

DAIC-WOZ: On the Validity of Using the *Therapist’s prompts* in Automatic Depression Detection from Clinical Interviews

Sergio Burdisso^{*,1}, Ernesto A. Reyes-Ramírez², Esaú Villatoro-Tello^{*,1},
Fernando Sánchez-Vega^{2,3}, A. Pastor López-Monroy² and Petr Motlicek^{1,4}

¹Idiap Research Institute, Martigny, Switzerland

²Mathematics Research Center (CIMAT), Gto, Mexico

³Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT), México

⁴Brno University of Technology, Brno, Czech Republic

Abstract

Automatic depression detection from conversational data has gained significant interest in recent years. The DAIC-WOZ dataset, interviews conducted by a human-controlled virtual agent, has been widely used for this task. Recent studies have reported enhanced performance when incorporating interviewer’s prompts into the model. In this work, we hypothesize that this improvement might be mainly due to a bias present in these prompts, rather than the proposed architectures and methods. Through ablation experiments and qualitative analysis, we discover that models using interviewer’s prompts learn to focus on a specific region of the interviews, where questions about past experiences with mental health issues are asked, and use them as discriminative shortcuts to detect depressed participants. In contrast, models using participant responses gather evidence from across the entire interview. Finally, to highlight the magnitude of this bias, we achieve a 0.90 F1 score by intentionally exploiting it, the highest result reported to date on this dataset using only textual information. Our findings underline the need for caution when incorporating interviewers’ prompts into models, as they may inadvertently learn to exploit targeted prompts, rather than learning to characterize the language and behavior that are genuinely indicative of the patient’s mental health condition.

1 Introduction

Recent advances in Artificial Intelligence (AI) have increased the existing enthusiasm among medical professionals and clinicians when considering the potential for AI-based solutions to make mental healthcare more accessible and to reduce the burden of psychiatric institutions (Passos et al., 2023). This possibility has led some psychiatrists to argue

that the use of AI might result in more standardized and objective measures of mental health (Pendse et al., 2022).

Consequently, the automatic analysis of clinical interviews has been recognized as a promising direction for the development of automatic solutions that will help to improve the diagnostic consistency of depression detection (Tao et al., 2023; Zou et al., 2022; Burdisso et al., 2019; Valstar et al., 2016). The Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) dataset (Gratch et al., 2014) stands out as the most representative multimodal resource which has been commonly used for training and validating depression classification models within a clinical setup. Most existing studies leverage the participant answers for depressive assessment, varying from single-modality methods, i.e., text transcripts, speech (Burdisso et al., 2023; Villatoro-Tello et al., 2021a; Xezonaki et al., 2020; Mallol-Ragolta et al., 2019), to multi-modal approaches (text + speech + video) (Zhuang et al., 2024; Fang et al., 2023; Shen et al., 2022; Yoon et al., 2022; Villatoro-Tello et al., 2021b). However, recent studies that incorporate therapist’s prompts during training, argue that such information works as supplementary context to better extract salient cues from participant answers (Zhuang et al., 2024; Shen et al., 2022; Niu et al., 2021; Dai et al., 2021), reporting high classification performances.

In this paper, we investigate the validity of using the interviewer’s prompts from the DAIC-WOZ dataset in automatic depression detection scenarios. We hypothesize that the reported results using both interviewer and participant information may be artificially inflated by a bias induced by the interviewer, failing to generalize to real-world scenarios where such biases may not exist. The impact of over-reporting performance in the DAIC-WOZ dataset has been already pointed by (Bailey and Plumbley, 2021) due to the presence of gender bias. Nevertheless, and to the best of our knowledge, this

^{*}Corresponding authors.

{sergio.burdisso, esau.villatoro}@idiap.ch

is the first work to report the existence of a strong bias in the interviewer’s prompts and to show that models can effectively exploit it as discriminative shortcuts.

