Analyzing Symptom-based Depression Level Estimation through the Prism of Psychiatric Expertise

Navneet Agarwal*[1], Kirill Milintsevich*[1,2], Lucie Metivier[3], Maud Rotharmel[4], Gaël Dias[1], Sonia Dollfus[5]
[1]Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France.
[2]Institute of Computer Science, University of Tartu, Tartu, Estonia.
[3]Normandie Univ, UNICAEN, PhIND, UMR-S 1237, GIP CYCERON, 14000 Caen, France
[4]Service hospitalo-universitaire de psychiatrie de l'adulte, Centre Thérapeutique d'Excellence, Centre Hospitalier du Rouvray, Sotteville-les-Rouen, France.
[5]CHU de Caen, Service de Psychiatrie; Normandie Univ, UNICAEN, ISTS, GIP Cyceron; Normandie Univ, UNICAEN, UFR de Médecine, 14000 Caen, France.
{navneet.agarwal,kirill.milintsevich}@unicaen.fr

Abstract

The ever-growing number of people suffering from mental distress has motivated significant research initiatives towards automated depression estimation. Despite the multidisciplinary nature of the task, very few of these approaches include medical professionals in their research process, thus ignoring a vital source of domain knowledge. In this paper, we propose to bring the domain experts back into the loop and incorporate their knowledge within the gold-standard DAIC-WOZ dataset. In particular, we define a novel transformer-based architecture and analyze its performance in light of our expert annotations. Overall findings demonstrate a strong correlation between the psychological tendencies of medical professionals and the behavior of the proposed model, which additionally provides new state-of-the-art results.

Keywords: Depression estimation, psychiatrist annotations, external knowledge introduction.

1. Introduction

Mental illness is a serious issue with high social and economic costs, yet a significant number of mental illness cases go undetected. Up to half of the patients with psychiatric disorders are not diagnosed as having mental illness by their primary care physicians (Higgins, 1994), a situation made worse due to a shortage of medical professionals (Butryn et al., 2017). As a consequence, artificial intelligence in psychiatry has been emerging as a general term that implies the use of computerized techniques and algorithms for the diagnosis, prevention, and treatment of mental illnesses (Fakhoury, 2019). Within clinical settings, semi-structured interviews are the common practice for evaluating a person's mental health. These interviews usually act as inputs for training automated models with self-assessment scores being used as the final ground truth (e.g. Patient Health Questionnaire PHQ-8 for depression estimation). Throughout the literature, different strategies have been proposed for the automated estimation of depression. Multimodal models combine inputs from different modalities (Ray et al., 2019; Qureshi et al., 2019). Multitask architectures simultaneously learn related tasks (Qureshi et al., 2019, 2020). Gender-aware models explore the impact of gender on depression estimation (Bailey and Plumbley, 2021; Qureshi et al., 2021). Hierarchical models process transcripts at different granularity levels (Mallol-Ragolta et al., 2019; Xezonaki et al., 2020). Attention models integrate external knowledge from lexicons (Xezonaki et al., 2020). Feature-based strategies compute multimodal characteristics (Dai et al., 2021). Graph-based systems aim to study complex structures within interview transcripts (Hong et al., 2022; Niu et al., 2021). Multiview architectures treat the input transcripts as a combination of different text views (Agarwal et al., 2022). Symptom-based models treat depression estimation as an extension of the symptom prediction problem (Milintsevich et al., 2023). Domain-specific language models are built (Ji et al., 2022) and large language models are prefix-tuned to automate depression level estimation (Lau et al., 2023).

Despite the multidisciplinary nature of the problem, most previous research initiatives have failed to include medical professionals in the learning process, except Yadav et al. (2020), who asked a psychiatrist to label tweets in terms of PHQ-9 symptoms. In this paper, we propose to follow this line of research by providing a clinically annotated version of the gold-standard DAIC-WOZ dataset[1] (Gratch et al., 2014) to allow the integration of domain expertise in artificial models. We also define a novel transformer-based model and examine ways to utilize psychiatric annotations within its learning

*These authors contributed equally to this work

[1]The Distress Analysis Interview Corpus (DAIC) is the only publicly available resource for interview-based distress analysis.

LREC-COLING 2024, pages 974–983
20-25 May, 2024.

process. Finally, we analogize the psychological tendencies of medical professionals against the proposed model in an attempt to validate its reliability as a predictive model in clinical settings. Overall results show that our model successfully aligns with medical experts thus being a trustful source of predictions for clinicians in psychiatry. Additionally, the proposed model provides new state-of-the-art results over the DAIC-WOZ test set.

