Speech emotion recognition in Spanish TV Debates

Irune Zubiaga, Raquel Justo, Mikel De Velasco, M. Inés Torres

Universidad del Pais Vasco UPV/EHU. Sarriena s/n. 48940 Leioa. Spain.

irune.zubiaga@ehu.eus, raquel.justo@ehu.eus, mikel.develasco@ehu.eus, manes.torres@ehu.eus

Abstract

Emotion recognition from speech is an active field of study that can help build more natural human–machine interaction systems. Even though the advancement of deep learning technology has brought improvements in this task, it is still a very challenging field. For instance, when considering real life scenarios, things such as tendency toward neutrality or the ambiguous definition of emotion can make labeling a difficult task causing the data-set to be severally imbalanced and not very representative.

In this work we considered a real life scenario to carry out a series of emotion classification experiments. Specifically, we worked with a labeled corpus consisting of a set of audios from Spanish TV debates and their respective transcriptions. First, an analysis of the emotional information within the corpus was conducted. Then different data representations were analyzed as to choose the best one for our task; Spectrograms and UniSpeech-SAT were used for audio representation and DistilBERT for text representation. As a final step, Multimodal Machine Learning was used with the aim of improving the obtained classification results by combining acoustic and textual information.

Index Terms: Acoustic Signal, Textual Information, Multimodal Machine Learning, Emotion Recognition

1. Introduction

The automatic detection of emotion from speech and language has gained popularity in recent years due to its capability to promote natural human-machine interaction, better comprehension of human interventions, etc. In order to be useful, the emotion detection systems need to work properly in real life scenarios, where emotions are not very extreme and only subtle expressions can be appreciated. Most of existing systems and approaches deal with emotions simulated by professional actors, leading to poor performances when trying to extrapolate to more realistic tasks.

Emotional responses result in changes in facial expression, in vocal expression, speaking style, in the way the language is used as well as in changes in physiological signals, such as the electroencephalographic signals (EEG) or galvanic skin responses, among others (GSR) [1]. The information provided by each signal can contribute to the selection of different features, which can be complementary. In this work, we focus on speech and language as information sources that can help in the automatic identification of emotions. Moreover, we will also explore whether the two sources can contribute together to a better system performance.

The six basic emotions defined by Eckman [2] (anger, surprise, disgust, enjoyment, fear, and sadness) can be represented by facial expressions that typically characterize these emotions [3]. However, spontaneous emotions that can be perceived from speech or language, are more varied and complex. Only a small set of complex and compound emotions [4] can be found in real scenarios [5, 6], and this subset is strongly dependent on the task. Therefore, a set of categories including the emotions that arise in each specific task has to be defined, according to perception experiments. However, this process is expensive and time consuming. In this work, we deal with a real life scenario; speech gathered from TV debates was considered to train an automatic emotion detection system.

For supervised learning, researchers need a ground truth to be used as a reference for automatic emotion identification. Usually, human annotators establish their own perception of the emotional data as the ground truth. So, in addition to being expensive and time consuming, these perceptual experiments also add subjectivity and complexity to the already complex and, to some extent, subjective emotional constructions, mainly in speech processing. In this work we carried out an annotation procedure based on crowdsourcing, that tries to gather the diversity from a bigger set of annotators [7].

As an alternative to working with categorical emotions, a number of researchers [8, 9] proposed a dimensional representation [10] of the emotional space. Thus, each affective state is represented by a point in a two-dimensional space, namely Valence and Arousal. This two dimensional model has been replaced by a three dimensional model, according to some authors work [11], including Dominance as a third dimension, to represent the complete range of human responses. This work employs both approaches to analyze emotional information.

The contribution of this work lies on the idea of using transformer-based representations for acoustic and textual information in a multimodal environment, in order to detect emotional information perceived in a real scenario. The achieved results show that multimodality is mainly helpful when considering Valence dimension and the categorical emotional information.

The manuscript is organized as follows: Section 2 deals with the specific task and corpus and the employed modelization of emotions. Section 3 describes the different features and methodologies employed to carry out the experiments and Section 4 summarizes the achieved results. Finally, Section 5 underlines extracted conclusions and future work.

