Mental Health Monitoring from Speech and Language

Irune Zubiaga, Ignacio Menchaca, Mikel de Velasco, Raquel Justo

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

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

Abstract

Concern for mental health has increased in the last years due to its impact in people life quality and its consequential effect on healthcare systems. Automatic systems that can help in the diagnosis, symptom monitoring, alarm generation etc. are an emerging technology that has provided several challenges to the scientific community. The goal of this work is to design a system capable of distinguishing between healthy and depressed and/or anxious subjects, in a realistic environment, using their speech. The system is based on efficient representations of acoustic signals and text representations extracted within the self-supervised paradigm. Considering the good results achieved by using acoustic signals, another set of experiments was carried out in order to detect the specific illness. An analysis of the emotional information and its impact in the presented task is also tackled as an additional contribution.

Index Terms: acoustic signal, textual information, mental health monitoring

1. Introduction

Mental Health is an essential component of overall health. However, depression and anxiety are still common disorders. As an example, in Europe the overall prevalence of current depressive disorder is between 5-10%, according to [1], with potentially large differences between countries and time. This kind of disorders are a major cause of disability, increasing the risk of premature mortality, decreasing life quality, and creating a substantial burden on health systems [2]. Many people may be described as “living with” a mental illness, and managing their own symptoms. However, they are often unsure of the thresholds for treatment, how to control their symptoms, which are the best coping strategies or the available resources.

AI based systems may facilitate watchful waiting and symptom monitoring by initiating contact and symptom checking at various times of the day and night. Thus, they can play a relevant role in the detection of the illness and in the patients care. These automatic systems can take advantage of different information sources like voice, gait, EEG, facial expressions, etc. [3, 4]. Speech can be an easily accessible, non-invasive marker, whose features can significantly change due to slight psychological or physiological changes [5]. Therefore, it can be considered a key marker for detecting depression and suicide risk [5, 6]. Different features like acoustic parameters extracted from fundamental frequency [7] or vocal-source-based features (jitter, shimmer, etc.) [5] have been used as successful cues for predicting depression. These features, among others, have been exploited by different machine learning algorithms. Spectral and prosodic features along with their statistics extracted using openSMILE toolkit [8] have been used to train support vector machines and random forest models for depression detection [9]. More recently deep Convolutional Neural Networks (CNNs) have been used to extract acoustic embeddings [10] and detect depression from speech [11, 12].

In this work, we focus on an alternative and more efficient way of representing the audio. The rise in the self-supervised learning paradigm and the recently proposed transformer architecture [13], have led to novel speech representations, such as wav2vec [14], HuBERT [15] or the most recent UniSpeech-Sat [16]. These models are inspired by deep Transformer-based text generation models[17] such as GPT[18, 19, 20] and BERT[21] that are able to extract features from simple text without additional annotation. Audio encoder models have also been used to extract speech representations. In this work, the HuBERT representation was selected to feed a simple neural network that can be trained with small amounts of data. This way, we can tackle an anxiety and depression detection task in a realistic environment where getting a large training set is time consuming and expensive. The HuBERT representation was also compared to the spectral and prosodic features achieved with openSMILE. Moreover, the transcriptions of the utterances were also considered as an information source. A BERT based representation of the text corresponding to the audio transcriptions, was also employed with the same aim. This way, an audio based system and a text based system were built to perform the anxiety and/or depression detection task. We also carried out an audio centered analysis in order to measure the impact of emotions, represented as a 2 dimensional model (Valence and Arousal), in the prediction of depression and/or anxiety.

The manuscript is organized as follows, Section 2 describes the task and corpus tackled in this work. Section 3 deals with the representation of the audio signal and the textual transcriptions. Section 4 and Section 5 detail the different sets of experiments that were carried out and Section 6 sums up the main conclusions and future work.

