2023 62nd Annual Conference of the Society of
Instrument and Control Engineers of Japan (SICE)
September 6-9, 2023. Tsu City, Japan

Advancing Mental Health Diagnostics: GPT-Based Method for Depression Detection

Michael Danner[1*], Bakir Hadzic[2*], Sophie Gerhardt[2], Simon Ludwig[2], Irem Uslu[2], Peng Shao[3], Thomas Weber[2], Youssef Shiban[4], Matthias Rätsch[2†]

[1] University of Surrey, Guildford, United Kingdom
[2]ViSiR, Reutlingen University, Reutlingen, Germany
[3]School of Management, Xi'an Polytechnic University, Xi'an, China
[4]Private University of Applied Sciences, Göttingen, Germany
*both authors contributed equally

Abstract: In this paper, we present a novel artificial intelligence (AI) application for depression detection, using advanced transformer networks to analyse clinical interviews. By incorporating simulated data to enhance traditional datasets, we overcome limitations in data protection and privacy, consequently improving the model's performance. Our methodology employs BERT-based models, GPT-3.5, and ChatGPT-4, demonstrating state-of-the-art results in detecting depression from linguistic patterns and contextual information that significantly outperform previous approaches. Utilising the DAIC-WOZ and Extended-DAIC datasets, our study showcases the potential of the proposed application in revolutionising mental health care through early depression detection and intervention. Empirical results from various experiments highlight the efficacy of our approach and its suitability for real-world implementation. Furthermore, we acknowledge the ethical, legal, and social implications of AI in mental health diagnostics. Ultimately, our study underscores the transformative potential of AI in mental health diagnostics, paving the way for innovative solutions that can facilitate early intervention and improve patient outcomes.

Keywords: Mental Health, Depression Detection, Deep Learning, NLP Transformer LLM, GPT-3.5, ChatGPT-4.

1. INTRODUCTION

According to the World Health Organisation's most recent report on mental health [1], one in eight people worldwide lives with some sort of mental health disorder. The most common ones are depressive and anxious disorders which are disorders correlated with a reduced quality of life and a very high risk of committing suicide. Around 800,000 people attempt suicide each year, and there are indications that there may have been more than 20 such attempts for every adult who died by suicide [2]. Many of them could have been avoided, and the lives of those impacted could have been improved if the mental health issue that caused them had been detected at the time. However, currently, available techniques for detecting depression are insufficiently effective. Unfortunately, almost two-thirds of those who require it do not obtain mental health care, primarily as a result of limitations to reaching mental health professionals, stigmatisation, high costs, or extensive waiting lists. All of these factors contribute to the vast majority of persons with mental health disorders remaining undetected [3]. This issue is particularly evident in economies that are developing, where mental health systems are not up to standard and the population lacks access to proper mental health treatments. However, this issue is also evident in developed regions, like Europe or USA. For example, more than 123 million Americans reside in mental health professional shortage areas [4]. The ability of a therapist to get the necessary diagnostic information from a patient, who mostly has a diminished outlook and motivation, depends mainly on their competence and experience. Both the diagnosis of depression and the assessment of the risk of suicide are quite challenging and time-consuming processes [5].

Due to the reasons previously mentioned, there is an immense demand for novel strategies to address these issues and improve existing methods so that they may become automated, faster, non-invasive, less costly, as well as accessible to larger populations. Great possibilities for that are arising from the area of artificial intelligence (AI) and machine learning where various methods for that purpose are currently being tested and developed. Some of them are presented under related work.

Our approach has the ultimate goal to develop an AI-based method for depression detection using textual inputs. Within this paper, our text-based model for depression detection is presented. In the training process, we used Wizard-of-Oz (DAIC-WOZ) and Extended-DAIC [6, 7, 8] datasets which are based on the PHQ-8 self-administered questionnaire used to assess and monitor symptoms of depression as ground truth values. In the test phase, we used the test dataset from the DAIC-WOZ but also simulated clinical interview data collected especially for the purpose of this study. This gave us the opportunity to evaluate our model using data gathered from standardised clinical interviews and this presents a novel approach in this field.

