Evaluating Lexicon Incorporation for Depression Symptom Estimation

Kirill Milintsevich[1,2] and **Gaël Dias**[1] and **Kairit Sirts**[2]
[1]Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, France
[2]Institute of Computer Science, University of Tartu, Estonia
{first_name}.{last_name}@{unicaen.fr[1]|ut.ee[2]}

Abstract

This paper explores the impact of incorporating sentiment, emotion, and domain-specific lexicons into a transformer-based model for depression symptom estimation. Lexicon information is added by marking the words in the input transcripts of patient-therapist conversations as well as in social media posts. Overall results show that the introduction of external knowledge within pre-trained language models can be beneficial for prediction performance, while different lexicons show distinct behaviours depending on the targeted task. Additionally, new state-of-the-art results are obtained for the estimation of depression level over patient-therapist interviews.

Illustration of the lexicon-based input marking
a) i'm pretty much good because see by me being a bus operator you run into circumstances and situations you gotta remain calm and still remain professional at the same time
b) i'm @ pretty @ much @ good @ because see by me being a bus operator you run into circumstances and situations you gotta remain @ calm @ and still remain professional at the same time
c) i'm @ pretty @ much @ good @ because see by me being a bus operator you run into circumstances and situations you gotta remain @ calm @ and still remain @ professional @ at the same @ time @

Table 1: Example of input marking. Text a) is the original text without markings, b) and c) show text with terms from AFINN and NRC lexicons.

1 Introduction

Considerable interest has emerged in using natural language processing to unobtrusively infer one's mental health condition (Chancellor and De Choudhury, 2020). A majority of studies have focused on predicting major depressive disorder (MDD) either as a symptom-based estimation (Yadav et al., 2020; Milintsevich et al., 2023) or a binary classification problem (Burdisso et al., 2023; Xezonaki et al., 2020). Both clinically motivated research initiatives and social media studies have emerged. In the latter case, Twitter (Zhang et al., 2023a), Reddit (Gupta et al., 2022) and depression-related forums (Yao et al., 2021) have fostered attention. In the former case, recorded patient-therapist conversations are transcribed and associated with self-assessment depression questionnaires, such as PHQ-8 (Kroenke et al., 2009) or BDI (Beck et al., 1988).

The DAIC-WOZ dataset (Gratch et al., 2014) has mostly been studied within the context of clinical research. Different works have been proposed to automatically infer depression level on this dataset: multi-modal (Qureshi et al., 2019; Wei et al., 2022) and text-based architectures (Li et al., 2023; Agarwal et al., 2022). The PRIMATE dataset (Gupta et al., 2022) has also received recent attention within the context of early symptom prediction on social media posts. The most comprehensive work on this dataset is proposed by Zhang et al. (2023a), which defines a context- and PHQ-aware transformer-based architecture.

People with MDD have shown increased use of negative emotional words and decreased use of positive emotional words (Rude et al., 2004; Savekar et al., 2023). In this line, Xezonaki et al. (2020) and Qureshi et al. (2020) used feature-level and task fusion of emotion and sentiment knowledge and showed improved performance for depression estimation. However, these works, along with other studies on social media mental health data (Zhang et al., 2023b), have used pre-transformer era neural architectures. Recent state-of-the-art approaches that rely on transformer-based pre-trained language models (PLMs) have not explored external knowledge fusion (Milintsevich et al., 2023).

In this paper, we investigate whether pre-trained language models could benefit from

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 322–328

Lexicon	PHQ-8	Train	Dev	Test
AFINN	≥ 10	8.4	7.6	8.0
	< 10	8.2	7.6	7.9
NRC	≥ 10	7.6	†6.8	†7.1
	< 10	7.7	†7.6	†7.6
SDD	≥ 10	†0.6	0.4	0.5
	< 10	†0.4	0.3	0.4

Table 2: Proportion of marked words for each lexicon over the DAIC-WOZ. Reported values are in percentage. † shows if the difference between the depressed and non-depressed populations is statistically significant.

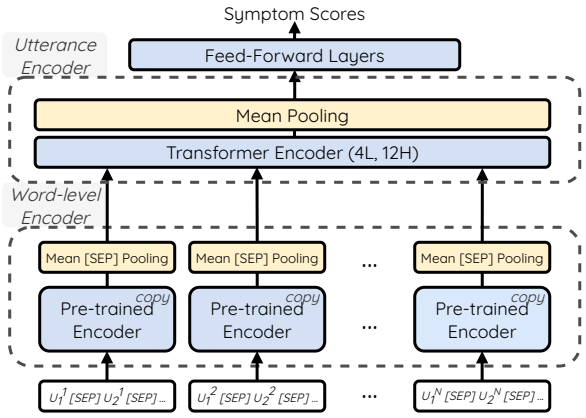

Figure 1: Overview of the model architecture. U_i^N stands for i-th utterance of N-th input. *Symptom Scores* are $||L||$ real numbers, where $||L||$ is the number of symptoms to predict.

the introduction of emotional, sentimental, and domain-specific external knowledge from the lexicons: AFINN (Nielsen, 2011), NRC (Mohammad and Turney, 2013) and SDD (Yazdavar et al., 2017). Introducing this external knowledge into a transformer-based model is feature-level and is achieved by modifying the input with specific markers that highlight spans of text, as shown in Table 1, inspired by the works of Wang et al. (2021) and Zhou and Chen (2022). This approach does not require any modification to the model's architecture, such as changing attention mechanism (Li et al., 2021; Wang et al., 2022) or adding new layers (Bai et al., 2022); it also keeps the model's vocabulary unchanged unlike Zhong and Chen (2021).