2. Task and Corpus

In this work a set of human–human conversations was gathered from TV debates. Specifically, the Spanish TV program “La Sexta Noche” was used. In this weekly broadcast show, news about hot topics from the week are addressed by social and political debate panels led by two moderators. A very wide range of talk-show guests (politicians, journalists, etc.) analyze social topics from their perspectives. Given that the topics under discussion are usually controversial, emotionally rich interactions can be expected. However, the participants are used to speaking in public so they do not lose control of the situation. Thus, even if they might overreact sometimes, this is a real scenario, where emotions are subtle. The spontaneity in this situation is vastly different from scenarios with acted emotions, as shown in [15]. The selected programs were broadcast during the electoral campaign of the Spanish general elections in December 2015. Table 1 shows a small excerpt of a dialogue taken from the TV Debate corpus.

Table 1: *Emotionally rich excerpt from the corpus in which three talk-show guests debate about politics. The excerpt is shown in Spanish (the original language) and in English.*

	Spanish
Speaker 1:	Sí, sí, efectivamente, efectivamente, cuatro que optan a ganar estas elecciones
Speaker 2:	Por eso
Speaker 1:	Pero hay muchos más partidos
Speaker 3:	Van a ganar, yo creo que un tanto a dos
Speaker 1:	Bueno, están en un pañuelo
	English
Speaker 1:	Yes, yes, indeed, indeed, four who opt to win these elections
Speaker 2:	That is why
Speaker 1:	But there are many more parties
Speaker 2:	They are going to win, I think that one to two
Speaker 1:	Well, they are too close to call

The whole audio signal associated to an specific show, was separated according to the interventions of the speakers. This way, an audio file was achieved for each speaker intervention. The example of Table 1 would correspond to 5 different audio files, associated to Speaker 1, Speaker 2, Speaker 1, Speaker 3, Speaker 1.

In contrast with previous research [12] in which audio segments between 2 and 5 seconds were considered, in this work we used the full audio of each speaker intervention without slicing it since we considered this could be a more representative unit for emotional recognition.The audio files in which speakers could not be told apart and the ones that were not related to the debates (music, ads, etc.) were removed from the corpus.

The diarization and transcription were carried out manually within the framework of the Affective Multimedia Analytics with Inclusive and Natural Communication (AMIC) project [13]. The labeling was done using crowd annotation by 5 annotators. This procedure provided a set of 2964 audio files from 2 to 20 seconds long. Their respective transcriptions were also gathered. Said transcriptions have a length between 1 and 86 words, with a mean sentence length of 33 words and a mode of 37.

Regarding speaker features, the gender distribution was 24.8% female and 75.2% male, with a total of 88 speakers.

Table 2: *Number of audio files for each VAD category.*

			Audio nº
V	Negative	$(v \leqslant 0.4)$	669
	Neutral	$(0.4 < v < 0.6)$	1597
	Positive	$(v \geqslant 0.6)$	698
A	Neutral	$(a \leqslant 0.15)$	2113
	Excited	$(a > 0.15)$	851
D	Intimidated	$(d \leqslant 0.75)$	1533
	Dominant	$(d > 0.75)$	1431

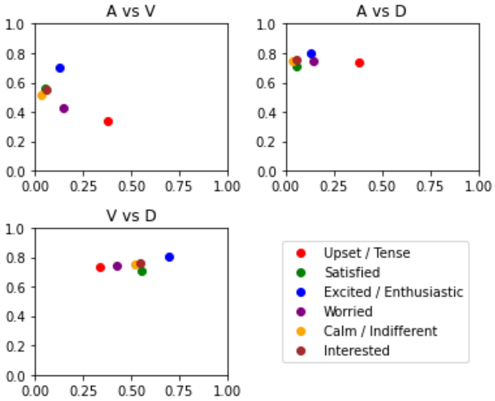

Figure 1: *Representation of the mean value of each emotion in the dimensional model.*

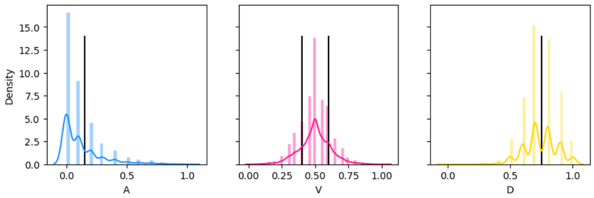

Figure 2: *Distribution of VAD values in the data-set*

.