The related work, methodology, results, and conclusion are presented in the following chapters.

2. RELATED WORK

Depressive symptoms can express themselves through both verbal and nonverbal channels. These channels can

† Matthias Rätsch is the presenter of this paper.

encompass a wide range of modalities, such as voice, prosody, speech content, facial expressions, body postures, and other behavioural indicators [9, 10]. Such modalities can serve as sources of information that reflect the emotional state of an individual. By leveraging the detection of depression across multiple modalities, more robust and accurate insights can be extracted from a multi-dimensional perspective [11]. This can be achieved through various methods, including the use of audio recordings, where the voice, prosody and spoken text content can be utilised as data sources. Alternatively, depression can be detected through the recording of facial expressions, head postures, or gaze directions [9]. This chapter outlines the state-of-the-art developments in the area of text-based and audio-based depression detection, leveraging deep learning methodologies. Furthermore, it highlights the relevance of the advancements in the field of AI-based dialogue systems, especially in the area of large language models.

2.1. Text-based

One approach for detecting depression is to analyse speech and language patterns in individuals. Specifically, the text-based recognition of depression involves extracting and transcribing audio data to reveal the content of natural language [9]. This content can then be used to extract many pieces of information related to the emotional state of depressed individuals [11]. For the classification of this data, feature vectors are extracted and learned by a deep learning model [12]. In recent years, several algorithms, including convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and transformers, have been utilised for text-based depression recognition [12, 13]. Among these, transformer-based deep learning models have shown exceptional performance in the context recognition of natural language [14]. Consequently, such models have become a promising research direction in the field of depression diagnosis and treatment [15, 16, 17, 12, 13]. Toto et al. [15] present an example of a smartphone application named EMU3 that aims to detect depression using a multimodal speech classification system called AudiBERT. The proposed AudiBERT model utilises dual self-attention mechanisms to establish the relationship between the structures of text and audio recordings. The model is trained and tested on the DAIC-WOZ dataset using BERT-based text recognition. The results demonstrate the effectiveness of AudiBERT in accurately detecting depression using both audio and text features. The proposed framework has promising potential for developing scalable, accessible, and cost-effective tools for the early detection and treatment of depression.

2.2. Audio

Deep learning models are capable of automatically extracting and learning complex patterns and relationships from audio data, which enables more accurate and reliable detection of depression [18]. The audio features used for training these models include pitch, intensity, and spectral features, which can be extracted from audio recordings using techniques such as Mel-frequency cepstral coefficients (MFCCs) [19], pitch analysis, and spectral feature analysis. The majority of current speech processing techniques first divide speech into brief (10–20 millisecond) frames before extracting low-level descriptors (such as spectral, prosodic, and glottal features) and high-level representations of those features (such as statistical functionals, such as mean and percentiles), vocal tract coordination (VTC) features, i-vectors, and Fisher vectors) [20]. Transformer-based methods have shown promising results in audio feature extraction and have become increasingly popular in recent years. McGinnis et al. [21] present an approach that is capable of recognising if children have internalised disorders like depression or anxiety. For that they need only 3 minutes of recorded speech and the accuracy of their approach is around 80%, which outperforms clinical thresholds on parent-reported child symptoms. Despite the fact that approaches using audio-based inputs have achieved quite promising results, Baileys and Plumbley [22] point out in their paper the presence of gender bias in depression detection models that utilise audio features. The study found that existing depression detection models trained on audio data exhibit significant gender bias, with higher accuracy for detecting depression in female voices compared to male voices. Additionally, the authors in their paper suggest potential solutions to mitigate the bias with a few different approaches.

2.3. Large-scale Language Models

The development of large-scale language models (LLMs) has revolutionised the field of natural language processing (NLP).