2. Related Work

Different architectures and strategies have been used throughout the literature to build models capable of estimating patients' depression level based on patient-therapist interviews. One promising research area is to leverage inputs from different modalities into one learning modal. Qureshi et al. (2019) explore the possibility of combining audio, visual, and textual input features into a single architecture using attention fusion networks. They further show that training the model for regression and classification simultaneously on the same dataset provides improvements in results. Ray et al. (2019) present a similar framework that invokes attention mechanisms at several layers to identify and extract important features from different modalities. The network uses several low-level and mid-level features from audio, visual and textual modalities of the participants' inputs. Another interesting approach aims at combining different tasks that share some common traits thus following the multi-task paradigm. Qureshi et al. (2020) propose to simultaneously learn both depression level estimation and emotion recognition on the basis that depression is a disorder of impaired emotion regulation. They show that this combination provides improvements in performance for the multiclass problem as well as the regression of the PHQ-8 score. Building on the success of hierarchical models for document classification, different studies (Mallol-Ragolta et al., 2019; Xezonaki et al., 2020) propose to encode patient-therapist interviews with hierarchical structures, showing boosts in performance. Xezonaki et al. (2020) further extend their proposal and integrate affective information (emotion, sentiment, valence, and psycho-linguistic annotations) from existing lexicons in the form of specific embeddings. Exploring a different research direction, Qureshi et al. (2021) study the impact of gender on depression level estimation and build four different gender-aware models that show steady improvements over gender-agnostic models. In particular, an adversarial multi-task architecture provides the best results overall. Along the same line, Bailey and Plumbley (2021) study gender bias from audio features as compared to (Qureshi et al., 2021), who target textual information. They find that deep learning models based on raw audio are more robust to gender bias than ones based on other common hand-crafted features, such as mel-spectrogram. Although most strategies rely on deep learning architectures, a different research direction is proposed by Dai et al. (2021), who build a topic-wise feature vector based on a context-aware analysis over different modalities (audio, video, and text). Niu et al. (2021) use graph structures within their architecture to grasp relational contextual information from audio and text modality. They propose a hierarchical context-aware model to capture and integrate contextual information among relational interview questions at word and question-answer pair levels. Milintsevich et al. (2023) treat binary classification as a symptom profile prediction problem and train a multi-target hierarchical regression model to predict individual depression symptoms from patient-therapist interview transcripts. Agarwal et al. (2022) highlight the importance of retaining discourse structure and define multi-view architectures that divide the input transcript into views based on sentence identities. The two views are processed both independently and co-dependently in order to account for intra-view and inter-view interactions. Building upon the success of language models in understanding textual data, Ji et al. (2022) fine-tune different BERT-based models on mental health data and provide a pre-trained masked language model for generating domain-specific text representations. Lau et al. (2023) further account for the lack of large-scale high-quality datasets in the mental health domain and propose the use of prefix-tuning as a parameter-efficient way of fine-tuning language models for mental health.

The gathering and assimilation of external knowledge into neural networks have garnered substantial attention in research endeavors in the domain of mental health. For the former case, Arseniev-Koehler et al. (2018) asked crowd workers to read excerpts of de-identified interview data from the DAIC-WOZ and rate how likely they thought a speaker had depression based on the transcribed utterances. Similarly, Yadav et al. (2020) work with Twitter data and employ four native English speakers from multiple disciplines to independently annotate tweets into the 9 categories of PHQ-9. For the latter case, various strategies have been proposed for the integration of external knowledge into neural network training. Outside the mental health domain, Soares et al. (2019) and Boualili et al. (2020) use special tokens to highlight information directly within the input text and rely on fine-tuning pretrained language models to understand the importance of marked text. Deshpande and Narasimhan (2020), (Stacey et al., 2022) and Wang et al. (2022) introduce additional loss terms during training as a means to guide the attention mechanism within the

Depression severity	Data split		
	Train	Val.	Test
No symptoms $[0..4]$	47	17	22
Mild $[5..9]$	29	6	11
Non-depressed Total	76	23	33
Moderate $[10..14]$	20	5	5
Moderately severe $[15..19]$	7	6	7
Severe $[20..24]$	4	1	2
Depressed Total	31	12	14
Total	107	35	47

Table 1: Number of interviews for each depressive class severity in the DAIC-WOZ dataset, distributed by train, validation and test sets.

neural networks towards the desired distributions. Within the mental health domain, only Xezonaki et al. (2020) generate custom context vectors using information from different lexicons, which are concatenated to word level representations.

3. Dataset and Psychiatric Annotations

3.1. Dataset

The Distress Analysis Interview Corpus (DAIC) is a multimodal corpus of semi-structured clinical interviews designed to simulate standard protocols for identifying people at risk of depression. Within our research, we focus on the textual input from the publicly available Wizard-of-Oz part of the corpus (DAIC-WOZ), which contains 189 interviews, where patients interact with an animated virtual agent controlled by a human therapist from a different room. Each session ranges from 7 to 33 minutes with an average time of 16 minutes. The dataset contains valuations for eight specific symptoms that are part of the PHQ-8 questionnaire: loss of interest, feeling of depression, sleeping habits, tiredness, loss of appetite, feeling of failure, lack of concentration and lack of movement. Table 1 shows the data splits between train, development and test sets, along with the class imbalance within the DAIC-WOZ dataset.

3.2. Psychiatrist Annotations

In our attempt to reintroduce domain expertise into the learning process, we carried out the clinical annotation of the DAIC-WOZ dataset[2]. In contrast to previous works that use crowd workers (Arseniev-Koehler et al., 2018) or native English speakers (Yadav et al., 2020) as annotators, we select mental health professionals for the annotation process. In particular, three psychiatrists from public hospitals were employed to undertake two major tasks: (1) span-based annotation of the transcripts and (2) PHQ-8 scoring based on interview transcripts.

