Exploring the Capabilities of a Language Model-Only Approach for Depression Detection in Text Data

Misha Sadeghi*[1], Bernhard Egger[2], Reza Agahi[3], Robert Richer[1], Klara Capito[4], Lydia Helene Rupp[4], Lena Schindler-Gmelch[4], Matthias Berking[4], and Bjoern M. Eskofier[1]

***Abstract*—Depression is a prevalent and debilitating mental health condition that requires accurate and efficient detection for timely and effective treatment. In this study, we utilized the E-DAIC (Extended Distress Analysis Interview Corpus-Wizard-of-Oz) dataset, an extended version of the DAIC-WOZ dataset, which consists of semi-clinical interviews conducted by an animated virtual interviewer called Ellie, controlled by a human interviewer in another room. With 275 participants, the E-DAIC dataset represents a valuable resource for investigating depression detection methods. Our aim is to predict PHQ-8 scores through text analysis. Leveraging state-of-the-art speech processing, LLM-based text summarization, and a specialized depression detection module, we demonstrate the transformative potential of language data analysis in enhancing depression screening. By overcoming the limitations of manual feature extraction methods, our automated techniques provide a more efficient and effective means of evaluating depression. In our evaluation, we achieve robust accuracy on the development set of the E-DAIC dataset, with a Mean Absolute Error (MAE) of 3.65 in estimating PHQ-8 scores from recorded interviews. This remarkable performance highlights the efficacy of our approach in automatically predicting depression severity. Our research contributes to the growing evidence supporting the use of LLMs in mental health assessment, showcasing the role of innovative technologies in advancing patient care for depression.**

***Index Terms*—Depression detection, mental health, large language models, text analysis, DAIC dataset, GPT-based models, DepRoBERTa**

I. Introduction

Depression is a widespread mental health condition with a significant global impact. It affects thoughts, behaviors, emotions, and overall well-being, impacting approximately 300 million individuals worldwide [1]. The subjectivity of the assessment process presents a significant challenge in providing effective treatment and care for depression, potentially leading to inaccurate evaluations [2]. Questionnaires, such as the Patient Health Questionnaire (PHQ) [3], heavily rely on patient responses, which may be compromised by the subjective nature of the questions. Efficient tools are crucial for practitioners who dedicate their valuable time to tasks like screening and seeking second opinions. However, relying solely on the PHQ as a screening tool can yield false positives or false negatives, hindering accurate results and challenging early depression diagnosis [2]. To address this challenge, researchers have explored behavioral indicators for automated depression detection and prediction [4], [5]. While cues like facial expressions and speech patterns have been investigated, text-based approaches leveraging advancements in natural language processing (NLP) and language models offer scalable and widely applicable depression screening. Enhancing language models allows for the automatic extraction of subtle features and patterns indicative of depression from text data.

Numerous studies on depression detection and analysis have recognized the DAIC-WOZ (Distress Analysis Interview Corpus-Wizard-of-Oz) dataset [6], [7] as a valuable and publicly available resource widely utilized in the field. Building upon this foundation, in our study, we leverage the Extended DAIC dataset (E-DAIC) [8], an extended version of the DAIC-WOZ dataset. Our objective is to demonstrate how Large Language Models (LLMs) can improve depression screening and diagnosis through text analysis. Notably, our contribution lies in the utilization of LLMs for automated feature extraction from the textual data within the dataset. This approach eliminates the need for manual cleaning of the dataset, as all processes are performed automatically. By leveraging the power of LLMs, we can extract features directly from the text, streamlining the analysis process and reducing manual intervention. The paper comprises a literature review in Section 2 and a proposed method for depression assessment in Section 3. Section 4 presents experimental results, while Sections 5 and 6 discuss the broader impact and implications, concluding the paper.