This has led to the emergence of several chatbots, based on LMMs, such as ChatGPT (based on LLM GPT 3.5/4.0 proposed by OpenAI/Microsoft [23]), BARD (based on LLM LaMDa, proposed by Google [24]), ERNIE (based on LLM ERNIE 2.3/3.0 [25] proposed by Baidu), or Dalai, Alpaca and others (based on LLM LLaMA proposed by Meta, Facebook [26]). These LMMs, based on the transformer architecture, have demonstrated significant advances in conversational AI and information retrieval.

It excels at generating coherent and contextually relevant responses, even for ambiguous queries. Chatbots based on LLMs are used in many fields of research and even in several everyday tasks. ChatGPT, an iteration of OpenAI's GPT series, has achieved outstanding performance in a wide range of tasks, such as translation, summarising, and question-answering. Its ability to generalise from few-shot learning examples is a testament to the model's adaptability. In parallel, Baidu's ERNIE model, which employs continual pre-training and knowledge distillation techniques, has shown impressive results in Chinese NLP tasks and achieved state-of-the-art performance on various benchmarks. Lastly, Meta's LLaMA model, utilising a combination of unsupervised and supervised learning, has been designed to handle

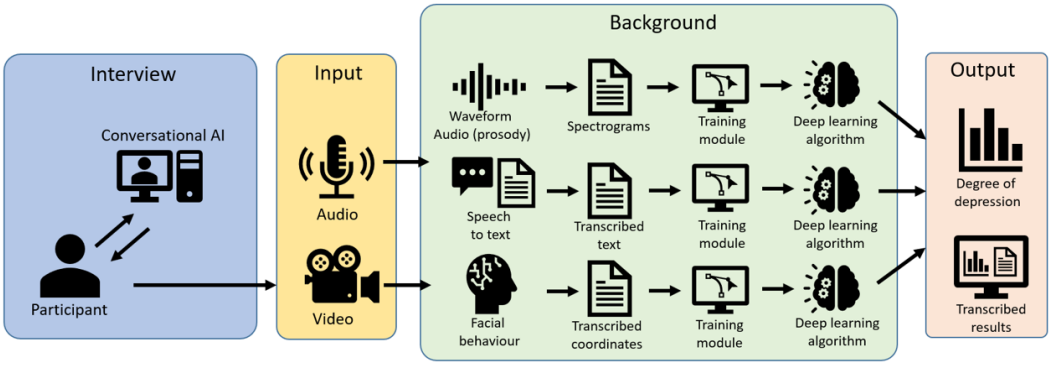

Fig. 1 Our comprehensive multimodal approach utilises text, audio, and video data from the DAIC interview process, seamlessly combining advanced methodologies to accurately determine depression severity and generate transcriptions.

low-resource languages and multi-modal data effectively. Each of these models offers unique contributions to the advancement of language understanding and generation, paving the way for more sophisticated AI applications.

3. METHODOLOGY

3.1. Multimodality

As evident in the previous chapters, several researchers have recently used deep learning techniques to identify depression using a multimodal approach. Most of the research relies on the Extended-DAIC dataset and DAIC WOZ, which we also employed in the present study. Content of fig. 1 visually demonstrates the core idea underlying the DAIC datasets. As can be seen, datasets include text, audio, and video inputs from interview respondents. Which data source provides the most successful symptom detection is a topic of continuing debate in the literature. Analysing data from the same dataset, Scherer et al. [27] focused only on analysing video inputs from the dataset. They found out that four behaviours linked with depression can be automatically recognised during the interview: angling of the head, eye gaze, duration and intensity of smiles, and self-touches. Despite the fact that nonverbal signs are offering quite valuable information, they alone are not sufficient. Cummins et al. [5] mainly focused on the effects of depression and suicidality on common paralinguistic speech characteristics (prosody, source features, formant features, as well as spectral features). As stated by Huang et al. [20], numerous studies have demonstrated that depression can affect speech production in a variety of ways, including cognitive impairment, phonation and expression errors, articulatory incoordination, disturbances in muscle tension, psychomotor delay, phoneme rates, as well as altered speech quality and prosody. Studies that use a multimodal approach tend to produce significantly better results than those that use one modality only [28, 29]. However, modalities like audio and video raise a variety of ethical and data protection concerns that remain quite a big challenge that we still can't tackle effectively [28]. Therefore, the data size and the efficiency for deep learning are the main reasons why most approaches and we focused on text-based inputs for the depression model. To fuse more channels is our aim for further studies.