2. Mental health monitoring task

This work deals with the data acquired within the framework of the H2020-MSCA-RISE project MENHIR [22] (Mental health monitoring through interactive conversations). In this project 60 conversations between a counsellor and a participant were recorded to form a corpus. Participants were divided in two groups: **AMH** and **Control**. AMH (32 members) consists of users of the Action Mental Health (AMH) foundation[1], diagnosed with some kind of mental illness, depression and anxiety being the most common ones. In contrast the Control group (28 members) is formed of people who have never been diagnosed with any kind of mental illness.

The interviews consist of three main sections; In the first section the counsellor asks the patient 5 questions that lead to non-emotional conversation. The second part consists of fourteen affirmations from the Warwick-Edinburgh Mental Well-Being Scale (WEMWBS) [23]. The participants have 5 possible answers that go from ”None of the Time” to ”All of the Time” to indicate how often they feel the way that these affirmations

[1] https://www.amh.org.uk/

Table 1: *Distribution of Anxiety/Depression in AMH and Control.*

	No. speakers	No. interventions
Depression	3	276
Anxiety	2	140
Both	16	824
Control	20	614

express. In this part, in addition to the answer, sometimes the interviewees added some dialogue of their own to explain further the answer they gave. In this section the counsellor perception of the speaker emotional status was also annotated according to the following questionnaire:

How do you perceive the client?

- Excited/activated/agitated
- Slightly Excited/activated/agitated
- Calm

His/her mood is:

- Positive
- Slightly positive
- Slightly negative
- Negative

Finally, in the last part, the participants were asked to read a text passage of the popular tale ”The Boy and the Wolf”.

2.1. Speech corpus

The second section of the interviews, with its corresponding emotional information, was used For generating the speech corpus. This section consists of 14 questions that the potential patients (speakers) respond to. The audio sessions were split according to each turn or intervention of the speaker. This way, the corpus was formed by the audio files corresponding to those interventions for which the counsellor annotated their perception of Valence and Arousal levels on the participants speech. Said interventions were extracted from 41 interview recordings; 21 regarding AMH and 20 regarding Control. The total corpus consists of 1854 audio files (interventions) and has a length of 4 hr 15 min and 45 sec. The number of speakers and interventions associated to each mental illness is given in Table 1.

2.2. Textual corpus

The textual corpus consists of transcriptions of the aforementioned interviews. These transcriptions include both dialogues written in a literal way and annotations regarding paralinguistic or acoustic information; *noises (music, footsteps etc.)*, *pauses*, *laughs* and such. We ignored annotations regarding noise and will refer to the rest of annotations as **paralinguistic tokens** from now on. For our task, we only considered the interviewee's conversational information and paralinguistic tokens. The transcription of the reading phase was also removed since there is no distinguishing semantic information associated with it. All the text associated to the remaining dialogues was gathered as a corpus consisting of 6741 sentences; 4743 sentences regarding AMH and 1998 regarding Control. Starting from this, we created four textual data sets to work with.

To create the first two data sets a cleaning process was carried out and thus we will refer to them as clean data. First, the typos in the text were corrected as far as possible. Then, the answers to the questions from the WEMWBE scale (fixed response like “None of the time”, etc.) and the yes/no answers were removed. In the next step, paralinguistic features that were gathered as plain text were represented using a unique token from the ones given in Table 2 (ex. em, ew, jeeze, oh... were represented as {*INTERJECTIONS*}). This process is explained in more detail in [24]. This way, two data sets were created; one considering clean plain text, which we will refer to as **CD**, and another one considering clean plain text and paralinguistic tokens from Table 2, which we will refer to as **CD+T**. These two data sets consist of 3955 sentences; 2810 sentences from the AMH group and 1145 from the Control group. The other two data sets were formed with a more natural approach; we did not do any kind of text cleaning, typo correction or word grouping. We used the literal transcriptions of the dialogues and the paralinguistic tokens were written between parentheses to process them as plain text. Their representations are shorted in Table 3. Two new data sets were created; one considering plain text, which we will refer to as **D**, and another one considering plain text and paralinguistic tokens from Table 3, which we will refer to as **D+T**.These data sets consists of 4872 sentences; 3308 regarding AMH and 1564 regarding Control.