2 The DAIC-WOZ Dataset

The DAIC-WOZ dataset contains clinical interviews in North American English, performed by an animated virtual (human-controlled, i.e., Wizard of OZ) interviewer, called Ellie, designed to support the diagnosis of different psychological distress conditions. The DAIC-WOZ stands as a valuable resource frequently utilized by the NLP community, attributed to its rigorous data collection methods and the scarcity of newer data sources exploring comparable phenomena. DAIC-WOZ is a multi-modal corpus, composed by audio and video recordings, and transcribed text from the interviews. To the date, the DAIC-WOZ corpus represents a unique and valuable resource, accumulating over 1K citations since its release.¹

Ellie conducts semi-structured interviews that are intended to create interactional situations favorable to the assessment of distress indicators correlated with depression, anxiety or post-traumatic stress disorder (PTSD). Theoretically, the advantage of Ellie over a human interviewer is the implicit replicability and consistency of the prompts and accompanying gestures. Thus, Ellie has a finite repertoire of 191 prompts, i.e., general questions (*what are you like when you don’t get enough sleep?*), neutral backchannels (*uh huh*), positive empathy (*that’s great*), negative empathy (*i’m sorry*), surprise responses (*wow!*), continuation prompts (*could you tell me more about that?*), and miscellaneous prompts (*don’t know; thank you*). Table 1 shows a few statistics from the dataset.²

3 Methodology

To assess the reliability of using Ellie’s prompts for automatic depression detection on DAIC-WOZ, we first examine some of the highest results reported in the recent past using this dataset, summarized in Table 2. We can categorize published works into two primary groups: (a) those using solely the participant (*P*) responses and, (b) those incorporating Ellie’s (*E*) prompts to the model. It seems that

¹Rough estimation based on the citation counts of (Gratch et al., 2014; DeVault et al., 2014) in Google scholar.

²Labels of the test set are not publicly available due to the AVEC competition (Valstar et al., 2016).

Speaker	Partition	Voc. size	Avg. #words	Avg. #tokens
Ellie (<i>E</i>)	<i>train</i>	232	190.3 (<i>sd</i> =26.9)	567.2 (<i>sd</i> =79.10)
	<i>eval</i>	216	184.8 (<i>sd</i> =50.2)	540.7 (<i>sd</i> =148.5)
Participant (<i>P</i>)	<i>train</i>	5858	621.1 (<i>sd</i> =326.2)	1606.2 (<i>sd</i> =893.9)
	<i>eval</i>	3268	664.2 (<i>sd</i> =281.7)	1756.3 (<i>sd</i> =814.7)

Table 1: DAIC-WOZ contains 107 training files (77 control [C] and 30 depressed [D]), an evaluation set of 35 files (23 [C] and 12 [D]). Table shows the vocabulary size and the average interview length measure in words and *WordPiece* tokens, with its corresponding standard deviation (*sd*) values.

works from group (b) exhibit an overall superior performance compared to those of group (a). To investigate whether this improvement may stem from a bias in Ellie’s prompts, before delving into a qualitative analysis, we proposed an initial ablation experiment. Concretely, we evaluated two versions of the same models: one employing only participant responses and another solely using Ellie’s prompts. Subsequently, we assess the performance difference between these versions, aiming to quantify the challenge in identifying depressed subjects based on participant responses versus Ellie’s prompts. Furthermore, we tested an ensemble approach to measure how complementary these two aspects are to each other.

In particular, we will conduct an ablation experiment using two models: a strong BERT-based baseline model and the Graph Convolutional Network (GCN) model described in Burdisso et al. (2023), which is the best-performing model that relies solely on the participant’s text (see Table 2). The choice of these two models aims to compare the baselines against the best-performing model, as well as to analyze models with different natures, namely a bidirectional sequential model and a sequence-agnostic one. Moreover, as will be described below, the GCN model has an attractive interpretability property that we will use in Section 5 for the qualitative analysis. Thus, by analyzing the differences between these two models, we can determine whether the observed patterns hold independently of the model’s nature. The models are described as follows:

- **LongBERT:** a BERT-based classification model. More precisely, we used a pre-trained BERT-based Longformer (Beltagy et al., 2020) model with a final linear layer added to classify the input using the encoding of the special *[CLS]* token, following common practice. The choice of using the

Longformer variant of BERT (Devlin et al., 2019), instead of the standard Transformer (Vaswani et al., 2017) version, stems from the fact that most interviews in DAIC-WOZ are long documents exceeding the 512 token limit (see Table 1).