II. Related work

To improve the detection of depression and related mental health conditions, researchers in the field of affective computing have investigated various behavioral cues, such as facial expressions, speech patterns, and other multimodal signals. These cues have shown promise as potential indicators of these disorders [9]. A study by Gong et al. [10] proposed a topic modeling-based approach for context-aware analysis of long interviews in the DAIC-WOZ dataset. It captured important temporal details and outperformed context-unaware methods. Additionally, Williamson et al. [11] found that analyzing the avatar's text was the most informative indicator of dialogue content and context, avoiding sources of inter-subject variability unrelated to depression. Ray et al. [12] proposed a multi-level attention-based network to predict depression

*Responsible author; Contact: misha.sadeghi@fau.de

[1]Machine Learning and Data Analytics Lab, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91052 Erlangen, Germany.

[2]Chair of Visual Computing, Department of Computer Science, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91058 Erlangen, Germany.

[3]Syenah GMBH, 65760 Eschborn, Germany.

[4]Chair of Clinical Psychology and Psychotherapy, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91052 Erlangen, Germany.

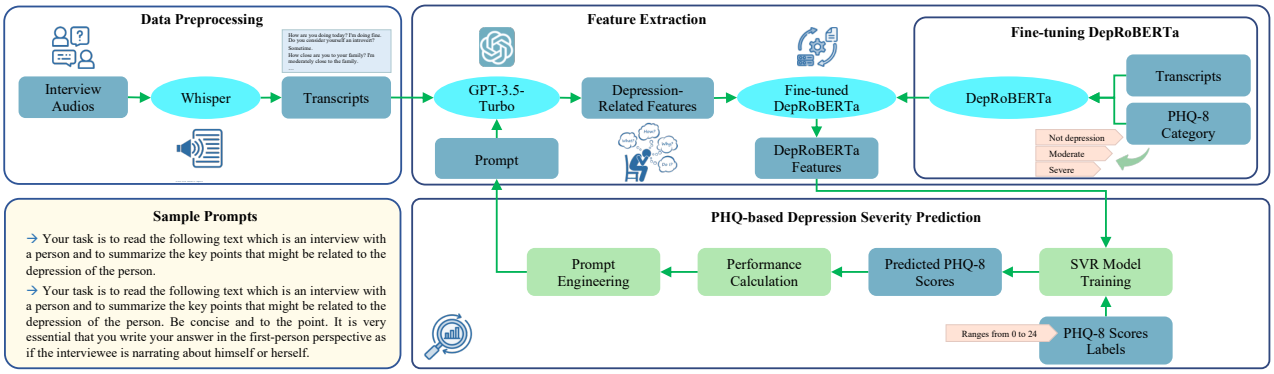

Fig. 1. Proposed framework overview: The figure presents the key blocks of our approach. The Data processing block extracts transcripts using the Whisper automatic speech recognition system. The Feature extraction block utilizes GPT-3.5-Turbo and a fine-tuned DepRoBERTa model. The Training process block encompasses PHQ-8 score prediction, method evaluation, and prompt engineering. The final block highlights the two best prompts for conclusive evaluation.

severity using audio, video, and text data in Audio/Visual Emotion Challenge and Workshop (AVEC) 2019 [8]. The text modality was assigned the highest weight, while the audio and video modalities were given equal weight. Stepanov et al. [13] used speech, language, and visual features to predict PHQ-8 scores on the DAIC-WOZ dataset. They analyzed visual features using 3D facial key points and computed Euclidean distances between normalized points. Yin et al. [14] based on the DAIC-WOZ dataset, predicted the PHQ-8 score using video, audio, and text features with two hierarchies of bidirectional LSTM networks. The authors used the adaptive sample weighting mechanism in AdaBoost to adjust the weight of the samples. Shen et al. [15] tackled the depression sub-challenge problem in AVEC 2017 using multiple modalities. Their findings showed that spectral features in the audio domain and behavioral cues extracted from transcripts were the most effective predictors of depression severity scores. Similarly, Nasir et al. [16] found that the i-vector system worked best for audio, and polynomial parameterization of facial landmarks combined with geometrical features was the best for video. Cohn et al. [5] used manual Facial Action Coding System (FACS) [17], active appearance modeling, and pitch extraction to measure facial and vocal expressions. The results showed that both facial and vocal expressions had moderate correlations with depression. In a different study, Prabhu et al. [18] conducted a study on depression detection using multiple modalities on the DAIC-WOZ dataset. They employed a CNN-LSTM model for facial expressions, an LSTM model for text, and another LSTM model for audio as well as transfer learning to enhance performance. Ensemble techniques were used to combine predictions from different modalities.