Span-based annotation: This task consists of highlighting information within transcripts that influences a psychiatrist's decision during an interview. Since it is a subjective task that lacks a definitive right or wrong answer, a common consensus on the importance of various utterances within the transcripts does not exist. Even within the field of medicine, professionals do not universally agree on the significance of various pieces of information, and subtle differences in opinion exist between psychiatrists based on their individual knowledge and experience. As such, after various meetings and discussions with the psychiatrists, it was agreed that the medical annotators should have complete freedom to annotate the transcripts without any constraints in order to capture their true judgment. As a consequence, we forgo defining detailed annotation protocols and rely on the annotator's judgment as experts in the field for the reliability of their annotations. However, they were encouraged not only to identify information that suggests the presence of depression, but also to pinpoint clues that indicate its absence. Furthermore, the inherent lack of consensus within the task eliminates the need for inter-annotator agreements. In case multiple annotators are assigned per transcript, a simple union of annotated spans would be used to capture knowledge from all assigned annotators. Unfortunately, at this stage of our research, only one annotator per transcript could be assigned due to the workload experienced by the annotators, particularly due to the radical increase of mental care demand after the covid pandemic coupled with the shortage of mental health professionals. The current annotation process lasted nearly 5 months and we anticipate this time frame to scale linearly with the increase in the number of annotators per transcript.

For the annotation purpose, we designed an online tool based on the doccano[3] project which was hosted on servers from the herokou platform[4] enabling the entire annotation process to take place remotely for the convenience of the psychiatrists. The tool was designed to allow the psychiatrists to annotate any span of text (word, phrase, sentence, text) within the transcript and assign a label of importance to each span: highly important, important (default) or minimally important. Upon analysis, it was found that these labels did not provide any information since more than 99% of the spans were marked with the default label (important), and were therefore not used in any further analysis. The annotation process gave rise to an

[2]The annotations can be accessed at https://github.com/navneet-agarwal/DAIC-WOZ-Annotations

[3]https://github.com/doccano/doccano

[4]https://www.heroku.com/

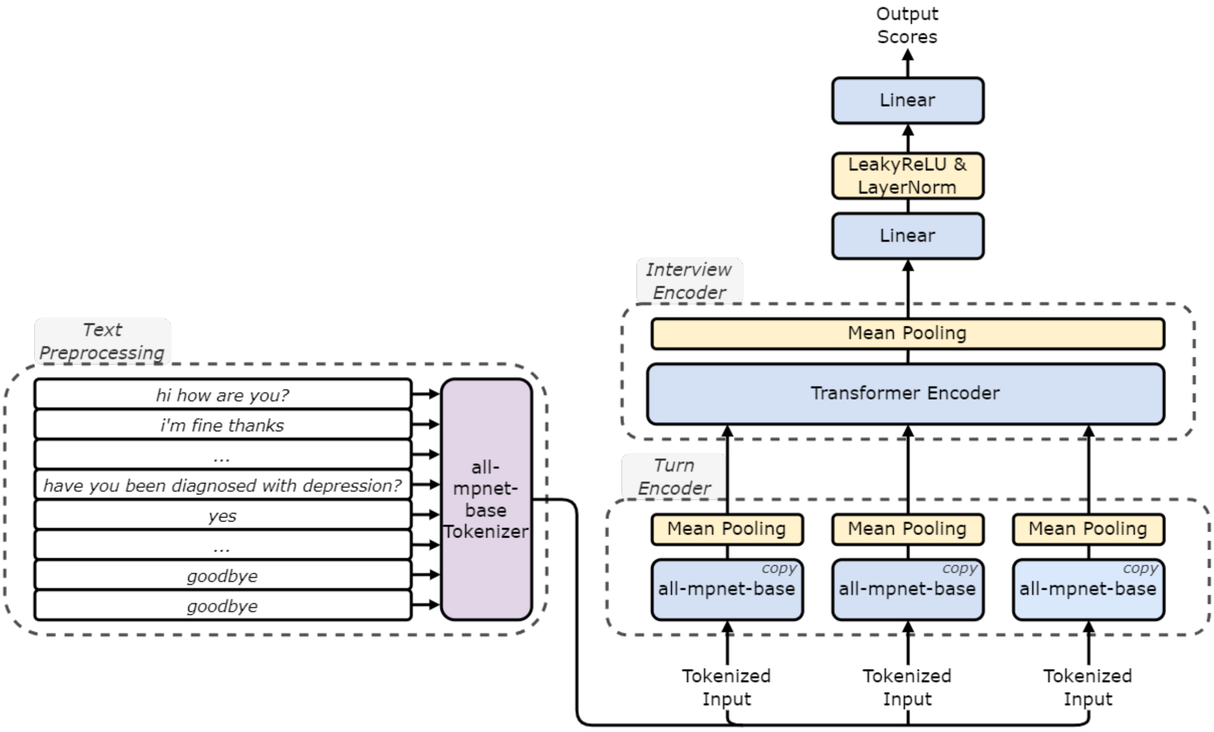


Figure 1: Hierarchical neural architecture for symptom-based prediction.

Span Level	Non-Depressed	Depressed
Word	467 (3.53)	227 (3.98)
Phrase	4101 (31.06)	1913 (33.56)
Sentence	0	0
Multi-sentences	77 (0.58)	42 (0.73)
Total	4645 (35.18)	2182 (38.28)

Table 2: Number of annotations for different levels of annotation spans. Figures in bracket indicate the average number of annotations per transcript.

average of 36.12 annotations per transcript (35.18 for the non-depressed class and 38.28 for the depressed class) with a mean length of 7.45 words (7.74 for the non-depressed class and 7.17 for the depressed class). The distribution of the annotations by patient class and span level is given in Table 2. Interestingly, complete sentences were not annotated by any of the psychiatrists, who mostly followed a ngram-based strategy, with a small number of annotations focusing on multiple sentences. Furthermore, none of the psychiatrists highlighted questions within the dataset with all the annotations contained within patient responses.