• **GCN:** The two-layer Graph Convolutional Network (GCN) described in Burdisso et al. (2023) that uses two types of nodes to characterize the interviews: word nodes and participant nodes. In this graph, nodes are represented at three distinct levels: one-hot encoded vectors, embeddings in a latent space (after applying the first convolution), and in a two-dimensional “output space,” (after the second convolution) where each dimension corresponds to the probability of belonging to the depression or the control group. Note that since the two type of nodes are represented in the same space, this last learned representation contains probabilities not only for the participants but also for *all the words*. This is an attractive quality of the model that allows us to track down Ellie’s bias to particular subset of words and prompts (as described in Section 5).

4 Experiments and Results

We trained and evaluated two variants of the GCN: one exclusively using the participant’s responses as in the original paper (Burdisso et al., 2023), denoted as *P*-GCN, and another one solely using Ellie’s prompts, referred to as *E*-GCN. Similarly, we also fine-tuned and evaluated the same two versions of the Longformer BERT model, referred to as *P*-longBERT and *E*-longBERT, respectively.³ Table 2 shows the obtained results. When using only the participant responses, *P*-GCN achieved a similarly high F1 score (0.85) to the score reported in the original paper (0.84), and *P*-longBERT a score (0.72) similar to other published works employing solely participant data (e.g. 0.69). On the other hand, when using Ellie, both *E*-GCN and *E*-longBERT achieve comparably higher F1 score. Notably, *E*-longBERT, by simply utilizing Ellie’s prompts, managed to achieve the same score (0.84) as the original GCN paper, and the *E*-GCN outperformed all main previously published works that solely rely on textual input, with a score of 0.88. This suggests that when employing Ellie’s prompts, the depression and control groups become more easily distinguishable. For instance, the F1

³Details are provided in Appendix A. Source code to replicate our study available at https://github.com/idiap/bias_in_daic-woz.

Model	Source			F ₁ score		
	<i>P</i>	<i>E</i>	<i>M</i>	Avg.	<i>D</i>	<i>C</i>
Mallol-Ragolta et al. (2019)	✓			0.60	-	-
Xezonaki et al. (2020)	✓			0.69	-	-
Villatoro-Tello et al. (2021a)	✓			0.64	0.52	0.77
Burdisso et al. (2023)	✓			0.84	0.80	0.89
Williamson et al. (2016)	✓	✓		0.84	-	-
Toto et al. (2021)	✓	✓		0.86	-	-
Shen et al. (2022)	✓	✓		0.83	-	-
Milintsevich et al. (2023)	✓	✓		0.80	-	-
Agarwal and Dias (2024)	✓	✓		0.77	-	-
Niu et al. (2021)	✓	✓	✓	0.92	-	-
Dai et al. (2021)	✓	✓	✓	0.96	-	-
Shen et al. (2022)	✓	✓	✓	0.85	-	-
Zhuang et al. (2024)	✓	✓	✓	0.88	0.85	0.91
<i>P</i> -longBERT	✓			0.72	0.64	0.80
<i>E</i> -longBERT		✓		0.84	0.80	0.89
<i>P</i> -longBERT \wedge <i>E</i> -longBERT	✓	✓		0.79	0.70	0.88
<i>P</i> -GCN			✓	0.85	0.81	0.88
<i>E</i> -GCN			✓	0.88	0.85	0.91
<i>P</i> -GCN \wedge <i>E</i> -GCN	✓	✓		0.90	0.87	0.94

Table 2: Main previously published results on DAIC-WOZ evaluation set along with our obtained results. Performance is reported in terms of the F₁ score for both control (*C*) and depression (*D*) classes, as well as their macro average (*Avg.*). Results are marked with the source data used: (*P*) and (*E*) text from the participant and Ellie; (*M*) multimodal, e.g., speech and video. The global-best result among models using only textual content is **underlined**, while the best results in each group is highlighted in **bold**.

score of the *longBERTs* for the depression group (*D*) improves from 0.64 to 0.80 when using Ellie’s prompts.

Finally, we performed a simple voting ensemble between the two variants of each model, denoted using the “and” symbol (\wedge). Participants are classified as positive (*i.e.*, in the depression group) only when both variants, Ellie *and* Participant, classify them as positive. As shown in Table 2, the ensemble approach enables the GCN-based model to achieve a remarkable F1 score of 0.90, the highest reported score to date among models exclusively utilizing textual content. These results suggest that the integration of both Ellie and participant content could be complementary for certain models, further exploiting Ellie’s bias to make the depression and control groups even more easily distinguishable.