2.1. Data-set for the VAD model

The Valence-Arousal-Dominance (VAD) model, also known as Pleasure-Arousal-Dominance (PAD), is a three-dimensional model that was introduced by Mehrabian and Russell in [14]. Mehrabian and Russell propose three independent dimensions for emotional representation; Valence (Pleasure), which ranges from displeasure to pleasure and expresses the pleasant or unpleasant feeling about something, Arousal, that ranges from nonarousal to arousal and represents the level of affective activation , and Dominance, which shows the level of control or influence on events and surroundings and goes from submissiveness to dominance.

To label Valence, Arousal and Dominance all 5 annotators were asked to answer the following set of questions for each intervention:

¿How do you perceive the speaker?

- Excited (1)
- Slightly excited (0.5)
- Neutral (0)

His emotional state is:

- Positive (1)
- Slightly positive (0.75)
- Neutral (0.5)
- Slightly negative (0.25)
- Negative (0)

¿How do you perceive the speaker in relation to the situation they are in?

- Rather dominant / Controlling the situation (1)

Table 3: *Number of audio files for each emotion.*

Emotion	**Audio nº**
Upset / Tense	361
Satisfied	221
Excited / Enthusiastic	27
Suprised	2
Worried	92
Calm / Indiferent	643
Bored	0
Interested	179
Total	1525

- Neither dominates the situation nor is intimidated (0.5)
- Rather cowed / Defensive (0)

These qualitative answers were encoded with the values in parentheses. Then, the mean of each set of labels (each set consisting of the answers of the 5 annotators for the intervention) was computed in order to have a single label for each intervention (ex. Labels of intervention 1 = positive, negative, neutral, neutral, negative = 1, 0, 0.5, 0.5, 0 = 0.4).

Our first approach was to carry out a set of regression experiments but the task was too complex and the obtained results were not satisfying. As a consequence, we decided to simplify the task by discretizing the data. We used Figure 2 as a guideline to choose the threshold values for each class, which are represented in the figure by black vertical lines. This way we were left with the classes shown in Table 2. We have **2964** samples for this task.

2.2. Data-set for the Categorical Emotion Model

To label categorical emotions annotators were to choose an emotion from Table 3 that, in their opinion, better suited the speakers state. Because of the perception of emotion being very ambiguous, for the emotion recognition task we only selected samples in which %60 of the annotators agreed in an emotional label with the goal of reducing noise in the data-set. The number of audio samples which belong to each emotion after applying this filter is shown in Table 3. As can be seen, there is not enough data regarding some of the classes for the model to learn a representation. Figure 1 presents the distribution of the emotions in our task within the dimensional emotional space, spanned by Valence, Arousal and Dominance. As seen there, when representing emotions in the VAD space some of them are difficult to tell apart.

Taking into account these facts we chose to try to discriminate between three different emotions: **Calm**, **Upset/Tense** and **Worried**. Even though *Excited/Enthusiastic* seems quite distinguishable from other emotions we did not work with it as a consequence of having very little data from this class (27 samples). The rest of samples were dismissed since merging them with the Calm class (the one they are closer to in the VAD space) would make the class imbalance even bigger than it already is (1:4:7). This way, we are left with **1096** samples for the emotion recognition task.

3. Experimental Setup

Both acoustic based and text based systems were built and trained with the aforementioned training corpus. In all of the cases 10 fold cross validation was used for validation.

3.1. Acoustic information

To analyze acoustic data Mel Spectrograms and the UniSpeech-SAT model were used.