III. Methods

A. Dataset

In this paper, we used the E-DAIC dataset, an extended version of the DAIC-WOZ dataset, to investigate our text-based approach for detecting depression. The E-DAIC dataset includes semi-clinical interviews conducted by a virtual interviewer named Ellie. These interviews aim to identify verbal and nonverbal indicators of mental illness. They are conducted in a "Wizard-of-Oz" setting, where the virtual agent is controlled either by a human interviewer or an AI-controlled agent. The interviews, lasting around 20 minutes, are transcribed and annotated with acoustic and visual features. The dataset comprises interviews with 275 participants (170 males, 105 females), divided into train, development (dev), and test sets with 163, 56, and 56 instances, respectively. Careful consideration was given to speaker diversity, including factors like age, gender distribution, and PHQ-8 scores [6], [7]. The PHQ-8 score measures depression severity using eight items scored on a scale of 0 to 3. Total scores range from 0 to 24, with higher scores indicating greater depression severity [3].

B. Automated Depression Assessment through Text Analysis and LLMs

The general framework of our proposed approach is illustrated in Figure 1. In the following subsections, we will delve into the specifics of each pipeline of the framework.

1) Data Preprocessing: The DAIC dataset is valuable for mental health research, but it has limitations due to incomplete transcripts from automatic speech-to-text extraction. This results in a lack of contextual information and important details being omitted. For example, questions such as "Have you ever been diagnosed with depression?" or "Have you ever been diagnosed with Post-traumatic Stress Disorder (PTSD)?" may only have a simple "yes" or "no" answer, without any indication of the question being asked. To address this problem, we employed the Whisper automatic speech recognition system [19] from OpenAI to extract transcripts from raw interview audios. This approach has been shown to be highly effective, allowing us to obtain transcripts for the entire interview with a high degree of accuracy [19]. Our assessment of several interview audios revealed that the Whisper "large" model consistently outperformed the "base" model in terms of accuracy and overall transcript quality. Whisper API surpasses the E-DAIC dataset's speech-to-text extraction, providing con-

siderably higher quality transcripts, thereby overcoming E-DAIC limitations and enabling comprehensive analysis with accurate and complete transcripts.

2) Feature Extraction: The small dataset size in this study poses a challenge for achieving high performance. To address this, we need to extract informative and distinctive features that yield the best results. While multi-modal approaches are generally effective, identifying discriminative features from a single modality (text, video, or audio) may lead to superior outcomes. Careful selection of relevant features is crucial to capture the critical aspects of the data, as having numerous features can lead to overfitting. Therefore, selecting a few significant features that provide valuable information is essential for optimal performance. To address this issue, our study proposes a two-part approach to detect depression severity. The first part focuses on exploring the transformation of transcripts to enhance their usefulness for depression detection. In the second part, we employ a pre-trained language model to analyze the transformed transcripts and predict an individual's PHQ score based on the text.

a) Transformation of Transcripts using GPT-3.5-Turbo: The first part of our approach involves transforming the interviews to make them more informative for detecting depression. To achieve this, we utilized GPT-3.5-Turbo [20], a state-of-the-art NLP model developed by OpenAI. In our method, we started by creating a prompt ourselves and then used GPT-3.5-Turbo for suggestions on alternative prompts. We calculated errors based on these prompts and selected the two best ones. These prompts served as valuable tools for transforming the transcripts, effectively highlighting depression-relevant features and extracting crucial information from the interviews. Through this transformation, we enhanced the accuracy of depression detection by emphasizing the patterns that are most relevant to the condition. To ensure optimal results, we discovered the importance of employing concise and focused prompts, which enable a more targeted analysis. Figure 1 illustrates two highly effective prompts that emerged from our careful evaluation process, taking into consideration the suggestions provided by GPT-3.5-Turbo. To select these two prompts, we conducted a performance evaluation based on their utilization. Error metrics, including Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), were calculated to measure the effectiveness of the approach. The choice of prompts was guided by the results of these error calculations.