5 Analysis and Discussion

Overall, experimental results suggest that Ellie’s prompts contain information that the models can exploit to more easily classify the participants. This

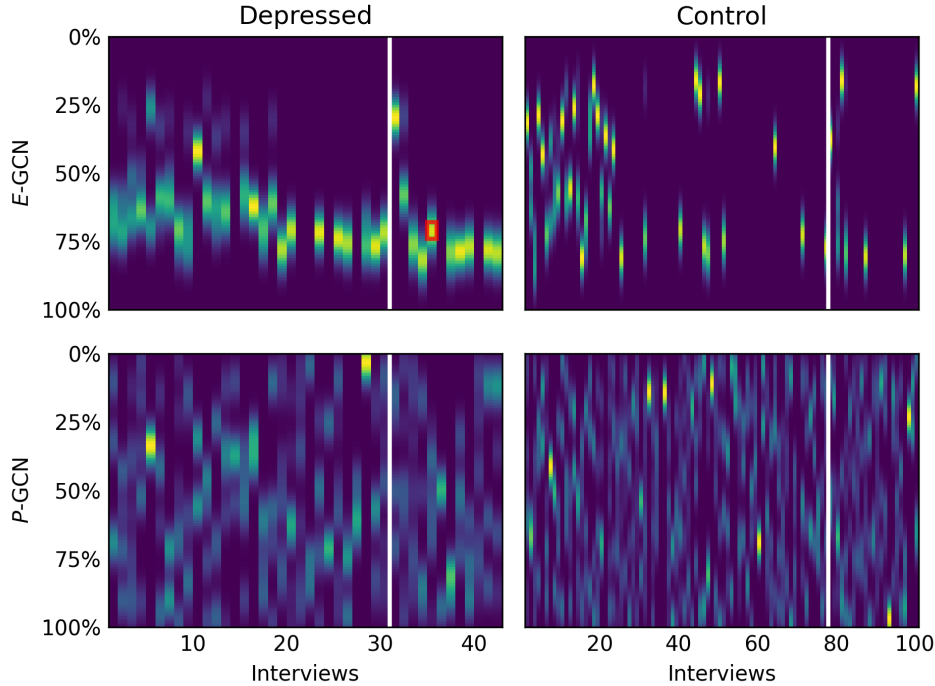


Figure 1: Heatmaps illustrating the distribution of learned keywords by each model across the progression of each interview. The x-axis represents individual interviews, while the y-axis denotes the percentage of the conversation from the beginning (0%) to the end (100%). The white vertical line in each plot indicates the training and evaluation splits respectively. Finally, in the *E-GCN* evaluation split region, the small red rectangle depicts the interview segment showed in Fig. 2.

is reasonable when considering that therapists adjust their questioning patterns based on the subjects’ responses and may adapt their inquiries to delve deeper into specific aspects *when detecting potential depressive symptoms*.

To explore this possibility further, as mentioned in Section 3, we leveraged the GCN-based model’s ability to learn a common representation for both participant and word nodes in the same output space. Firstly, we extracted the words that both GCN models learned to use to identify the depressed group, which we will refer to as keywords.⁴ Subsequently, we analyzed the distribution of these keywords throughout the progression of each interview to contrast the depressed group against the control group, allowing us to visualize how easily distinguishable the two groups are from the perspectives of both Ellie (*E-GCN*) and the participant (*P-GCN*) models. Figure 1 illustrates the distributions obtained from our analysis, highlighting the contrasting behavior of the *E-GCN* and *P-GCN* models. The *P-GCN* distribution exhibits variability across interviews, with no distinct pat-

tern emerging from the distribution of keywords. In contrast, the *E-GCN* model displays a clear and consistent pattern, with concrete regions where keywords concentrate. That is, the participant model gathers evidence from various parts of the conversations, whereas Ellie’s model focuses mainly on very specific segments, *i.e.* specific questions, to classify the participants. Furthermore, by contrasting the distributions for the depressed group against the control group, we observe that it is easier to distinguish between them using *E-GCN* than *P-GCN*. This suggests that Ellie’s keywords are not only more localized but also possess greater discriminatory power. Note that for *E-GCN*, in contrast with the control group, almost all the interviews in the depressed group have colored regions, and they are mostly concentrated in a single segment that appears *after halfway the interviews*.⁵ Interestingly, most of these segments correspond to a phase in the interview where Ellie begins to ask more personal questions about past experiences with mental

⁴Words w such $P(\text{depressed} | w) > P(\text{-depressed} | w)$

⁵As shown in Table A2, to validate this observation further, we fine-tuned *E-longBERT* on the second half of interviews, achieving 0.84 F1 (same as full interviews). Using only the first half dropped F1 to 0.60, highlighting the importance of this latter portion.