Table 2: *Paralinguistic tokens in CD and their meanings.*

Token	Meaning
{INTERJECTIONS}	uses an interjection
{CUT}	stops speaking mid-word
{inhaling}	inhales
{laugh}	laughs
<pause>WORDS</pause>	says something between pauses
{STALLING}	makes a sound denoting they are thinking
<laugh>WORDS</laugh>	says something while laughing
{breathing}	takes a deep breath
{tsk}	makes a flicking sound with their mouth
{STUTTER}	stutters when saying a word
{PUZZLEMENT}	makes a sound denoting puzzlement
{cough}	coughs
<breathing>WORDS</breathing>	says something while taking a deep breath
<pause>empty</pause>	makes a long pause
pause at start	marks if the sentence starts with a pause

Table 3: *Paralinguistic tokens in D and their meanings.*

Token	Meaning
(inhaling)	inhales
(laugh)	laughs
(breathing)	takes a deep breath
(tsk)	makes a flicking sound with their mouth
(cough)	coughs

3. Data processing

3.1. Acoustic Features

Affective processes can change Arousal and tension influencing both voice and speech production. These changes can be estimated with different parameters of the acoustic waveform. In this research, we used GeMAPS and Hubert to extract features of the vocal expressions, from acoustic waveforms, that provide this kind of information.

The GeMAPS (Genova Minimalistic Parameter Set) [25] feature set is divided into two main blocks.; the first block is composed of some prosodic, excitation and vocal tract descriptors, and the second one of some dynamic and cepstral descriptors. The first descriptors are classified depending on their relation to different physical characteristics like frequency (pitch or jitter) energy (shimmer, loudness or harmonics-to-noise ratio) and spectrum (alpha ratio, hammaberg index, spectral slope or harmonic differences). The second ones are parameters such as mel-frequency cepstral coefficients (MFCC) or spectral flux. Furthermore, features such as arithmetic mean, coefficients of variation, 20/50/80 percentiles and other temporal parameters have been added to the set, making a total of 88 different features. The OpenSmile Python library [8] was used to achieve this set of 88 features from an audio file in wav format.

Alternatively, Hubert was chosen with the pre-trained parameters *hubert-large-ll60k*, which have been fitted with the Libri-Light [26] corpus at 16k hertz. Such models trained only on raw audio have proven to be able to extract very representative features that have been widely used for different tasks. The model extracts a total of 1024 features every 20ms which have then been reduced by averaging to 1024 features.

3.2. Text representation

For text representation we worked with BERT (Bidirectional Encoder Representations from Transformers) [21]. BERT is based on a deep Transformer encoder network [17]. This kind of network can process long texts efficiently by using self-attention. BERT is bidirectional, which means that it uses the whole input text to understand the meaning of each word. In its base size (the one presented in [21]), BERT is composed of 12 successive transformer layers, each having 12 attention heads and has a total number of 110 million parameters. The BERT Encoder block calculates a 768-long vector representing an embedding of each input token.

4. Prediction of a mental disorder

Regarding the detection of mental illness, two different experiments were conducted. The first one tried to differentiate AMH and Control group members (Section 4.1).The second one will try to identify which participants suffer from anxiety and which ones from depression within the AMH members (Section 4.2).

For this purpose, 8 folds of the data set have been built in order to create a Cross-Validation in which no intervention from the interviews that form the test subset appears in the train partition. This way more robust and reliable results were achieved.

4.1. Prediction of interventions related to Healthy and Ill subjects

The task to discriminate healthy and ill subjects was carried out using both acoustic and textual information separately.

As for the acoustic information, the speech corpus described in Section 2.1 was considered. The GeMAPS and HuBERT feature sets were used to feed a Deep Neural Network and a random oversampling method was used in the training set to balance the data. In both cases the network consisted of two hidden layers, the first one with a ReLu activation function and an output layer with the softmax activation function. Due to the different feature dimension in each set the hidden layer was 32 dimensional when using GeMAPS and 128 dimensional when using HuBERT. Adam optimizer was used with a learning rate of 1e-4. The batch size was set to 32 and the cross-entropy loss function was used. The training was done over 250 epochs.