PHQ-8 scoring: This task involves completing the self-assessment PHQ-8 questionnaire on behalf of each patient only based on their interview transcripts. Although the PHQ-8 screening tool is widely used as a measure of depression and has been found to be precise (Shin et al., 2019), it relies on the subjective assessment of the patient about his/her condition outside the context of the interview. As such, an interview transcript might not contain enough information to accurately express the intensity of individual symptoms. Furthermore, since the interviews are conducted with the aim of depression estimation and not specifically for fulfilling the PHQ-8 questionnaire, information on some symptoms might be missing altogether within individual transcripts depending on the questions asked during the interview. In order to verify these propositions, we asked the clinicians to fulfill the PHQ-8 questionnaires on behalf of each patient based on their understanding of the given transcripts. This task consists of evaluating each of the 8 symptoms within the PHQ-8 questionnaire on a Likert scale ranging from 0 to 3. The statistics about this task, illustrated in Table 3, show that 5 out of 8 symptoms (i.e. loss of interest, feeling of depression, sleeping habits, feeling of tiredness, and feeling of failure) are steadily mentioned in most transcripts, while 3 of them (i.e. loss of appetite, lack of concentration and lack of movement) could not be measured reliably by the psychiatrists. This confirms our claims regarding the lack of symptom-level information within individual interviews. This annotation task also acts as a human expert performance baseline, that defines an achievable learning goal for correctly inferring PHQ-8 scores for each symptom based on information present within the transcripts.

Symptoms	No interest	Depressed	Sleep	Tired	Appetite	Failure	Concentration	Movement
# annotations	178	188	179	160	47	176	48	10

Table 3: Nb. of psychiatrist scorings for each PHQ-8 symptom over the 189 interviews of the DAIC-WOZ.

ELLIE: *how close are you to your family*
PARTICIPANT: *@@ very close @@ even though i don't live with them @@ i try to see them as much as possible @@*
ELLIE: *mhm*
ELLIE: *how do you like your living situation*
PARTICIPANT: *uh it's ok*

Figure 2: Example of annotation marking.

Model	MAE Dev.	MAE Test
SOTA		
ASP MT. DLC+DLR+EIR (Qureshi et al., 2020)		3.69
HCAG-T (Niu et al., 2021)	3.73	-
SGNN (Hong et al., 2022)	3.76	-
Symptom prediction (Milintsevich et al., 2023)	3.61	3.78
Dual encoder (warm start) (Lau et al., 2023)	2.76	3.80
Our Configurations		
Baseline model	4.08	**3.52**
Marked-up model	3.49	3.60

Table 4: Comparison of overall model performance against current state-of-the-art results. The results are averaged over 5 random initializations.

4. Model and Mark-up Strategy

4.1. Neural Network Architecture

To learn the 8 symptom values of the PHQ-8, we design the transformer-based hierarchical model illustrated in Figure 1. The architecture is based on the model defined by Milintsevich et al. (2023), which has been updated to have access to sentence-level attention and take advantage of recent sentence representation models. In particular, the architecture has undergone two significant alterations compared to the definition in §3.2 of (Milintsevich et al., 2023): (1) the BiLSTM cells are replaced by a transformer-based encoder at the interview level (interview encoder), and (2) the pre-trained turn encoder is based on the all-mpnet-base model[5] in place of *S-RoBERTa*[6], both using a contrastive learning objective (Reimers and Gurevych, 2019). In particular, the model consists of two encoders: the turn encoder that encodes each sentence and the interview encoder that encodes sentence level representations into an interview level embedding. The interview level embedding is then passed through a feed-forward network that maps it to a prediction vector $m = [m_1, m_2, ..., m_8]$, where each predicted label $m_k \in [0, 3]$ represents a symptom score for the corresponding question in the PHQ-8 questionnaire. The interview encoder contains 4 layers containing 12 attention heads each with an intermediate size of 1536 and an hidden size of 768. This model acts as the base architecture for the different experiments and model configurations explored within our research and is referred to as the **Baseline model**.

[5]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[6]https://huggingface.co/sentence-transformers/all-distilroberta-v1