3.1.1. Mel Spectrograms

The Mel Spectrogram is a spectrogram where the frequencies are converted to the Mel Scale [15], this being a perceptual scale of pitches judged by listeners to be equal in distance from one another. Mel Spectrograms have been proved to be a good audio representation for several tasks including emotion recognition [16].

Our first approach to carry out the emotion classification task was to use Mel Spectrogram representations of each audio file as an input to a Deep Convolutional Neural Network (DCNN). Zero padding was used on the spectrograms for them to have the same length, the achieved shape being 128x625. The network consisted of three convolutional layers with 3, 5 and 10 filters and three linear layers with 70, 40 and n neurons and ReLu as the activation function, n being the number of classes in each task. The model was trained for 300 epochs with Adam as the optimizer a learning rate of 1e-4 and a batch size of 16.

3.1.2. UniSpeech-SAT

Another outlook to deal with this task was to use speech representation models such as Wav2vec, Hubert, WavLM and UniSpeech-SAT. The best outcome was achieved when using the UniSpeech-SAT model architecture, specifically, when working with the *microsoft/unispeech-sat-large* [17] pre-trained model. This being the case, we will only focus in the results that were obtained with this setting. The Universal Speech Representation Learning with Speaker Aware Pre-Training model (UniSpeech-Sat) [18] performs specially well on speaker verification, speaker identification, and speaker diarization tasks. UniSpeech-SAT has been pre-trained on 16kHz sampled speech audio with utterance and speaker contrastive loss. The model is pre-trained on 94k hours of public English audio data; 60K hours of Libri-Light [19], 10K hours of GigaSpeech [20] and 24K hours of VoxPopuli [21].

To carry out the classification experiments, we froze the UniSpeech-SAT model and added a 1024 dimensional and a n dimensional linear layer to the last hidden layer, n being the number of classes we want to predict in each case. After that, the model was trained for 80 epochs using Adam as an optimizer and with a batch size of 8 and learning rate of 5e-5.

3.2. Textual information

To work with textual information the DistilBert [22] model was used. DistilBERT is a small, fast, cheap and light Transformer model. By leveraging knowledge distillation during the pre-training phase, the reduction of BERTs size by 40 % is achieved along with the model running 60% faster while preserving 97% of its language understanding capabilities.

We used the *CenIA/distillbert-base-spanish-uncased* pre-trained model. Said model is the *distilbert-base-uncased* model trained in *The Large Spanish Corpus* [23], which is a compilation of 15 unlabelled Spanish corpora spanning Wikipedia to European parliament notes.

Then the model was fine-tuned for two epochs using Adam as an optimizer with a batch size of 8 and a learning rate of 3e-5.

distilbert-base-uncased consists of 6 layers of transformers block with a hidden size of 768 and 12 self-attention heads and has a total of 66M trainable parameters.

3.3. Multimodal Machine Learning

Multimodal machine learning (MMML) is a multi-disciplinary research field that addresses some of the original goals of artificial intelligence by building models that can process and relate information from multiple modalities, including linguistic, acoustic and visual information. This approach outperforms single modal AI in many real-world problems [24] [25].

In this research we created a model that takes textual and acoustic information as an input with the aim of improving emotion classification results.The model architecture is shown in Figure 3. In this model we concatenated the logits from the audio model described in Section 3.1.2 and the text model described in Section 3.2 and used them as an input to a Neural Network that consisted of a n*2 dimensional and a n dimensional linear layer, n being the number of classes in each task.

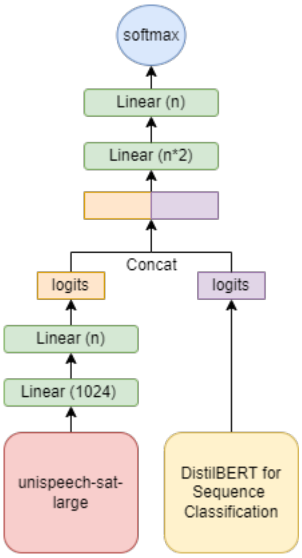

Figure 3: *MMML model architecture.*

4. Results

The obtained results are shown in Table 4.