b) Language Models for Depression Detection: In the second phase of our approach, we employ a language model for analyzing the transformed transcripts and predicting an individual's PHQ score. Specifically, we utilize a fine-tuned RoBERTa language model [21] called DepRoBERTa (RoBERTa for Depression Detection) [22]. DepRoBERTa is a specialized model designed to extract relevant features from text and detect depression. It is based on RoBERTa-large and was pre-trained on depressive posts from Reddit. The model, known as "deproberta-large-depression", has demonstrated remarkable performance in identifying the level of depression in English social media posts [22]. It can detect three different levels of depression: "not depression", "moderate", and "severe". The model was developed as part of the winning solution for the Shared Task on Detecting Signs of Depression from Social Media Text at LT-EDI-ACL2022 and is available on the Hugging Face model hub [23]. Domain experts labeled the training dataset used for this model, assigning one of the three labels to each instance. The label "not depression" indicates the absence of depression signs, while the other two labels indicate the presence of moderate or severe depression symptoms, respectively [22].

In our approach, we used the DepRoBERTa model to extract three distinct features ("not depression", "moderate", and "severe") from text data. These features provide a measure of depression severity on a scale of 0 to 1. To achieve this, we fine-tuned the DepRoBERTa model using a low learning rate (lr = 5e-6) on the transformed transcriptions from the E-DAIC dataset. The transcripts were labeled according to the original categories of the DepRoBERTa model. Assigning transcripts to each category was based on their PHQ scores, where PHQ scores of 14 or higher were labeled as "severe", scores between 7 and 13 as "moderate", and scores of 7 or lower as "not depression". To fine-tune the Deproberta model, we trained it on the train set and monitored the loss on the dev set. Early stopping was employed to prevent overfitting. After fine-tuning, we used the model to extract the three features from each text instance in our dataset. Subsequently, a Support Vector Regression (SVR) machine learning model was trained using the PHQ scores as the target variable. This model utilizes the extracted features to predict the PHQ score in each text instance. The model's hyperparameters were optimized using GridSearchCV, with MAE serving as the optimization metric. For the SVR model, we trained it on the train set and evaluated its performance on the dev set and test set. This allows us to compare our results with previous works.

IV. Results and Discussion

Table 1 presents a comprehensive comparison of methods employed in the studies on the DAIC and EDAIC datasets. The majority of these studies participated in the AVEC challenges, with a few conducted independently. Since our study focuses on PHQ_Score prediction, we specifically looked at other studies that address the same problem. Different studies consider different data modalities, including text, video, and/or audio, or a combination of them. In Table 1, we ordered papers according to the modalities they considered. Most studies calculated the errors on the dev set, while some also considered the test set. As utilizing a method capable of automatically extracting features from transcripts is crucial in order to streamline the analysis process and ensure efficiency, the final column of the table indicates whether manual processing of the transcripts was conducted or not. Additionally, in line with our study, only the studies indicated with a light gray highlight in the table made use of the updated version of the DAIC dataset (E-DAIC).

In Proposed Method 1, we chose not to fine-tune the DepRoBERTa model. Instead, we solely extracted the embed-

TABLE I
COMPARATIVE ANALYSIS OF PHQ SCORE PREDICTION MODELS BASED ON THE DAIC (OR E-DAIC) DATASET.