When regarding textual information, BERT was used to classify mentally ill and healthy people. To carry out this task, we used a model that consisted of a BERT main layer and a classification head [27]. Three different pre-trained models were used as a checkpoint for the BERT main layer and for the tokenizer: *bert-base-cased* and *bert-base-uncased* witch were presented in [21] and *bert-base-uncased-emotion* [28] which is a pre-trained *bert-base-uncased* model fine-tuned with the *emotion* data set [29]. Then, the whole model was fine-tuned using the four variations of the textual corpus described in Section 2.2. The chosen optimizer was AdamW and a linear scheduler was used. The learning rate was set to 5e-5, the batch size to 16 and the fine-tuning was done in 3 epochs. These parameters were chosen following the recommendations in [21]. In the case of **CD+T** the paralinguistic tokens were added to the tokenizer as *additional special tokens* to try and learn a representation for them. In contrast, when working with **D+T**, we processed paralinguistic tokens as if they were plain text. The obtained results for this task are shorted in Table 4.

Table 4: *Macro F1-scores of the results obtained in the audio based and text based approaches to the AMH and Control group discrimination problem.*

	GeMAPS	HuBERT
Acoustic DNN	0.88	**0.94**

	CD	CD+T	D	D+T
bert-base-cased	0.61	0.56	0.66	0.60
bert-base-uncased	0.60	0.55	0.67	0.56
bert-base-uncased-emotion	0.60	0.60	0.65	**0.68**

As seen in Table 4, when working with text the best result is obtained when using *bert-base-uncased-emotion*, which can be a cause of this model being fine-tuned on emotional data. The general lack of accuracy in BERT's predictions is probably a consequence of the majority of sentences having a length of one or two words. In addition, most long sentences were difficult to classify even by humans (ex."two rounds of granary toast with a banana") and the data set was noisy, which has shown to significantly degrade BERT's performance when fine-tuning it for tasks such as sentiment analysis [30]. The results are better when working with non clean data. This might be because BERT uses features to provide context information that are being removed by cleaning words and representing paralinguistic features with tokens (ex. replacing *wh-what?* with {*STUTTER*}). It can also be observed that when using *bert-base-cased* and *bert-base-uncased* taking paralinguistic tokens into account makes the results worse. In the case of **CD+T**, this can be because there was not a pre-learned representation for these tokens and the performed training is not enough as to learn one. In the case of **D+T** this might be because they contain no semantic information. In contrast, when working *bert-base-uncased-emotion*, **D+T** has a better outcome than **D**. This might mean that there is some kind of emotional information in the tokens from Table 3. In the case of **CD** and **CD+T** it appears that while a representation for tokens in Table 2 still seems hard to learn, their presence does not worsen the results as seen when using other model checkpoints. It is worth noting that in this case the transcriptions were done by humans and if a full automatic system is required, an ASR for text detection would have to be introduced, which would lead to slightly worse results.

When working with acoustic signals the system performance was significantly better. This shows that the acoustic model is able to identify some features that are not detectable in text and that help differentiating one group from the other. The GeMAPS features turned out to be worse than those extracted with the Hubert model. This highlights the power of the acoustic embeddings achieved making use of semi-supervised learning. Besides, the number of features extracted with GeMAPS (88) is smaller than those extracted with HuBERT (1024).

Table 5: *Averaged F1-Score for the Cross-Validation on anxiety and depression detection problems.*

	GeMAPS	HuBERT
Anxiety	0.64	**0.71**
Depression	0.53	**0.70**

4.2. Prediction of Depression and Anxiety Interventions

The results achieved with speech signal encouraged us to try to discriminate between depression and anxiety. These sets of experiments were only performed on the AMH group.

With this aim, a system was designed to predict whether an intervention corresponds to a patient with or without depression. Alternatively, an additional network was designed to predict whether it corresponds to a patient with or without anxiety.