Participant: it's hard to say that's it's down when it's consistently that way
 Ellie: yeah
 Ellie: have you been diagnosed with depression
 Participant: yes
 Ellie: how long ago were you diagnosed
 Participant: a long
 Participant: oh a long time ago i was um
 Participant: about ten years old or
 Participant: around there about
 Ellie: what got you to seek help
 Participant: uh i was just missing a lot of school
 Participant: i didn't wanna go to school so eventually uh they sent me to
 Participant: to a therapist and then to a psychologist and then to a psychiatrist and uh
 Participant: it went from there
 Ellie: do you still go to therapy now
 Participant: no i haven't gone to therapy
 Participant: since i was a teenager
 Ellie: why did you stop
 Participant: well
 Participant: i didn't feel it was helping me at all
 Participant: so
 Participant: one day i just
 Participant: decided i didn't wanna go anymore
 Participant: eh
 Ellie: how has seeing a therapist affected you
 Participant: uh well like i said i didn't feel that it helped me um
 Participant: so i guess it i don't know that it affected me much
 Ellie: do you have disturbing thoughts
 Participant: no i wouldn't say
 Participant: disturbing
 Participant: no

Figure 2: Illustrative segment from interview "381" in the evaluation set, highlighted in Figure 1. Conversation turns are color-coded based on the proportion of keywords present, with keywords underlined for emphasis.

health issues. Figure 2 shows one such segment. Here, we see the segment containing the only four questions that Ellie’s model used to classify the participant, disregarding everything else in the conversation, including the question “*Have you been diagnosed with depression?*” Note that such questions may be asked to different participants, but an affirmative answer triggered Ellie to delve deeper into specific questions, questions that models could easily learn to identify and exploit to correctly classify the participants.

5.1 Implications in Clinical Practice

In clinical practice the final psychiatric diagnosis is typically determined through a clinical interview, often semi-structured, where rating scales serve as additional sources of information to aid in diagnosis. However, these rating scales have limitations, as responses can be influenced by factors such as the patient’s emotional state, comorbidities, relationship with the clinician, and patient self-bias (e.g., participants may be more likely to exaggerate their symptoms (Mao et al., 2023)).

Accordingly, the final goal of screening tools such as Ellie, is to contribute towards the replicability, consistency, standardization and the construction of objective measures that support the diag-

nosis of different mental disorders (Pendse et al., 2022).

As shown, the overall analysis described in this paper uncovers interesting biases in the data and shows how ostensibly good performance of NLP models can be deceiving and stress the importance of paying attention to the data and the rationales of the models rather than simply focusing on the superficial performance numbers. Thus, for automatic depression detection systems to be applicable in real-life clinical practice, systems must be able to provide practitioners with interpretable and transparent insights to validate systems decisions. There are complex interactions happening during a clinical interview, and accurately modeling is still an open challenge, highlighting the need to develop robust and ethical AI systems for this important and sensitive application domain.

6 Conclusions

Our analysis reveals that the prompts posed by the interviewer, Ellie, contain biases that allow models to more easily distinguish between depressed and control participants in the DAIC-WOZ dataset. By analyzing the keywords learned by the models, we discover that Ellie’s model tends to focus on highly localized segments of the interviews, primarily concentrated in the latter portion where more personal mental health questions are asked. In contrast, the model using participant responses alone does not exhibit such localization, instead gathering evidence from across the entire conversations. More broadly, our findings underline the need for caution when incorporating interviewers’ prompts into mental health diagnostic models. Interviewers often strategically adapt their questioning to probe for potential symptoms. As a result, models may learn to exploit these targeted prompts as discriminative shortcuts, rather than learning to characterize the language and behavior that are truly indicative of mental health conditions.

7 Ethical Considerations

In this section, we elaborate on the potential ethical issues.

