification task, since it represents the instances where the predicted emotions match the true ones. According to this, it could be inferred that especially for the fear emotion the weighted model does not show a sufficient difference in the classification to the unweighted model, hence we propose the study of an optimization of the weighted vector w, which could include a higher weighting for this class and verify a possible improvement in the performance of the model.

Table 4 reports the metrics evaluated with the test data by the two models. In general, the weighted one achieves higher scores in macro f1-score and recall with 0.63 and 0.61, respectively, in contrast to the unweighted model that achieved an f1-score of 0.61 and recall of 0.60. Furthermore, considering that the dataset is unbalanced, the weighted model is considered to have a more robust classification performance. This is also highlighted taking into consideration that they used the same computational



(a) Unweighted model

(b) Weighted model

Figure 4: Confusion matrixes

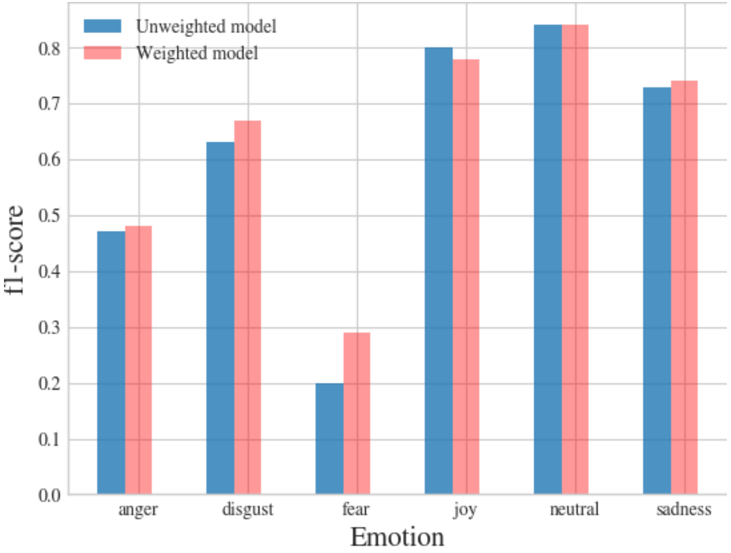


Figure 5: Comparison of F1-score between models for each emotion

resources.

To have some insights into emotion specific performance with test data. We show f1-score achieved for each class by models in Figure 5. We can see that the weighted model achieves better results in most cases, except for the emotion joy. In general, there are not high differences in the f1 scores obtained by the models for each emotion. Nevertheless, it can be identified that the models perform better in classifying the emotions joy, neutral, sadness, and disgust with an f1-score above 0.6. This is in contrast to the identification of the emotions anger and fear. This could not necessarily be related to the amount of data per class, since in the training phase, the emotions joy and sadness were covered in lower percentages, (12.06% and 11.50%) with respect to disgust (23.50%) and even so, the f1-score obtained is higher than 0.7 in both cases. It is evident that the two models face a high challenge when classifying the emotion fear, for which f1-scores is below 0.3 in both cases, being even lower for the unweighted model. Hence, it is proposed to make use of class balancing techniques such as oversampling of the minority class or undersampling of the majority ones, as well as to evaluate synthetic data generation techniques that could help to improve the robustness of the model.

Table 4
F1-score and recall with test data

Emotion	Unweighted Model		Weighted Model	
	recall	F1	recall	F1
neutral	0.86	0.84	0.87	0.84
disgust	0.63	0.63	0.71	0.67
anger	0.49	0.47	0.44	0.48
joy	0.79	0.80	0.76	0.78
sadness	0.67	0.73	0.71	0.74
fear	0.17	0.20	0.17	0.29
Macro avg.	0.60	0.61	**0.61**	**0.63**

Table 5
Class weighting

Emotion	%	w_i
neutral	38.86	0.4
disgust	23.50	0.7
anger	13.30	1.2
joy	12.06	1.3
sadness	11.50	1.4
fear	00.76	21.7