Model	Modality[1]	MAE (dev)	RMSE (dev)	MAE (test)	RMSE (test)	Manual Proc.[2]
Williamson et al. [11]	T+A+V	4.18	5.3	-	-	Yes
Yang et al. [24]	T+A+V	3.98	4.65	5.163	5.97	Yes
Yang et al. [25]	T+A+V	2.48	3.09	4.36	5.4	Yes
Rohanian et al. [26]	T+A+V	-	-	3.61	4.99	Yes
Qureshi et al. [27]	T+A+V	3.07	4.14	-	-	Yes
Ray et al. [12]	T+A+V	-	4.28	-	-	Yes
Yin et al. [14]	T+A+V	-	4.94	-	5.50	No
Song et al. [28]	T+A+V	5.95	7.15	-	-	No
Niu et al. [29]	T+A	2.94	3.80	-	-	Yes
Al Hanai et al. [30]	T+A	4.97	6.27	-	-	No
Oureshi et al. [31]	T+V	3.49	-	-	-	Yes
Nasir et al. [16]	A	5.82	6.73	-	-	No
Stepanov et al. [13]	A	-	-	4.11	4.94	No
Niu et al. [32]	A	-	-	7.48	9.79	No
Niu et al. [29]	A	3.45	4.91	-	-	No
Nasir et al. [16]	V	6.48	7.87	-	-	No
Stepanov et al. [13]	V	-	-	5.36	6.72	No
Williamson et al. [11]	T	**3.34**[3]	**4.46**	-	-	Yes
Gong et al. [10]	T	**2.77**	**3.54**	**3.96**	**4.99**	Yes
Stepanov et al. [13]	T	-	-	4.88	5.83	Yes
Ray et al. [12]	T	-	**4.37**	**4.02**	**4.73**	Yes
Oureshi et al. [31]	T	3.78	-	-	-	Yes
Niu et al. [29]	T	3.73	**4.80**	-	-	Yes
Rohanian et al. [26]	T	-	-	4.98	6.05	No
Al Hanai et al. [30]	T	5.18	6.38	-	-	No
Proposed Method 1	T	3.78	**5.23**	4.79	6.05	No
Proposed Method 2	T	-	-	4.41	5.44	No
Proposed Method 3	T	**3.65**	5.27	**4.26**	**5.36**	No

[1] T: Text, A: Audio, and V: Video.
[2] Manual Processing: Yes if authors performed manual cleaning of the transcripts.
[3] Bold fonts indicate our best results and results from studies that focused on text modality and outperformed our method.
- Light gray rows use the E-DAIC dataset, while others use the DAIC dataset.
- Dark gray rows are the only previous studies on text data without manual transcript processing.

dings from its last layer. This approach resulted in an MAE of 3.78 on the dev set. In Proposed Method 2, we performed fine-tuning of the DepRoBERTa model on the entire train set and dev set using an lr of 5e-5. Our objective was to incorporate more data during model training, considering the limited size of the train set, in order to improve the results on the test set. This approach led to an MAE of 4.41 and an RMSE of 5.44, which are superior to the results obtained in Proposed Method 1. In Methods 1 and 2, we employed the prompt "Your task is to read the following text which is an interview with a person and to summarize the key points that might be related to the depression of the person". In Proposed Method 3, we utilized the other prompt: "Your task is to read the following text which is an interview with a person and to summarize the key points that might be related to the depression of the person. Be concise and to the point. It is very essential that you write your answer in the first-person perspective as if the interviewee is narrating about himself or herself". With this prompt, we achieved the best MAE results, specifically an MAE of 3.65 on the dev set and 4.26 on the test set. To accomplish this, we fine-tuned the DepRoBERTa model using an lr of 5e-6 and incorporated the polynomial (poly) kernel.