3.2. Datasets

The datasets used in this paper (DAIC-WOZ, Extended-DAIC and simulated data) are labelled with the PHQ-8 Score [30]. The Patient Health Questionnaire-8 (PHQ-8) is a widely used and validated self-report questionnaire designed to assess the severity of depressive symptoms in individuals. Consisting of eight items, the PHQ-8 is derived from the longer PHQ-9 by excluding the question related to suicidal ideation, which makes it a more suitable tool for research and settings where discussing suicidal thoughts may not be appropriate. Participants rate the frequency of experiencing each symptom over the past two weeks on a scale ranging from 0 (not at all) to 3 (nearly every day). The total score, ranging from 0 to 24, is calculated by summing the individual item scores, with higher scores indicating more severe depression. The PHQ-8 is valued for its brevity and simplicity, making it an efficient and reliable tool for screening depression in various populations and healthcare settings. Questionnaires and examination methods with more robust psychometric properties and research on how successful our work can support the diagnosis of medical specialists in the clinical practice will be focused on in the next papers.

The Distress Analysis Interview Corpus (DAIC) from the University of California - Institute for Creative Technologies (USCICT) offers two variants of datasets. The DAIC-WOZ [6] is based on the Wizard-of-Oz experiment and was developed for the research of anxiety, depression and post-traumatic stress disorder (PTSD). And the Extended-DAIC [7] is an extension of the DAIC-WOZ with a specialisation on depression and PTSD.

The data consists of transcribed text, audio and video recordings and was collected through interviews with the virtual interviewer Ellie, controlled by a human interviewer in another room, but participants were not aware of that. See fig. 2 and fig. 3.

A novel approach to the test phase in our study is the use of simulated data from standardised clinical interviews that were conducted especially for the purpose of

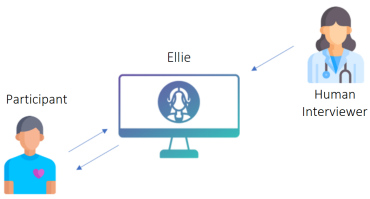

Fig. 2 DAIC Interview method (Ellie was controlled).

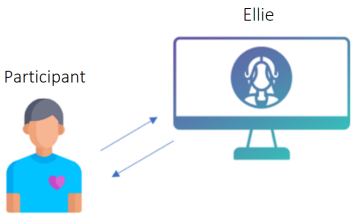

Fig. 3 Extended-DAIC interview method.

this study. The interviews are conducted to measure variables from the PHQ-8 questionnaire. Conducting clinical interviews with people who are actually experiencing a mental health disorder would be inappropriate in light of ethical and data protection rules. Because of this, psychology students were interviewed for our interviews by trained mental health experts. Before the interviews, students received comprehensive education on how depressed or non-depressed individuals would react and behave during the interview. The so-called "simulated data" in this study consists of transcribed interviews collected in this way.

3.3. Faulty Data from DAIC-WOZ

When evaluating the data, the following files are not included in the training, because they are described in the documentation of the dataset as noisy or interrupted transcriptions: 373 and 444. The files where Ellie is missing (451, 458 and 480) are not removed because only the statements of the participants are considered. The score after filling out the PHQ-8 questionnaire from file 409 is 10. This was wrongly listed as not depressive. For this reason, this label is corrected manually.