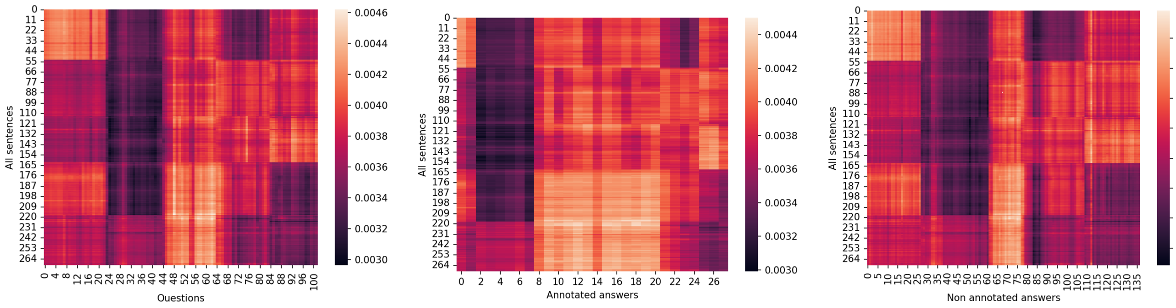
4.2. External Knowledge Integration

In our effort to reintroduce domain expertise into depression estimation tasks, we incorporate psychiatrist annotations into the learning process of our neural network model. We align our work with the research approach taken by Soares et al. (2019) and Boualili et al. (2020), and introduce special markers into the input text to directly highlight clinical annotations within the transcripts. The underlying idea is that explicitly marking spans in the input text may allow the model to carefully identify the annotations and make a more informed prediction. Consequently, all annotations provided by the psychiatrists are encompassed in between the @@ markers within the transcripts, giving rise to a marked-up corpus (example in figure 2). We use the Baseline architecture defined earlier and fine-tune it using the marked-up corpus. Specifically, the pre-trained *all-mpnet-base* model is fine-tuned by unfreezing only the final layer. The resulting model is referred to as the **Marked-up model**.

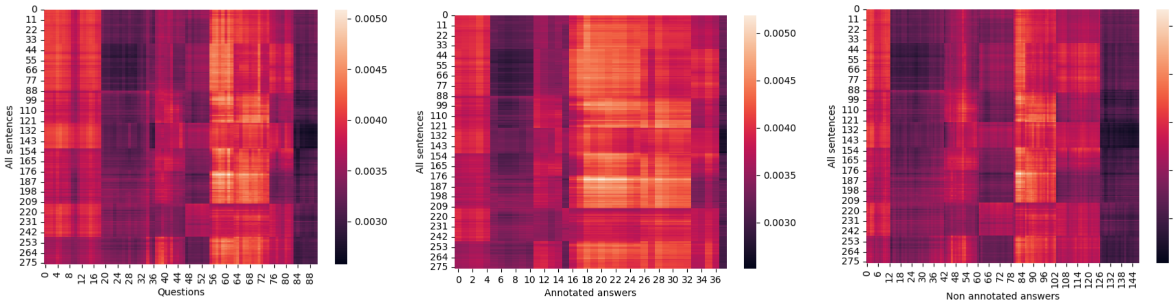
5. Overall Results

Table 4 provides overall results for the various model configurations considered in our experiments and puts them into perspective by comparison against current state-of-the-art results. Our baseline model provides new state-of-the-art performance for the Mean Absolute Error (MAE) metric on the test set of the DAIC-WOZ on an average over 5 runs. It is interesting to notice that the marked-up model does not improve over the baseline model despite containing extra information, although it does outperform all previous research initiatives. This issue is further discussed in detail in §7.

Ablation study: We conduct an ablation study to analyze the amount of information contained within



(a) Patient id: 335 (Depressed). Average attention scores (Q, A, N) = (1.994e-05, 2.030e-05, 1.966e-05)



(b) Patient id: 307 (Non-Depressed). Average attention scores (Q, A, N) = (1.955e-05, 2.044e-05, 1.905e-05)

Figure 3: Heat maps of sentence level attention scores from the Baseline model for two different patients.

Ablation	MAE on Test set
Baseline model	**3.52**
Baseline$_{\text{ann.}}$ inference	4.02
Baseline$_{\text{non-ann.}}$ inference	3.84

Table 5: Ablation study with baseline model for exclusively non-annotated and annotated sentences.

the clinical annotations. Given the complete set of information required for estimating depression, we seek to understand the role played by our clinical annotations within this set. For that purpose, we define two new input configurations and use them with the trained baseline model at the inference stage to generate new predictions over the modified inputs. The two versions in this input ablation study are defined as follows:

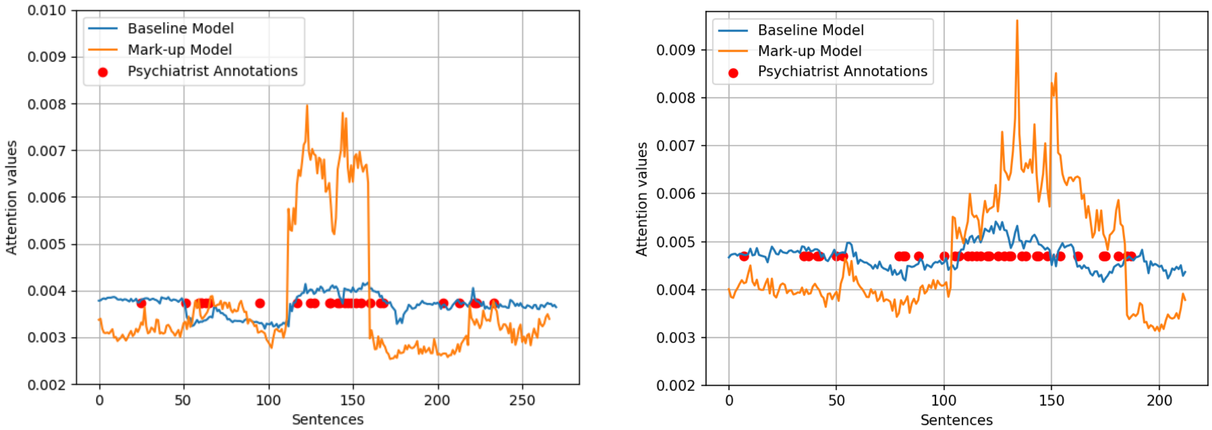
Baseline$_{ann}$ inference: only question-answer pairs with at least one annotation are kept within the input transcripts.