1. **Data privacy, participant demographics, and consent.** All the experiments reported in this paper were made on the publicly available DAIC-WOZ dataset, a valuable resource used for training and validating depression detection systems from clinical interviews.

This particular dataset was collected by the Institute for Creative Technologies at the University of Southern California. According to the original paper, the DAIC-WOZ dataset received approval from Institutional Ethics Board. All the participants, including the U.S. armed forces veterans and general public from the Greater Los Angeles metropolitan area, were informed that their interviews will be used for academic purposes. All personal details like names, ages, and professions are either removed or anonymized, eliminating any risk of personal information exposure. Original videos from the interviews are not provided, but instead vector features of facial actions and eye gaze are given, making it impossible to reconstruct the participants' appearance. In general, the information of participants was rigorously protected.

2. **The role of AI-based diagnosis.** Our performed experiments aimed at highlighting the importance of using interpretable AI-based solutions as an assistant tools. Thus, the goal is not to replace human experts (psychologists and psychiatrists) but to develop systems that should be used only as support tools. The principle of leaving the decision to the machine would imply major risks for decision making in the health field, a mistake that in high-stakes healthcare settings could prove detrimental or even dangerous. The experiments reported in this paper represent a step forward on the development of bias-aware models in the context of clinical interviews analysis.

8 Limitations

In this section we discuss the limitations of the study described in this paper.

1. **Task configuration.** In this paper we only focused on the task of depression detection from clinical interviews, i.e., a controlled scenario where a mental health expert (therapist) conducts an interview with the goal to identify different psychological distress conditions present in the interviewed participant. This setup is significantly different from the so called "wild setting", which refers to the analysis of daily messages, e.g., social media posts. Thus, the findings and claims made in this paper are limited to a clinical setup, and might

not be applicable to different setups. As part of our future work, we plan to validate the impact of prompts generated by a fully automatic therapist in similar setups, in particular in the E-DAIC (DeVault et al., 2014) corpus.

2. **Corpus and modality specific.** Our study is limited to textual modality present in the DAIC-WOZ corpus. Given that the acoustic modality contains also Ellie's interventions, we would like to confirm the presence of the same bias in the acoustic modality. Thus, as part of our future work, we plan to extend our analysis to the additional modalities present in the selected corpus. Similarly, our findings apply specifically to the DAIC-WOZ corpus, hence we cannot confirm the presence of the same type biases in similar corpora. As part of our immediate work, we will replicate our analysis with other datasets like E-DAIC (DeVault et al., 2014), EATD (Shen et al., 2022), or the recently released ANDROIDS (Tao et al., 2023) dataset.

Acknowledgements

This work was supported by Idiap Research Institute's internal funds and computational resources. In addition, we thank CONAHCYT for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies and CIMAT Bajío Super-computing Laboratory (#300832). Reyes-Ramírez (CVU 1225869) thanks CONAHCYT for the support through the master's degree scholarship at CIMAT. Sanchez-Vega acknowledges CONAHCYT for its support through the program "Investigadoras e Investigadores por México" (Project ID.11989, No.1311).

References

- Navneet Agarwal and Gaël Dias. 2024. *Analysing Relevance of Discourse Structure for Improved Mental Health Estimation*. In *9th Workshop on Computational Linguistics and Clinical Psychology (CLPSYCH) associated to 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Saint Julian, Malta.
- Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