3.4. BERT

The emergence of NLP transformer models, particularly Bidirectional Encoder Representations from Transformers (BERT), has marked a significant milestone in the field of AI. BERT, developed by researchers at Google, is a pre-trained language model that can be fine-tuned for a wide array of NLP tasks, such as sentiment analysis, machine translation, and question-answering systems. BERT's architecture leverages the transformer, which is an attention mechanism that learns contextual relationships between words or tokens in a text. Unlike traditional, unidirectional language models, BERT is designed to process input sequences bidirectionally, enabling it to capture both past and future context simultaneously. This bidirectional approach allows BERT to outperform its predecessors in numerous NLP benchmarks, setting new standards in the field. One of the key advantages of BERT is its ability to benefit from transfer learning, where the knowledge acquired from pre-training on vast amounts of data can be transferred to specific tasks with relatively small datasets. This characteristic not only reduces the need for extensive labelled data but also accelerates model training and improves overall performance.

In this work, we applied the following algorithms in a BERT model described in table 1:

`BertTokenizer:` to split the text and tokenize sentences into subwords or word pieces for the BERT model given a vocabulary generated from the *Wordpiece* algorithm.

`BertForSequenceClassification:` a BERT model transformer with a sequence classification and regression linear layer on top of the pooled output.

Table 1 our BERT model parameters

Model	BERT-Base uncased
Environment	12-layer, 768-hidden, 12-heads
Parameters	110 M
Batch size	16
Length embedding	27
Epochs	15
Optimizer	AdamW
Learning rate	3×10^{-5}
Dropout hidden	0.3
Dropout attention	0.5
Weight-decay	4×10^{-2}

3.5. Data preprocessing

The data should be preprocessed for a more efficient classification process. First, the text from the interviewer was deleted and the rest of the transcript of the participants was saved in a string. The number of words in the transcribed text per participant ID in a string is more than 1500 words. BERT itself can process a maximum of 512 tokens, which is the reason why the text of the participants is divided into fractions of 25 words. Then contractions like "it's" and "don't" that often occur in the English language will get written out with the Python library contractions. Then punctuation and the resulting double spacing, Zero values and numbers are removed.

3.6. Uneven class distribution

The classes for depressed and non-depressed participants are unevenly distributed as shown in table 2. However, the class distribution should be even for the perception of a classification. To align the class distribution, an up-sampling of the minority class was performed. Additionally, we created more data by extracting the minority class and combining the second half of the current sentence and the first half of the next sentence of the ex-

tracted class, to get a new sentence with the same participant ID.

Table 2 Dataset class distribution

Dataset	depressed	non-depressed
DAIC-WOZ	56	133
Extended-DAIC	66	209

3.7. Fine-tuning

For the fine-tuning we started with a set of recommended hyperparameters from Delvin et al. [31]: AdamW optimiser with learning rate 3e-5, batch size of 32, and a number of epochs 4. With our model, we had a strong overfitting and to counteract that, we added a few regularisations. First, we tried to change the hidden and attention dropout rates after the work of El Anigri et al. [32]. On top of that, we also integrated a weight-decay over the AdamW optimiser.

After some fine-tuning, we ended up with the parameters from the table 1. These parameters are not ideal and there is still room for improvement.

3.8. Metrics

For the evaluation of the results, the classification report from the sklearn-library is used on the test data. The following metrics are displayed:

Precision: it refers to the proportion of true positive cases, or correctly diagnosed cases, out of all the cases that were identified as positive. Precision can be particularly important for medical diagnosis as misdiagnosing a condition can have serious consequences for a patient's treatment and overall well-being. A high precision rate indicates that the diagnosis is likely to be correct, and therefore, the patient can receive appropriate treatment and care.