Table 6
F1-score and recall achieved by New Weighted Model

Emotion	Weighted Model	
	recall	F1
neutral	0.82	0.85
disgust	0.63	0.63
anger	0.45	0.45
joy	0.83	0.79
sadness	0.78	0.77
fear	1.00	0.29
Macro avg.	0.75	0.63

3.2.1. Out-of-competition results

In addition, out-of-competition we evaluate another common class weighting, where the vector $\boldsymbol{w}$ is set inversely proportional to the frequency of classes in the data, according to (2). Where n refers to the total number of samples, n_c corresponds to the total number of classes (emotions) and n_i is the total number of samples belonging to the class i. The resulting weights for each class are shown in Table 5.

$$w_i = \frac{n}{n_c n_i} \tag{2}$$

Fig. 6 shows the confusion matrix resulting with test data (not seen in the training phase). It is possible to appreciate that according to *recall* (main diagonal), a higher weighting in the *fear* class improves the model performance in its categorisation. However, it is necessary to highlight that as well as improving the classification, the prediction is also affected in the recognition of other emotions such as *Anger* and *Disgust*, in relation to the weighted model I (See Fig. 7). The recall and F1-score results achieved by this model for each emotion and their respective averages are shown in Table 6. Notice clearly that in relation to the weighted model I, this new class weighting improves the classification in the rate of true positives, especially for the classes *anger*, *fear*, *joy*, and *sadness*.

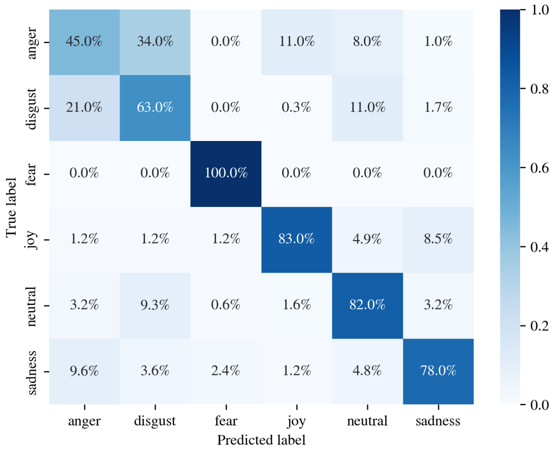


Figure 6: Confusion Matrix of the New Weighted Model

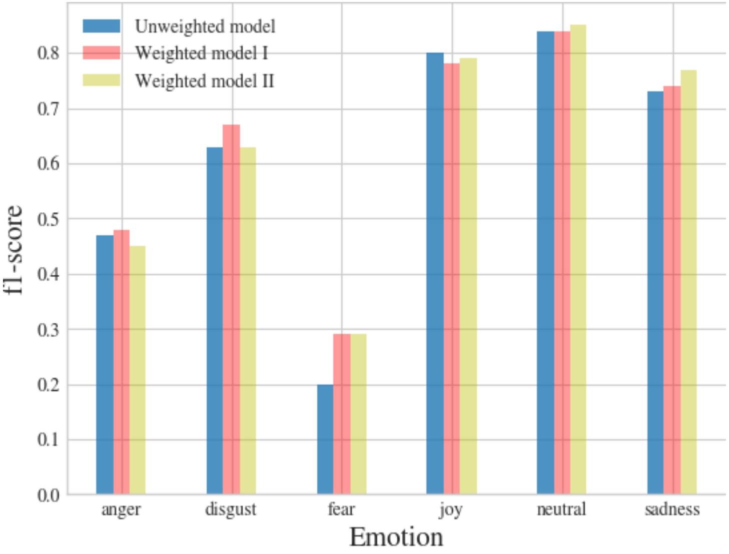


Figure 7: Comparison of F1-score achieved by models for each emotion

We can observe that according to the results achieved by the two models with class weighting, the use of balancing techniques such as the one used in this study, allows us to achieve a better result in the task of emotion classification, highlighting that the computational cost both in the training and test phases used by the weighted models does not present a major difference with respect to the unweighted one. We further emphasise the importance of using optimization techniques to find optimal weights that could improve the model performance in this task.