- Andrew Bailey and Mark D Plumbley. 2021. Gender bias in depression detection using audio features. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 596–600. IEEE.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.
- Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.
- Sergio Burdisso, Marcelo Luis Errecalde, and Manuel Montes y Gómez. 2019. Towards measuring the severity of depression in social media via text classification. In *XXV CACIC*, pages 577–588.
- Sergio Burdisso, Esaú Villatoro-Tello, Srikanth Madikeri, and Petr Motlicek. 2023. Node-weighted Graph Convolutional Network for Depression Detection in Transcribed Clinical Interviews. In *Proc. INTERSPEECH 2023*, pages 3617–3621.
- Ilias Chalkidis, Xiang Dai, Manos Fergadiotis, Prodromos Malakasiotis, and Desmond Elliott. 2022. An exploration of hierarchical attention transformers for efficient long document classification. *arXiv preprint arXiv:2210.05529*.
- Zhijun Dai, Heng Zhou, Qingfang Ba, Yang Zhou, Lifeng Wang, and Guochen Li. 2021. Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis. *Journal of affective disorders*, 295:1040–1048.
- David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhomme, et al. 2014. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Ming Fang, Siyu Peng, Yujia Liang, Chih-Cheng Hung, and Shuhua Liu. 2023. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomedical Signal Processing and Control*, 82:104561.
- Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Strattou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- Adria Mallo-Ragolta, Ziping Zhao, Lukas Stappen, Nicholas Cummins, and Björn W. Schuller. 2019. A Hierarchical Attention Network-Based Approach for Depression Detection from Transcribed Clinical Interviews. In *Proc. Interspeech 2019*, pages 221–225.
- Kaining Mao, Yuqi Wu, and Jie Chen. 2023. A systematic review on automated clinical depression diagnosis. *npj Mental Health Research*, 2(1):20.
- Kirill Milintsevich, Kairit Sirts, and Gael Dias. 2023. Towards automatic text-based estimation of depression through symptom prediction. *Brain Informatics*, 10.
- Meng Niu, Kai Chen, Qingcai Chen, and Lufeng Yang. 2021. 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)*, pages 4235–4239.
- Ives Cavalcante Passos, Francisco Diego Rabelo-da Ponte, and Flavio Kapczinski. 2023. *Digital mental health: a practitioner's guide*. Springer.
- Sachin R Pendse, Daniel Nkemelu, Nicola J Bidwell, Sushrut Jadhav, Soumitra Pathare, Munmun De Choudhury, and Neha Kumar. 2022. From treatment to healing: Envisioning a decolonial digital mental health. In *CHI Conference on Human Factors in Computing Systems*, pages 1–23.
- Ying Shen, Huiyu Yang, and Lin Lin. 2022. 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)*, pages 6247–6251. IEEE.
- Fuxiang Tao, Anna Esposito, and Alessandro Vinciarelli. 2023. The Androids Corpus: A New Publicly Available Benchmark for Speech Based Depression Detection. In *Proc. INTERSPEECH 2023*, pages 4149–4153.
- Ermal Toto, ML Tlachac, and Elke A. Rundensteiner. 2021. Audibert: A deep transfer learning multimodal classification framework for depression screening. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21*, page 4145–4154, New York, NY, USA. Association for Computing Machinery.

Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10. ACM.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Esaú Villatoro-Tello, Gabriela Ramírez-de-la Rosa, Daniel Gática-Pérez, Mathew Magimai.-Doss, and Héctor Jiménez-Salazar. 2021a. Approximating the mental lexicon from clinical interviews as a support tool for depression detection. In *Proc. ICMI’21*, page 557–566.

Esaú Villatoro-Tello, S. Pavankumar Dubagunta, Julian Fritsch, Gabriela Ramírez de-la Rosa, Petr Motlicek, and Mathew Magimai-Doss. 2021b. Late Fusion of the Available Lexicon and Raw Waveform-Based Acoustic Modeling for Depression and Dementia Recognition. In *Proc. Interspeech 2021*, pages 1927–1931.

James R Williamson, Elizabeth Godoy, Miriam Cha, Adrienne Schwarzenruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F Quatieri. 2016. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 11–18.

Danai Xezonaki, Georgios Paraskevopoulos, Alexandros Potamianos, and Shrikanth S. Narayanan. 2020. Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews. In *Interspeech*.

Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. 2022. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12226–12234.

Chen Zhuang, Deng Jiawen, Zhou Jinfeng, Wu Jincenzi, Qian Tieyun, and Minlie Huang. 2024. Depression detection in clinical interviews with LLM-empowered structural element graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Bochao Zou, Jiali Han, Yingxue Wang, Rui Liu, Shenghui Zhao, Lei Feng, Xiangwen Lyu, and Huimin Ma. 2022. Semi-structural interview-based chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. *IEEE Transactions on Affective Computing*.

Model	Learning Rate	Epoch	Features	Macro F ₁
P-GCN	1.022e-06	10	top-250	0.85
E-GCN	1.124e-06	10	auto	0.88

Table A1: Best hyperparameters obtained for the GCN models after optimization along with the obtained macro averaged F₁ score.