As depicted in Table 1, among the studies focusing solely on text analysis, the two highlighted rows in dark gray (Rohanian et al. [26] and Al Hanai et al. [30]) stand out as the only ones that did not employ any manual processing of the transcripts. Notably, our results outperformed theirs; however, it is worth mentioning that they utilized the previous version of the DAIC dataset. Table 1 also demonstrates that Gong et al. [10] achieved the highest performance in all error parameters except RMSE on the test set. However, their method involved a substantial amount of manual processing on the data. They manually cleaned the transcripts, extracted discussed topics, and created interview questions. They also manually clustered the context into different topics. As previously mentioned, the transcripts have missing parts, which require significant effort to clean and extract useful information manually. Furthermore, the dataset used in their paper is an older version of DAIC and differs from the E-DAIC dataset. Therefore, our results cannot be directly compared to theirs. Additionally, Williamson et al. [11] also achieved notable performance on the dev set by adopting an approach that involved extensive manual cleaning of the transcripts and extracting question-answer pairs manually. They also separately extracted the most critical questions, such as those related to diagnoses of depression and attendance of therapy sessions, along with their corresponding answers. They additionally incorporated non-verbal cues such as laughter, coughing, sighing, and deep breathing, and filled the pauses in the interview with corresponding non-verbal cues. This approach allowed them to analyze beyond just text data and consider other important aspects of communication. In addition, their analysis was also based on an older version of the DAIC dataset, making it difficult to directly compare their results with ours. Among the other studies that considered text modality, only Ray et al. [12] achieved a slightly better MAE on the test set compared to ours. They also reported lower RMSE on both dev and test sets. However, they as well performed manual cleaning of the transcripts, although the extent of their cleaning process was not explicitly specified. Niu et al. [29] also achieved one of the best results in RMSE (dev). However, they also performed manual cleaning of the transcripts and extracted question-answer pairs manually.

Our framework, which sets itself apart from previous works, stands out due to its unique approach of automatically extracting all features from the transcripts without any manual cleaning. This novel feature distinguishes our work from most of the previous studies in the field. By employing this methodology on the E-DAIC dataset, we make a significant contribution to the field. While our preliminary results demonstrate promising outcomes, there remain ample opportunities for further exploration. Future endeavors can concentrate on incorporating additional data modalities to gain deeper insights into the speaker's emotional state and behavior during the interview, thereby enhancing the accuracy of predictions. Additionally, to further refine our approach, an ensemble implementation of different prompts could prove advantageous. By combining the insights and predictions generated from a variety of prompts, we can leverage the strengths of each prompt and potentially achieve more precise predictions. This avenue of investigation holds promise for enhancing the overall performance of our framework and advancing the field in a meaningful way.

V. BROADER IMPACT STATEMENT

The impact of current AI technology, particularly LLMs, is a topic of ongoing discussion, both within the broader public and the research community. These tools possess great capabilities, as demonstrated in this study, where we utilized the publicly available and fully anonymized E-DAIC dataset for depression detection. We understand the significance of

handling data responsibly and obtaining consent to ensure ethical research practices. This study highlights a potential application of LLMs in depression detection. However, it is crucial to recognize that the capabilities and limitations of these tools are not yet fully understood, and thus, a careful approach to their integration is necessary. It is important to emphasize that while such a tool may be valuable for screening or assisting practitioners, it is not intended to replace human involvement in the process. Therefore, it is essential to approach the integration of LLMs with careful consideration of their ethical implications and potential biases.

VI. Conclusion and Outlook

In conclusion, our study demonstrates the potential of LLMs to assess depression through text analysis, setting us apart from most other studies that rely on manual feature extraction. By leveraging LLMs and focusing solely on text data, we improve the accuracy and efficiency, defined as reduced manual work, of PHQ score prediction on the E-DAIC dataset compared to previous traditional manual feature extraction methods. The results indicate that our LLM-based method outperforms alternative techniques, as evidenced by lower RMSE and MAE values. Future research can further enhance accuracy by exploring the incorporation of other modalities and diverse datasets. Overall, our study contributes to the growing evidence supporting the use of LLMs in mental health assessment and highlights the significance of innovative technologies in improving patient care for depression.

Acknowledgment

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1483 – Project-ID 442419336, EmpkinS.