Baseline$_{non-ann}$ inference: only question-answer pairs without any annotation are retained within the input transcripts.

Results of the ablation study are shown in table 5. We see a significant drop in performance on removing annotated question-answer pairs from the input transcripts, highlighting the validity of the psychiatrists' annotations. Surprisingly, we also see a drop in performance when only annotated question-answer pairs are used as inputs. This behavior can be attributed to the fact that in this case the number of sentences within the interviews is severely reduced and as such the coherence of the discourse is undermined, affecting the performance of the automated models.

6. Attention and Annotated Spans

Psychiatrist annotations highlight text spans that hold relevance for depression estimation as per clinicians' knowledge and medical guidelines. Given their importance from the medical point of view, we propose to verify whether automated models attend to the same annotated text spans or look for information that complements clinical knowledge. Psychiatrist annotations are analyzed against sentence-level attention scores from the model, the sentence being the atomic textual element for this analysis. In particular, we focus on 3 different sentence types: questions (Q), non-annotated turns (N) that contain answers without any annotations, and clinically-annotated turns (A) that contain patient responses with at least one annotation. Thus, each attention head $H^{s \times s}$ of the interview encoder is converted into three attention sub-matrices $H^{s \times q}$, $H^{s \times n}$ and $H^{s \times a}$, where s is the number of sentences in a given transcript, q the number of questions, a the number of annotated turns and n the number of non-annotated turns, such that $s = q + n + a$. For each interview, we average the sentence-level attention scores for Q, N and A sentence types for all attention heads contained in the interview encoder as defined in equation 1, where h and l stand for the number of



(a) Patient id 335 (Depressed)

(b) Patient id 307 (Non-Depressed)

Figure 4: Attention scores for the baseline and marked-up models plotted against clinical annotations.

Class	Metric	Q	N	A
Non-depressed	min.	12.84	12.93	13.60
	max.	137.50	136.76	135.35
	med.	42.03	42.10	**42.25**
	avg.	30.85	31.01	**31.25**
Depressed	min.	15.29	15.02	15.37
	max.	103.88	102.83	110.89
	med.	37.96	38.50	**38.82**
	avg.	12.18	12.18	**12.29**

Table 6: Sentence-level attention scores calculated over the DAIC-WOZ dataset for **Q**uestions, **N**on-annotated and **A**nnotated turns. Values are with a precision of 10^{-4}. Med. and avg. stand for median and arithmetic mean.

heads and layers respectively.

$$\overline{X} = \frac{1}{l.h} \sum_{l,h} \frac{1}{i.j} \sum_{i,j} H_{i,j}^{s \times x}, \forall x \in \{q, n, a\} \quad (1)$$

Finally, we average these values over the 189 interviews of the DAIC-WOZ to get the overall picture. Results with the baseline model are given in Table 6 and show that the transformer-based model focuses more on clinically annotated spans compared to other parts of the transcripts, independently of the patient class. This provides the first evidence that the baseline model targets clinically motivated spans for its decision process without the introduction of any external knowledge or use of specific architectures tuned towards guiding the attention values.

To complement this analysis, figure 3 plots three attention heatmaps $\overline{Q}$, $\overline{A}$ and $\overline{N}$ with brighter regions representing higher attention scores. Plots are provided for a depressed patient as well as a non-depressed patient. This illustration exemplifies overall results and shows that although model attention is distributed over all three categories, clinically-annotated turns receive higher average attention as compared to non-annotated turns and questions. Finally, figure 4 illustrates the attention scores in perspective of the psychiatrists' annotations for the same patients. Following the blue line corresponding to the baseline model, we observe an increase in attention scores in the vicinity of psychiatrist annotations, while the opposite is true in the absence of annotations. These plots represent a general trend observed throughout the dataset with some exceptions.

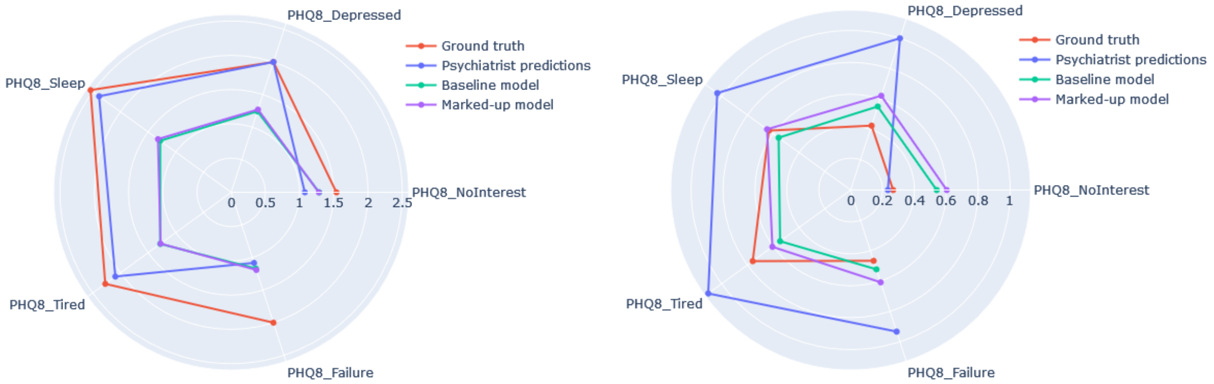
7. Performance Analysis against Knowledge Introduction