The first observation when focusing on audio models is that UniSpeech-SAT outperforms the DCNN with Mel Spectrograms as an input. For this reason, from now on we will only compare UniSpeech-SAT (we will call it the audio model from now on) and the text based model.

The audio model outperforms the text based model when focusing on the categorical emotion and arousal while having very similar results in valence and dominance classification.

Looking more closely into each emotional feature, in the categorical emotion classification task the best results are the ones obtained with the audio model. Even in that case, the results are not very good. One of the facts that may cause this is that while the model has a high classification performance when regarding *Calm* and *Upset/Tense*, it performs very poorly when focusing on *Worried*. For example,in the case of the audio model, the obtained F-scores for *Calm* and *Upset/Tense* are respectively **0.78** and **0.81** while the F-score for *Worried* is as low as **0.19**. This makes sense considering that *Worried* is the minority class with only 92 samples vs the 643 and 361 that *Calm* and *Upset/Tense* have, which makes it hard to learn a representation for this class. However when using MML there is an improvement in the classification.

When considering Valence, results are similar when working with text and audio, being, in both cases, pretty low. It is the feature of the VAD with the lowest F-score values, which might be related to having to predict three labels instead of two (as is the case for Arousal and Dominance) since this reduces the quantity of training data for each class. Valence has also shown to generally perform worse in audio centered researches than arousal [26]. However, when using MMML we can see a very slight improvement. The results in [27] show that even the best-performing HuBERT representation under-performs on Valence prediction compared to a multimodal model that also incorporates text representation. The results in [26] also show that, while Valence is hard to detect in audio, text based features do add to the accuracy of prediction of Valence for speech stimuli. Our results might be in line with these observations and it might be interesting to further look into it.

Arousal is the feature that has the highest prediction accuracy. This can also be seen in other researches considering VAD [26]. This result is achieved when working with audio, which is in line with the results in [28] that show that Valence is better estimated using semantic features while Arousal is better estimated using acoustic features.

Regarding Dominance the achieved results are very similar when working with all the tested models. This might be a consequence of not having enough data as to learn a representation.

Table 4: *Classification results F-score.*

	E	V	A	D
Spectrograms	0.49	0.36	0.63	**0.57**
UniSpeech-SAT	0.59	0.46	**0.73**	**0.57**
DistilBERT	0.52	0.46	0.59	0.56
MMML	**0.61**	**0.47**	0.70	0.56

5. Conclusions and Future Work

We can remark the value of MMML models for categorical emotion recognition in our task. It would also be interesting to further analyze their use for Valence classification.

We observed that Valence is the best recognized dimension when using textual information, while Arousal has better outcomes when working with audio. This might make sense considering that positive or negative feelings about something are easier to detect in semantics than Arousal or Dominance [29] which might be features that are more related to acoustics.

For future work it would be interesting to explore other classification architectures and label more data to improve the results and make it possible to learn representations for more classes. For example, it would be interesting to have more data of the *Enthusiastic* class, since, as seen in Figure 1 and in [12], it is quite distinguishable from other emotions in our corpus.

6. Acknowledgements

The research presented in this paper was conducted as part of the AMIC PdC project, which received funding from the Spanish Ministry of Science under grants TIN2017-85854-C4-3-R, PID2021-126061OB-C42 and PDC2021-120846-C43 and it was also partially funded by the European Union's Horizon 2020 research and innovation program under grant agreement No. 823907 (MENHIR).

7. References

[1] A. Raheel, M. Majid, M. Alnowami, and S. M. Anwar, "Physiological sensors based emotion recognition while experiencing tactile enhanced multimedia," *Sensors*, vol. 20, no. 14, p. 4037, 2020.

[2] P. Ekman, *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Henry Holt and Company, 2004.

[3] M. A. Nasri, M. A. Hmani, A. Mtibaa, D. Petrovska-Delacrétaz, M. B. Slima, and A. B. Hamida, "Face emotion recognition from static image based on convolution neural networks," in *5th International Conference on Advanced Technologies for Signal and Image Processing, ATSIP 2020, Sousse, Tunisia, September 2-5, 2020*. IEEE, 2020, pp. 1–6.