Recall: also known as sensitivity, measures the proportion of actual positive cases that are correctly identified by the model. A higher recall value means that the model is able to identify a larger proportion of positive cases, which is generally desirable in applications such as medical diagnosis, where the goal is to identify all cases of a disease, even if it means some false positives are identified

F_1 **score:** it is calculated as the harmonic mean of precision and recall. This means that the F_1 score gives equal weight to both precision and recall. A high F_1 score indicates that a model has both high precision and high recall, which means it is able to correctly identify and classify cases accurately.

4. EXPERIMENTAL RESULTS

After fine-tuning and testing, we came to the conclusion that the best way to enhance our system lies within data augmentation because the amount of data we have from the DAIC-WOZ dataset is too low. Therefore in further experiments, we included the training data from the Extended-DAIC, see table 4. On top of that, we tested the DAIC-WOZ test data on untrained GPT-3.5 and GPT-4.

The questions in the simulated data are identical to those in the PHQ-8 questionnaire, but the conversation's form and content are different. While for the DAIC-WOZ dataset, participants believed that they communicated with the "artificial intelligence" (avatar Ellie was controlled by the human interviewer), simulated data was collected through the simulated clinical interview where participants were interviewed by trained mental health professionals. The results from the simulated test data are likely inferior to the test data from the DAIC database since the dialogue with the human interviewer is more natural and contains more filler words, which could make the evaluation more challenging. The novel approach proposed in this work and our suggestion for further studies on this subject is the inclusion of this kind of test data in the test phase. An algorithm will be superior to others if it can interpret input from different sources with the same accuracy. Such a model might be useful in a wide range of settings and could be utilised in many use cases. Future research in the field of human-machine interaction is necessary to investigate and compare individual willingness to interact with AI and humans in regard to their emotional and psychological conditions. In table 3, we present our results and compare them to those of other relevant studies on this topic.

Table 3 Experimental results and comparison.

Work	Precision	Recall	F_1 score
DAIC-WOZ dataset:			
Villatoro-Tello [33]			0.53
Villatoro-Tello [34]	0.59	0.59	0.59
Senn et al. [35]			0.60
Ours (BERT-based)	0.63	0.66	0.64
Ours (GPT3.5-based)	**0.78**	**0.79**	**0.78**
Simulated dataset:			
Ours (BERT-based)	0.68	0.41	0.43

It's important to note that some pertinent studies were left excluded from this comparison because they used different or various datasets, training methods, and models, or they presented metrics that are not comparable to our scores.

To the best of our knowledge, our model outperforms the most recent state-of-the-art results when compared under the previously described comparison standards.

Table 4 Further Experiment.

Model	Precision	Recall	F_1 score
DAIC-WOZ dataset:			
BERT	0.63	0.66	0.64
GPT-3.5	**0.78**	**0.79**	**0.78**
ChatGPT-4	0.70	0.60	0.61
DAIC-WOZ and Extended-DAIC dataset:			
BERT	0.83	0.82	0.82

In additional experiments outlined in table 4, we evaluated the performance of the DAIC-WOZ dataset using GPT-3.5-turbo, ChatGPT-4, and our custom BERT model with extended training on the larger dataset. Due to limited access to the ChatGPT-4 API, we resorted to testing the dataset with the chatbot. The GPT-3.5 model demonstrated remarkable improvement, significantly surpassing our initial experiments and previous results reported by other researchers.

The extended training data consists of the DAIC-WOZ [6] training data (107 entities) and the Extended-DAIC [7] training data (163 entities) with which we have over 2.5 times the size of the training data.

We conducted experiments using various configurations of the GPT-3.5 API to assess the impact of different settings. As illustrated in figure 4, the temperature setting, ranging from 0 (conservative) to 1 (creative), influences the results. However, the distribution is not entirely consistent, as it exhibits some irregularities and fluctuations.