Although the baseline model attends to parts of the interviews that psychiatrists find relevant, we explore the impact of the introduction of clinician expertise directly in the learning process and analyze the performance of the marked-up model. Overall results are illustrated in Table 7 and do not evidence gains in performance resulting from the knowledge added by the psychiatrist annotations. Indeed, the baseline model outperforms the marked-up model 5 times out of 8 for both the depressed and non-depressed classes. This confirms our previous findings from section §6, showing that the baseline architecture already attends to clinically annotated sentences, thus reducing the impact of the marked-up strategy. Figure 4 compares both baseline and marked-up models, with plots showing similar behaviors of attending to the annotated sentences although with different amplitude. In particular, the marked-up model tends to pay high attention to the middle of the transcripts thus failing to highlight important information from other regions. This is not the case for the baseline model, which has more evenly distributed attention values, while still being consistent with psychiatrist annotations.

In order to put prediction results into perspective, we calculate the Mean Absolute Error (MAE) between the psychiatrists' PHQ-8 scores and pa-

Symptoms	Psychiatrist Pred.		Baseline model		Marked-up model	
	Depr.	Non-Depr.	Depr.	Non-Depr.	Depr.	Non-Depr.
Loss of interest	0.615	0.366	**0.611**	**0.431**	0.699	0.485
Feeling of depression	0.571	0.696	**0.884**	**0.443**	0.939	0.465
Sleeping habits	0.615	0.533	0.761	**0.691**	**0.651**	0.808
Tiredness	0.727	0.689	**0.797**	0.711	0.812	**0.666**
Feeling of failure	1.083	0.800	0.820	**0.543**	**0.786**	0.573
Lack of concentration	-	-	**1.332**	0.521	1.361	**0.475**
Loss of appetite	-	-	**0.932**	0.745	1.037	**0.628**
Lack of movement	-	-	1.008	**0.105**	**0.964**	0.125

Table 7: MAE calculated against patients' self-assessments scores by symptoms over the DAIC-WOZ test set. Results are averaged over 5 runs for the automated models. Psychiatrist prediction evidences the difference between the patients' assessments and the psychiatrists' ones.



(a) Depressed class

(b) Non-Depressed class

Figure 5: Radar plots showing symptom-wise average scores for the different automated models, the patient self-assessments and the psychiatrists' ratings over the test set of the DAIC-WOZ. Note that only 5 symptoms are illustrated, which refer to the ones that psychiatrists could reliably annotate.

tients' self-assessments. Results in Table 7 show that psychiatrist predictions outperform automated models in most cases, albeit by a small margin for most of the symptoms (feeling of failure being an exception where the baseline model performs better). Further analysis of psychiatrist scoring confirms findings from the medical domain (Domken et al., 1994), showing that clinicians tend to under-evaluate the PHQ-8 scores for the depressed class while over-evaluating those for the non-depressed class. Intriguingly, we observe the same behavior for the automated models as illustrated in Table 8. The figures show that both the baseline model and the marked-up model exhibit the same behavior as psychiatrists, which further strengthens our claim of shared psychological tendencies between our proposed model and psychiatrists. As expected, the number of transcripts misdiagnosed by the automated models far exceeds those misdiagnosed by psychiatrists. This is due to the fact that models generate floating point predictions whereas psychiatrists' predictions are based on a Likert scale ranging from 0 to 3.

In order to further analyze the behavior of over and under-evaluation, we plot the symptom-wise average scores for the different automated models, the patient self-assessments and the psychiatrists' ratings in figure 5. The illustrations show a high correlation between the results from the two automated models. Both baseline and marked-up models generate the same average scores for the depressed

Symptoms	Depr.		Non-Depr.	
	Over	Under	Over	Under
Psychiatrist Prediction				
Loss of Interest	1	5	3	6
Feeling of depression	3	3	16	2
Sleeping habits	3	3	10	2
Tiredness	2	3	12	5
Feeling of failure	1	8	13	5
Baseline Model				
Loss of Interest	4	9	24	5
Feeling of depression	2	12	24	9
Sleeping habits	1	12	19	10
Tiredness	1	10	14	14
Feeling of failure	1	11	20	9
Marked-up model				
Loss of Interest	4	9	27	3
Feeling of depression	3	11	26	7
Sleeping habits	1	12	19	11
Tiredness	1	10	15	14
Feeling of failure	2	10	23	7

Table 8: Number of over- and under-evaluated transcripts in the test set for the baseline model, the marked-up model and the psychiatrists' scorings.

class, while for the non-depressed class the values are very close. This confirms that the introduction of annotations into the learning process through the markup strategy does not provide significant performance gain. These plots also support the claims of over and under-evaluation of PHQ-8 scores, and showcase a similar pattern as seen in table 8.

8. Conclusion

In this paper, we examine automated depression estimation through the prism of psychiatric expertise and compare the behavior of automated models against clinical annotators. The analysis of sentence-level attention scores shows that the baseline model learns to analyze the transcripts in ways similar to trained psychiatrists despite the lack of medical knowledge in the training process. Our analysis further establishes a strong correlation between the psychological tendencies of our automated model and medical professionals, thus validating its role as a credible source of predictions for clinicians in psychiatry. Additionally, the proposed architecture provides new state-of-the-art results over the DAIC-WOZ test set. The source code and the clinically annotated DAIC-WOZ dataset will be publicly released upon acceptance.