[4] K. R. Scherer, *Approaches To Emotion. Chapter: On the nature and function of emotion: A component process approach*. K. R. Scherer & P. Ekman. Taylor and Francis Group, 1984.

[5] M. deVelasco, R. Justo, A. López-Zorrilla, and M. Torres, "Can spontaneous emotions be detected from speech on tv political debates?" in *Proceedings of the 10th IEEE International Conference on Cognitive Infocommunications*, Naples, 2019.

[6] M. deVelasco, R. Justo, A. López-Zorrilla, and M. I. Torres, "Automatic analysis of emotions from speech in spanish tv debates," *Acta Polytechnica Hungarica (In Press)*, vol. 19, pp. 149–171, 2022.

[7] R. Justo, M. I. Torres, and J. M. Alcaide, "Measuring the quality of annotations for a subjective crowdsourcing task," in *Pattern Recognition and Image Analysis*, L. A. Alexandre, J. Salvador Sánchez, and J. M. F. Rodrigues, Eds. Springer International Publishing, 2017, pp. 58–68.

[8] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions, IJSE*, pp. 68–99, 2010.

[9] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062 – 1087, 2011, sensing Emotion and Affect - Facing Realism in Speech Processing.

[10] J. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[11] I. Bakker, T. Van der Voordt, J. Boon, and P. Vink, "Pleasure, arousal, dominance: Mehrabian and russell revisited," *Current Psychology*, vol. 33, pp. 405–421, 10 2014.

[12] M. de Velasco, R. Justo, and M. Inés Torres, "Automatic identification of emotional information in spanish tv debates and human-machine interactions," *Applied Sciences*, vol. 12, no. 4, 2022.

[13] A. Ortega, E. Lleida, R. S. Segundo, J. Ferreiros, L. F. Hurtado, E. S. Arnal, M. I. Torres, and R. Justo, "Amic: Affective multimedia analytics with inclusive and natural communication." *Proces. del Leng. Natural*, vol. 61, pp. 147–150, 2018.

[14] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, pp. 273–294, 1977.

[15] S. S. Stevens, J. E. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *Journal of the Acoustical Society of America*, vol. 8, pp. 185–190, 1937.

[16] K. Venkataramanan and H. R. Rajamohan, "Emotion recognition from speech," *arXiv preprint arXiv:1912.10458*, 2019.

[17] "Unispeechsatlarge," accessed: 2022-08-10. [Online]. Available: https://huggingface.co/microsoft/unispeechsatlarge

[18] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li, and X. Yu, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022.

[19] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-light: A benchmark for asr with limited or no supervision," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020.

[20] G. Chen, S. Chai, G.-B. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *Interspeech 2021*, Aug 2021.

[21] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021.

[22] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *ArXiv*, vol. abs/1910.01108, 2019.

[23] "Datasets: large_spanish_corpus," accessed: 2022-08-10. [Online]. Available: https://huggingface.co/datasets/large_spanish_corpus

[24] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, "Multimodal deep learning models for early detection of alzheimer's disease stage," *Scientific reports*, vol. 11, no. 1, pp. 1–13, 2021.

[25] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski *et al.*, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[26] M. Asgari, G. Kiss, J. van Santen, I. Shafran, and X. Song, "Automatic measurement of affective valence and arousal in speech," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 965–969.

[27] S. Srinivasan, Z. Huang, and K. Kirchhoff, "Representation learning through cross-modal conditional teacher-student training for speech emotion recognition," 2021.

[28] S. G. Karadoğan and J. Larsen, "Combining semantic and acoustic features for valence and arousal recognition in speech," in *2012 3rd International Workshop on Cognitive Information Processing (CIP)*, 2012, pp. 1–6.

[29] Z. Yao, X. ru Zhu, and W. Luo, "Valence makes a stronger contribution than arousal to affective priming," *PeerJ*, vol. 7, 2019.