Both experiments replicate the procedure mentioned in Section 4.1 but with an adapted output to the new classification problem. The results are shown in Table 5. The results are not as good as those in Section 4.1 because the task is now more challenging as a consequence of the two classes being more similar between them. All samples belong to the AMH group, which means that even if the subject does not have the target illness, it does suffer from another one. Additionally, the class imbalance problem is more relevant in this case. However, the results with HuBERT features are still promising.

5. Detection of Emotional Information

As a first step to include affective information, an analysis of the counselors emotional annotations was carried out. Figure 1 shows the histogram for the interventions annotated with different values of Valence and Arousal from the AMH and the Control group. As expected, lower values of pleasure were perceived in AMH group, meaning that their status is generally more negative than the status of the members of Control group, whose values are more positive. Something similar can be observed in the Arousal levels, where AMH members show lower levels of excitement. However, the difference is not as significant in this case, revealing that Valence might be a more informative feature for the detection of anxiety and depression.

When focusing on specific illnesses and differences between them, the analysis shown in Figure 2 was carried out. This figure shows the percentage of interventions from each group (Depression, Anxiety, Both, None) that were annotatedwith an emotional label for Valence and Arousal. When regarding Valence, patients with depression are the most negative ones and those with no illnesses the most positive. Patients with anxiety who are not also diagnosed with depression are in between, suggesting that they might be differentiated from depressed ones using Valence as a cue for illness detection. On the other hand, the differences in Arousal levels are not as meaningful when comparing depression and anxiety.

Figure 1: *Percentage of interventions per group.*

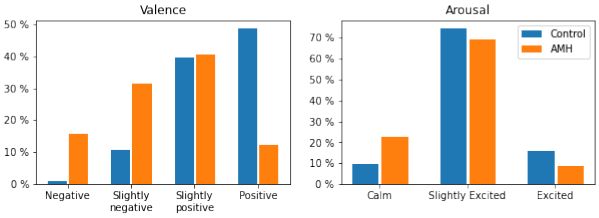

Figure 2: *Percentage of interventions per illness.*

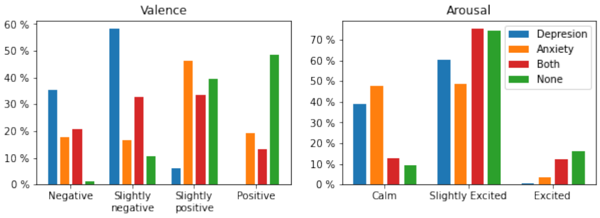

Table 6: *F1-Score on Valence and Arousal detection problems.*

	GeMAPS	HuBERT
Valence	0.35	0.46
Arousal	0.41	**0.57**

Finally, an experiment was conducted in order to predict Valence and Arousal associated to each intervention. Thus, a classifier was designed to predict among the 4 different Valence categories and the 3 statuses of Arousal. Once again the DNN architecture and training paradigm explained in Section 2 were implemented for the classification, but considering a random 90%-10% train-test split instead of the designed 8-folds Cross-Validation, due to a lack of emotion distribution across the folds. The achieved results are given in Table 6.

The task of identifying emotions is even more complicated since it involves a larger number of classes and the emotional annotation is a perception of the interviewer, what makes the task very subjective. Even so, we still draw the same conclusions as in Sections 4.1 and 4.2. Looking at the achieved results and by looking at Figures 1 and 2 it might be interesting to focus on Valence for future work, since it seems to provide more relevant information related to depression and anxiety than Arousal despite having worse results. Moreover, simplifying the Valence information into two different classes (positive and negative) might lead to more accurate results.

6. Conclusions

This manuscript provides a system capable of detecting depression and/or anxiety from speech signal uttered by potential patients in an interview. Our experiments show that acoustic features based on HuBERT transformer significantly outperform the classical GeMAPS extended set. Thus, an additional experiment was carried out in order to distinguish between anxiety and depression. Although the achieved results are not as impressive, HuBERT features still provide promising results. The text associated to transcriptions is also taken into account to build an alternative system that, although provides worse results, can be considered as an alternative information source. Finally, an analysis of emotional information associated to the interventions was conducted to study its potential use for future work showing that Valence might be an interesting marker.