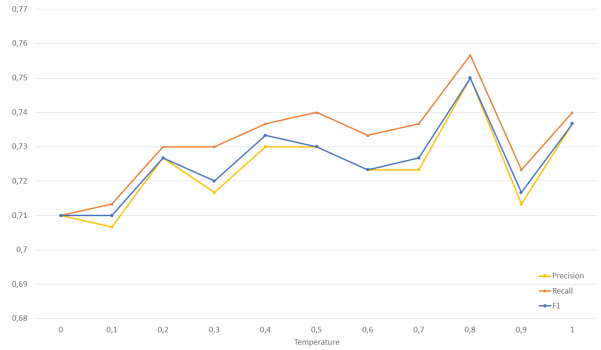

Fig. 4 GPT-3.5 results with different temperatures.

5. CONCLUSION

In conclusion, this paper has demonstrated the impressive potential of our model in detecting depression from text-based data and text-based simulated data. Our findings reveal that, compared to other methods, we could outperform other approaches in both accuracy and efficiency, thereby offering a robust and reliable means of identifying depression in individuals.

The superior performance of GPT-3.5-based models underscores the importance of continued research into and development of large-scale language models for depression detection. Given the increasing prevalence of mental health issues globally, developing automated tools capable of accurately detecting such conditions is crucial to providing timely and targeted support for those affected. As soon as we can access the GPT 4.0 API, fine-tuning this model will further improve the results.

Looking ahead, we believe that adopting larger large-scale language models could yield even more accurate and effective results in depression detection. As such, we encourage further exploration of this promising avenue, with a particular emphasis on refining these models to understand better and identify the nuanced manifestations of depression. Ultimately, this research may pave the way for creating powerful diagnostic tools, which could significantly enhance our ability to support individuals in need and make great strides in the ongoing battle against mental health challenges.

Acknowledgements

This work is partially supported by a grant of the BMWi ZIM-FuE programs, no. KK5007201LB0

REFERENCES

[1] World Health Organisation, “World mental health report: Transforming mental health for all”, *World Health Organisation*, 2022.

[2] World Health Organisation, “Preventing suicide: A global imperative”, World *Health Organisation*, 2014.

[3] J. Radez, T. Reardon, C. Creswell, P.J. Lawrence, G. Evdoka-Burton, and P. Waite, “Why do children and adolescents (not) seek and access professional help for their mental health problems? A systematic review of quantitative and qualitative studies”, *European child & adolescent psychiatry*, 30, pp. 183–211, 2021.

[4] M. Smith-East and D.F. Neff, “Mental health care access using geographic information systems: An integrative review”, *Issues in Mental Health Nursing*, 41(2), pp. 113–121, 2020.

[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*, 71, pp. 10–49, 2015.

[6] J. Gratch, R. Artstein, G.M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, and D.R. Traum, “The distress analysis interview corpus of human and computer interviews”, *LREC* pp. 3123–3128, 2014.

[7] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, and G. Lucas, “Simsensei kiosk: A virtual human interviewer for healthcare decision support”, *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems* pp. 1061–1068, 2014.

[8] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E. M. Messner, and S. Song, “Avec 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition”, *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, pp. 3–12, 2019.

[9] R. A. Calvo, S. D’Mello, J. M. Gratch, and A. Kappas, eds., “Cyberpsychology and affective computing”, *The Oxford Handbook of Affective Computing*, Oxford University Press, 2015.

[10] K.R. Scherer, “What are emotions? And how can they be measured?”, *Social Science Information*,

44(4), 695–729, 2005.

[11] S. Alghowinem, R. Goecke, J. Epps, M. Wagner, and J. Cohn, "Cross-cultural depression recognition from vocal biomarkers", *Interspeech 2016, ISCA* pp. 1339–1343, 2016.

[12] J. Park and N. Moon, "Design and implementation of attention depression detection model based on multimodal analysis", *Sustainability*, 14(6), p. 3569, 2022.

[13] I. Uslu, "Deep Learning im Mental Health Kontext", *Reutlingen University*, Master thesis, 2023.

[14] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer", *International Journal of YYY*, 21(1), pp.5485–5551, 2020.