References

[1] M. Carey *et al.*, "Accuracy of general practitioner unassisted detection of depression," *Australian & New Zealand Journal of Psychiatry*, vol. 48, no. 6, pp. 571–578, 2014.

[2] D. American Psychiatric Association *et al.*, *Diagnostic and statistical manual of mental disorders: DSM-5*. American psychiatric association Washington, DC, 2013, vol. 5, no. 5.

[3] K. Kroenke *et al.*, "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.

[4] M. Valstar *et al.*, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 3–10.

[5] J. F. Cohn *et al.*, "Detecting depression from facial actions and vocal prosody," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–7.

[6] J. Gratch *et al.*, "The distress analysis interview corpus of human and computer interviews," University of Southern California Los Angeles, Tech. Rep., 2014.

[7] D. DeVault *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 2014, pp. 1061–1068.

[8] F. Ringeval *et al.*, "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, 2019, pp. 3–12.

[9] S. Scherer *et al.*, "Automatic audiovisual behavior descriptors for psychological disorder analysis," *Image and Vision Computing*, vol. 32, no. 10, pp. 648–658, 2014.

[10] Y. Gong and C. Poellabauer, "Topic modeling based multi-modal depression detection," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 69–76.

[11] J. R. Williamson *et al.*, "Detecting depression using vocal, facial and semantic communication cues," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 11–18.

[12] A. Ray *et al.*, "Multi-level attention network using text, audio and video for depression prediction," in *Proceedings of the 9th international on audio/visual emotion challenge and workshop*, 2019, pp. 81–88.

[13] E. A. Stepanov *et al.*, "Depression severity estimation from multiple modalities," in *2018 ieee 20th international conference on e-health networking, applications and services (healthcom)*. IEEE, 2018, pp. 1–6.

[14] S. Yin *et al.*, "A multi-modal hierarchical recurrent neural network for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 65–71.

[15] Y. Shen *et al.*, "Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6247–6251.

[16] M. Nasir *et al.*, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 43–50.

[17] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[18] S. Prabhu *et al.*, "Harnessing emotions for depression detection," *Pattern Analysis and Applications*, pp. 1–11, 2022.

[19] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[20] OpenAI, "Openai gpt-3.5 model documentation," Accessed: Apr. 25, 2023. [Online]. Available: https://platform.openai.com/docs/models/gpt-3-5

[21] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[22] R. Poświata and M. Perełkiewicz, "Opi@ lt-edi-acl2022: Detecting signs of depression from social media text using roberta pre-trained language models," in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 2022, pp. 276–282.

[23] "deproberta-large-depression," Hugging Face, Accessed: Apr. 25, 2023. [Online]. Available: https://huggingface.co/rafalposwiata/deproberta-large-depression

[24] L. Yang *et al.*, "Multimodal measurement of depression using deep learning models," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 53–59.

[25] L. Yang *et al.*, "Hybrid depression classification and estimation from audio video and text information," in *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, 2017, pp. 45–51.

[26] M. Rohanian *et al.*, "Detecting depression with word-level multimodal fusion." in *Interspeech*, 2019, pp. 1443–1447.

[27] S. A. Qureshi *et al.*, "The verbal and non verbal signals of depression–combining acoustics, text and visuals for estimating depression level," *arXiv preprint arXiv:1904.07656*, 2019.

[28] S. Song *et al.*, "Spectral representation of behaviour primitives for depression analysis," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 829–844, 2020.

[29] M. Niu *et al.*, "Hcag: A hierarchical context-aware graph attention model for depression detection," in *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2021, pp. 4235–4239.

[30] T. Al Hanai *et al.*, "Detecting depression with audio/text sequence modeling of interviews." in *Interspeech*, 2018, pp. 1716–1720.

[31] S. A. Oureshi *et al.*, "Gender-aware estimation of depression severity level in a multimodal setting," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[32] M. Niu *et al.*, "Automatic depression level detection via lp-norm pooling," *Proc. INTERSPEECH, Graz, Austria*, pp. 4559–4563, 2019.