9. Acknowledgements

This research is supported by the FHU A^2M^2P project funded by the G4 University Hospitals of Amiens, Caen, Lille and Rouen (France). The calculations for model's training and inference were carried out in the High Performance Computing Center of the University of Tartu (University of Tartu, 2018).

10. Bibliographical References

Navneet Agarwal, Gaël Dias, and Sonia Dollfus. 2022. Agent-based splitting of patient-therapist interviews for depression estimation. In *Workshop on Participatory Approach to AI for Mental Health (PAI4MH) associated to 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Alina Arseniev-Koehler, Sharon Mozgai, and Stefan Scherer. 2018. What type of happiness are you looking for? - a closer look at detecting mental health from language. In *5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (CLPSYCH) associated to 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Andrew Bailey and Mark D. Plumbley. 2021. Gender bias in depression detection using audio features. In *29th European Signal Processing Conference (EUSIPCO)*, pages 596–600.

Lila Boualili, José G. Moreno, and Mohand Boughanem. 2020. Markedbert: Integrating traditional IR cues in pre-trained language models for passage retrieval. In *43rd International ACM SIGIR conference on research and development in Information Retrieval (SIGIR)*, pages 1977–1980.

Tracy Butryn, Leah Bryant, Christine Marchionni, and Farhad Sholevar. 2017. The shortage of psychiatrists and other mental health providers: causes, current state, and potential solutions. *International Journal of Academic Medicine*, 3(1):5–9.

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.

Ameet Deshpande and Karthik Narasimhan. 2020. Guiding attention for self-supervised learning with transformers. In *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4676–4686. Association for Computational Linguistics.

Marc Domken, Jan Scott, and Peter Kelly. 1994. What factors predict discrepancies between self and observer ratings of depression? *Journal of Affective Disorders*.

Marc Fakhoury. 2019. *Artificial Intelligence in Psychiatry*, pages 119–125. Springer Singapore, Singapore.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, 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 *9th International Conference on Language Resources and Evaluation (LREC)*.

Edmund S Higgins. 1994. A review of unrecognized mental illness in primary care: prevalence, natural history, and efforts to change the course. *Archives of family medicine*, 3(10):908.

Simin Hong, Anthony Cohn, and David Crossland Hogg. 2022. Using graph representation learning with schema encoders to measure the severity of depressive symptoms. In *International Conference on Learning Representations (ICLR)*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *13th Language Resources and Evaluation Conference (LREC)*, pages 7184–7190.

Clinton Lau, Xiaodan Zhu, and Wai-Yip Chan. 2023. Automatic depression severity assessment with deep learning using parameter-efficient tuning. *Frontiers in Psychiatry*, 14:1160291.

Adria Mallol-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 *Interspeech (INTERSPEECH)*, pages 221–225.

Kirill Milintsevich, Kairit Sirts, and Gaël Dias. 2023. Towards automatic text-based estimation of depression through symptom prediction. *Brain Informatics*, 10(1):1–14.

Meng Niu, Kai Chen, Qingcai Chen, and Lufeng Yang. 2021. Hcag: A hierarchical context-aware graph attention model for depression detection. In *IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4235–4239.

Syed Arbaaz Qureshi, Gaël Dias, Mohammed Hasanuzzaman, and Sriparna Saha. 2020. Improving depression level estimation by concurrently learning emotion intensity. *IEEE Computational Intelligence Magazine*, 15(3):47–59.

Syed Arbaaz Qureshi, Gaël Dias, Sriparna Saha, and Mohammed Hasanuzzaman. 2021. Gender-aware estimation of depression severity level in a multimodal setting. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Syed Arbaaz Qureshi, Sriparna Saha, Mohammed Hasanuzzaman, and Gaël Dias. 2019. Multitask representation learning for multimodal estimation of depression level. *IEEE Intelligent Systems*, 34(5):45–52.

Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg. 2019. Multi-level attention network using text, audio and video for depression prediction. In *9th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, page 81–88.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3980–3990.

Cheolmin Shin, Seung-Hoon Lee, Kyu-Man Han, Ho-Kyoung Yoon, and Changsu Han. 2019. Comparison of the usefulness of the phq-8 and phq-9 for screening for major depressive disorder: Analysis of psychiatric outpatient data. *Psychiatry Investigation*, 16(4):300–305.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *57th Conference of the Association for Computational Linguistics (ACL)*, pages 2895–2905.

Joe Stacey, Yonatan Belinkov, and Marek Rei. 2022. Supervising model attention with human explanations for robust natural language inference. In *AAAI conference on artificial intelligence (AAAI)*, volume 36, pages 11349–11357.

University of Tartu. 2018. UT rocket.

Shanshan Wang, Zhumin Chen, Zhaochun Ren, Huasheng Liang, Qiang Yan, and Pengjie Ren. 2022. Paying more attention to self-attention: Improving pre-trained language models via attention guiding. *arXiv preprint arXiv:2204.02922*.

Danai Xezonaki, Georgios Paraskevopoulos, Alexandros Potamianos, and Shrikanth Narayanan. 2020. Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews. In *Interspeech (INTERSPEECH)*, pages 4556–4560.

Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. In *28th International Conference on Computational Linguistics (COLING)*, pages 696–709.