4. Conclusion

In this work, we developed a methodology based on transfer learning to explore the advantage of employing pre-trained models for emotion recognition from speech transcriptions. The Spanish MEA Corpus 2023 was used as a benchmark dataset for the training and test phases. We propose evaluate the performance of two models, a weighted that uses a technique based on class weighting to address the problem of emotion imbalance, to avoid biasing the model towards the majority class (neutral), and the unweighted one that does not make an adjustment of weights between classes. Through experimentation, we observed favorable results with F1 scores of 0.63 and 0.61 for weighted and unweighted models, respectively. Even though the two models exhibit comparative behaviors, the present research determined key factors in the use of the class-weighting technique that could yield potential improvements in handling the emotion imbalance problems. These findings highlight the potential of leveraging pre-trained models as a viable approach for emotion recognition from text. Moving forward, efforts should focus on refining and optimizing the weights for the weighted model to enhance emotion recognition prediction. Additionally, exploring multimodal methods that incorporate other input sources such as speech patterns would offer alternative ways to improve model performance.

Acknowledgments

This work was supported by the FrailAlert (SBPLY/21/180501/000216, co-financing from both the Junta de Comunidades de Castilla-La Mancha and the European Union through the European Regional Development Fund); and ActiTracker (TED2021-130867B-I00) and INDRI (PID2021-122642OB-C41 /AEI/ 10.13039/501100011033/FEDER, UE) .

References

[1] E. Cambria, D. Das, S. Bandyopadhyay, A. Feraco, Affective computing and sentiment analysis, A practical guide to sentiment analysis (2017) 1–10.

[2] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, M. R. Wróbel, Emotion Recognition and Its Applications, Springer International Publishing, Cham, 2014, pp. 51–62. URL: https://doi.org/10.1007/978-3-319-08491-6_5. doi:10.1007/978-3-319-08491-6_5.

[3] W. Wang, K. Xu, H. Niu, X. Miao, Emotion recognition of students based on facial expressions in online education based on the perspective of computer simulation, Complexity 2020 (2020) 1–9.

[4] D. Yang, A. Alsadoon, P. C. Prasad, A. K. Singh, A. Elchouemi, An emotion recognition model based on facial recognition in virtual learning environment, Procedia Computer Science 125 (2018) 2–10.

[5] S. G. Koolagudi, K. S. Rao, Emotion recognition from speech: a review, International journal of speech technology 15 (2012) 99–117.

[6] M. Ragot, N. Martin, S. Em, N. Pallamin, J.-M. Diverrez, Emotion recognition using physiological signals: laboratory vs. wearable sensors, in: Advances in Human Factors in Wearable Technologies and Game Design: Proceedings of the AHFE 2017 International Conference on Advances in Human Factors and Wearable Technologies, July 17-21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8, Springer, 2018, pp. 15–22.

[7] N. Sebe, I. Cohen, T. S. Huang, Multimodal emotion recognition, in: Handbook of pattern recognition and computer vision, World Scientific, 2005, pp. 387–409.

[8] A. B. Ingale, D. Chaudhari, Speech emotion recognition, International Journal of Soft Computing and Engineering (IJSCE) 2 (2012) 235–238.

[9] P. Tarnowski, M. Kołodziej, A. Majkowski, R. J. Rak, Emotion recognition using facial expressions, Procedia Computer Science 108 (2017) 1175–1184.

[10] V. Ramanarayanan, A. C. Lammert, H. P. Rowe, T. F. Quatieri, J. R. Green, Speech as a biomarker: Opportunities, interpretability, and challenges, Perspectives of the ASHA Special Interest Groups 7 (2022) 276–283.

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

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

[13] 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.

[14] P. Ekman, et al., Basic emotions, Handbook of cognition and emotion 98 (1999) 16.

[15] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, N. Amir, The automatic recognition of emotions in speech, Springer, 2011.

[16] F. Perez-Sorrosal, Distilbert base uncased fine-tuned with spanish tweets, https://huggingface.co/francisco-perez-sorrosal/distilbert-base-uncased-finetuned-with-spanish-tweets-clf-cleaned-ds, 2023.

[17] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).