[15] J. C. Cheng and A. L. P. Chen, "Multimodal time-aware attention networks for depression detection", *Journal of Intelligent Information Systems*, 59(2), pp. 319–339, 2022.

[16] E. Toto, M. Tlachac, and E. A. Rundensteiner, "Audibert: A deep transfer learning multimodal classification framework for depression screening", *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, ACM, pp. 4145–4154, 2021.

[17] A. Sharma, K. Sharma, and A. Kumar, "Real-time emotional health detection using fine-tuned transfer networks with multimodal fusion", *Neural Computing and Applications*, pp. 1–14, 2022.

[18] N. Aloshban, A. Esposito, and A. Vinciarelli, "What you say or how you say it? Depression detection through joint modeling of linguistic and acoustic aspects of speech", *Cognitive Computation*, 14(5), pp. 1585–1598, 2022.

[19] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), pp. 357–366, 1980.

[20] Z. Huang, J. Epps, and D. Joachim, "Investigation of speech landmark patterns for depression detection", *IEEE Transactions on Affective Computing*, 13(2), pp. 666–679, 2019.

[21] E. W. McGinnis, S. P. Anderau, J. Hruschak, R. D. Gurchiek, N. L. Lopez-Duran, K. Fitzgerald, K. L. Rosenblum, M. Muzik, and R. S. McGinnis, "Giving voice to vulnerable children: Machine learning analysis of speech detects anxiety and depression in early childhood", *IEEE Journal of Biomedical and Health Informatics*, 23(6), pp. 2294–2301, 2019.

[22] A. Bailey, and M. D. Plumbley "Gender bias in depression detection using audio features," *29th European Signal Processing Conference, EUSIPCO 2021 IEEE*, pp.596–600, 2021.

[23] OpenAI, "GPT-4 technical report", *arXiv preprint arXiv:2303.08774*, 2023.

[24] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, et al., "Lamda: Language models for dialog applications", *arXiv preprint arXiv:2201.08239*, 2022.

[25] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu, W. Liu, et al., "ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation", *arXiv preprint arXiv:2107.02137* , 2021.

[26] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, et al., "Llama: Open and efficient foundation language models", *arXiv preprint arXiv:2302.13971*, 2023.

[27] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, and L. P. Morency, "Automatic audiovisual behavior descriptors for psychological disorder analysis", *Image and Vision Computing*, vol. 32, no. 10, pp. 648–658, 2014.

[28] P. Lopez-Otero, and L. Docio-Fernandez, "Analysis of gender and identity issues in depression detection on de-identified speech", *Computer Speech & Language*, vol. 65, pp. 101–118, 2021.

[29] S. A. Qureshi, S. Saha, M. Hasanuzzaman, and G. Dias, "Multitask representation learning for multimodal estimation of depression level", *IEEE Intelligent Systems*, vol. 34, no. 5, pp. 45–52, 2019.

[30] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The phq-8 as a measure of current depression in the general population", *Journal of Affective Disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

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

[32] S. El Anigri, M. M. Himmi, and A. Mahmoudi, "How Bert's dropout fine-tuning affects text classification?", *Business Intelligence*, pp. 130–139, 2021.

[33] E. Villatoro-Tello, S. P. Dubagunta, J. Fritsch, G. Ramirez-de-la Rosa, P. Motlicek, and M. Magimai-Doss, "Late fusion of the available lexicon and raw waveform-based acoustic modeling for depression and dementia recognition", *Interspeech*, pp. 1927–1931, 2021.

[34] E. Villatoro-Tello, G. Ramirez-de-la Rosa, D. Gática-Pérez, M. Magimai-Doss, and H. Jiménez-Salazar, "Approximating the mental lexicon from clinical interviews as a support tool for depression detection", *Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 557–566, 2021.

[35] S. Senn, M. Tlachac, R. Flores, and E. Rundensteiner, "Ensembles of bert for depression classification", *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 4691–4694, 2022.