A Technical details

A.1 Graph Convolutional Network

A Graph Convolutional Network (GCN) is a multi-layer neural network that operates directly on a graph and induces embedding vectors of nodes based on the properties of their neighbors. In this work we use the inductive two-layer GCN described in [Burdisso et al. \(2023\)](#). Let $A \in \mathcal{R}^{n \times n}$ be the weighted adjacency matrix of the graph connecting words and interviews of the DAIC-WOZ training set, the GCN is defined as:

$$H^{(1)} = \sigma(\tilde{A}H^{(0)}W^{(0)}) \quad (1)$$

$$Z = \text{softmax}(\tilde{A}H^{(1)}W^{(1)}) \quad (2)$$

where $\tilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ represents the normalized symmetric adjacency matrix, $W^{(0)}$ is the learned node embeddings lookup table, and $W^{(1)}$ represents the learned weight matrix in the second layer. Loss is computed by means of the cross-entropy between Z_i and the one-hot encoded ground truth label Y_i for all i -th interview in the training set. Following the original paper, we set $k = 64$ for the k -dimensional feature matrix $H^{(1)} \in \mathcal{R}^{n \times k}$. The adjacency matrix is defined as follows:

$$A_{ij} = \begin{cases} mi(i, j) & \text{if } i, j \text{ are words \& } mi(i, j) > 0 \\ pr(i, j) & \text{if } i, j \text{ are words \& } i = j \\ \text{tf-idf}_{i,j} & \text{if } i \text{ is interview \& } j \text{ is word} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where mi is the point-wise mutual information and pr the PageRank ([Brin and Page, 1998](#)) score for node i .

Finally, in Section 5 we extracted all the words that the model learned to associate to the depressed category. To select these keywords we selected all words i such that $P(\text{depressed} \mid \text{word}_i) > P(\text{control} \mid \text{word}_i)$, that is, $\text{keywords} = \{\text{word}_i \mid Z_{i,\text{depressed}} > 0.5\}$.

Model	Learning Rate	Epoch	Macro F ₁
P-longBERT	2.497e-03	10	0.72
<i>first half</i>	1.352e-03	10	0.67
<i>second half</i>	6.051e-03	10	0.73
E-longBERT	1.044e-03	6	0.84
<i>first half</i>	8.209e-04	9	0.60
<i>second half</i>	5.075e-04	7	0.84

Table A2: Best hyperparameters obtained for the long-BERT models after optimization along with the obtained macro averaged F₁ score.

A.2 Longformer BERT

The Longformer (Beltagy et al., 2020) replaces the quadratic self-attention mechanism of Transformers (Vaswani et al., 2017) with a combination of global and local windowed attention, scaling linearly with sequence length. This modification enables efficient processing of documents with thousands of tokens, consistently outperforming Transformer-based models on long document tasks. In particular, we used the version of Longformer described in Chalkidis et al. (2022) which has been warm-started re-using the weights of BERT, and continued pre-trained for MLM following the paradigm described in the original Longformer paper. This pre-trained model is available in Hugging Face at <https://huggingface.co/kiddothe2b/longformer-mini-1024>.

A.3 Implementation details

All models were implemented using PyTorch and were optimized using *Optuna* (Akiba et al., 2019) with 100 trials for hyperparameter search maximizing the macro averaged F1 score. In each trail, models were trained using AdamW (Loshchilov and Hutter, 2019) optimizer ($\beta_1=0.9, \beta_2=0.999, \epsilon=1e-8$) with *learning rate* and number of epochs n searched in $\gamma \in [1e-7, 1e-3]$ and $n \in [1, 10]$, respectively. In addition, for GCN, the optimization also tried the three feature selection techniques described in the original paper, *auto*, *top-k*, *none* for, respectively, automatic selection based on term weights learned using Logistic Regression, top- k best selection based on ANOVA *F-value* between words and labels with $k \in \{100, 250, 500, 1000, 1500\}$, and no feature selection (full vocabulary). Best obtained hyperparameters for the GCN models are shown in Table A1. Finally, Table A2 presents the parameters obtained for the *longBERT* models,

along with the results of the complementary ablation experiments mentioned at the end of Section 5. Specifically, we divided each interview into two equal parts and performed fine-tuning and evaluation using either the first or the second half. The objective was to reinforce our conclusions regarding the existence of a bias, particularly in the second half of the interviews, as detected by the keywords from the GCN model (Figure 1).