7. Acknowledgements

This work was partially funded by the European Commission, grant number 823907 and the Spanish Ministry of Science under grant TIN2017-85854-C4-3-R.

8. References

[1] J. Arias-de la Torre, G. Vilagut, A. Ronaldson, A. Serrano-Blanco, V. Martín, M. Peters, J. M. Valderas, A. Dregan, and J. Alonso, "Prevalence and variability of current depressive disorder in 27 european countries: a population-based study," *The Lancet Public Health*, vol. 6, no. 10, pp. e729–e738, 2021.

[2] G. Archer, D. Kuh, M. Hotopf, M. Stafford, and M. Richards, "Association Between Lifetime Affective Symptoms and Premature Mortality," *JAMA Psychiatry*, vol. 77, no. 8, pp. 806–813, 08 2020.

[3] H. Cai, X. Zhang, Y. Zhang, Z. Wang, and B. Hu, "A case-based reasoning model for depression based on three-electrode EEG data," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 383–392, 2020.

[4] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear, "Head pose and movement analysis as an indicator of depression," *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, 09 2013.

[5] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[6] Z. Zhao, Z. Bao, Z. Zhang, J. Deng, N. Cummins, H. Wang, J. Tao, and B. Schuller, "Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 423–434, 2020.

[7] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: features and normalization," in *Proc. Interspeech 2011*, 2011, pp. 2997–3000.

[8] F. Eyben, M. Wöllmer, and B. W. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor." in *ACM Multimedia*, A. D. Bimbo, S.-F. Chang, and A. W. M. Smeulders, Eds. ACM, 2010, pp. 1459–1462.

[9] M. F. Valstar, B. Schuller, K. Smith, T. R. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *AVEC '14*, 2014.

[10] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of Biomedical Informatics*, vol. 83, pp. 103–111, 2018.

[11] K. Chlasta, K. Wołk, and I. Krejtz, "Automated speech-based screening of depression using deep convolutional neural networks," *Procedia Computer Science*, vol. 164, pp. 618–628, 12 2019.

[12] S. H. Dumpala, S. Rempel, K. Dikaios, M. Sajjadian, R. Uher, and S. Oore, "Estimating severity of depression from acoustic features and embeddings of natural speech," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7278–7282.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[14] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *CoRR*, vol. abs/2006.11477, 2020.

[15] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[16] 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," 2021.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[18] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[20] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[22] Z. Callejas, K. B. Akhlaki, M. Noguera, M. I. Torres, and R. Justo, "MENHIR: mental health monitoring through interactive conversations," *Proces. del Leng. Natural*, vol. 63, pp. 139–142, 2019.

[23] R. Tennant, L. Hiller, R. Fishwick, S. Platt, S. Joseph, S. Weich, J. Parkinson, J. Secker, and S. Stewart-Brown, "The warwick-edinburgh mental well-being scale (wemwbs): Development and uk validation," *Health and Quality of Life Outcomes*, vol. 5, no. 1, 2007.

[24] I. Zubiaga and R. Justo, "Multimodal feature evaluation and fusion for emotional well-being monitorization," in *Pattern Recognition and Image Analysis: 10th Iberian Conference, IbPRIA 2022, Aveiro, Portugal, May 4–6, 2022, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 242–254.

[25] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[26] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.

[27] "Bertforsequenceclassification," accessed: 2022-06-15. [Online]. Available: https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertForSequenceClassification

[28] "bhadresh-savani/bert-base-uncased-emotion," accessed: 2022-06-15. [Online]. Available: https://huggingface.co/bhadresh-savani/bert-base-uncased-emotion

[29] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "CARER: Contextualized affect representations for emotion recognition," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 3687–3697.

[30] A. Kumar, P. Makhija, and A. Gupta, "Noisy text data: Achilles' heel of bert," 2020.