Results on the DAIC-WOZ dataset show that the performance of transformer-based models is impacted by the added lexicon information (especially sentiment), and new state-of-the-art values can be obtained from the combination of the three lexicons. However, such results are less expressive for the PRIMATE dataset, with slight improvements induced by the introduction of external information. Overall, the improvement in predicting particular symptoms evidences that lexicon information can be helpful, provided that its content closely corresponds to the targeted task.

2 Methodology

Data. In this work, we use two depression datasets: DAIC-WOZ (Gratch et al., 2014) and PRIMATE (Gupta et al., 2022). The DAIC-WOZ dataset contains 189 clinical interviews in a dialogue format. Each interview has two actors: a human-controlled virtual therapist and a participant. The dataset is distributed in pre-determined splits, such that 107 interviews are used for training, 35 for validation, and 47 for testing. Each interview in the dataset is accompanied with a PHQ-8 assessment, which consists of eight questions inquiring about symptoms. Each question is scored from 0 to 3 on a Likert scale, and the total PHQ score ranging from 0 to 24 is the sum of the eight symptom scores. According to a standard cutoff score of 10, the interviews can be divided into diagnostic classes, where subjects with PHQ-8 total score < 10 are considered non-depressed, and those with score ≥ 10 are categorized as depressed. The eight listed symptoms are: `LOI` (lack of interest), `DEP` (feeling down), `SLE` (sleeping disorder), `ENE` (lack of energy), `EAT` (eating disorder), `LSE` (low self-esteem), `CON` (concentration problem), `MOV` (hyper/lower activity).

The PRIMATE dataset is based on Reddit posts from depression-related communities, or subreddits, in which people describe their health conditions. A total of 2003 posts were manually annotated with binary labels for each individual symptom from the PHQ-9 (Kroenke et al., 2001), each label signifying whether the corresponding symptom is discussed in the post or not. PHQ-9 has the same first eight symptoms as PHQ-8 and one additional `SUI` (suicidal thoughts). The data was labeled by five crowd workers and verified by a mental health professional. The dataset is not pre-split into the train, validation, and test sets, so we randomly take 1601, 201, and 201 posts for each split accordingly.

Model architecture. To encode the interview transcripts, we adopt the hierarchical model from (Milintsevich et al., 2023). In their model, the interview is first split utterance-by-utterance, with each utterance processed by a word-level encoder.

Model	LOI	DEP	SLE	ENE	EAT	LSE	CON	MOV	PHQ-8
BERT	$0.56_{\pm.05}$	$\mathbf{0.63_{\pm.02}}$	$0.77_{\pm.05}$	$0.87_{\pm.04}$	$\mathbf{0.81_{\pm.03}}$	$0.78_{\pm.06}$	$0.74_{\pm.01}$	$0.34_{\pm.01}$	$4.38_{\pm.21}$
+SDD	$0.70_{\pm.02}$	$0.88_{\pm.05}$	$0.94_{\pm.05}$	$0.94_{\pm.04}$	$1.00_{\pm.07}$	$0.97_{\pm.04}$	$0.87_{\pm.02}$	$0.34_{\pm.00}$	$5.60_{\pm.18}$
+AFINN	$\mathbf{0.50_{\pm.03}}$	$0.70_{\pm.03}$	$0.79_{\pm.03}$	$0.81_{\pm.04}$	$0.85_{\pm.03}$	$0.72_{\pm.02}$	$0.77_{\pm.02}$	$0.34_{\pm.00}$	$4.56_{\pm.22}$
+NRC	$\mathbf{0.50_{\pm.03}}$	$0.66_{\pm.05}$	$\mathbf{0.73_{\pm.05}}$	$0.77_{\pm.03}$	$\mathbf{0.81_{\pm.05}}$	$0.71_{\pm.07}$	$\mathbf{0.73_{\pm.05}}$	$0.34_{\pm.00}$	$\mathbf{4.31_{\pm.18}}$
+ALL	$\mathbf{0.50_{\pm.04}}$	$0.69_{\pm.03}$	$0.81_{\pm.12}$	$\mathbf{0.74_{\pm.06}}$	$\mathbf{0.81_{\pm.07}}$	$\mathbf{0.69_{\pm.05}}$	$0.74_{\pm.03}$	$0.34_{\pm.00}$	$4.56_{\pm.42}$
MEBERT	$0.59_{\pm.02}$	$0.64_{\pm.06}$	$0.91_{\pm.05}$	$0.92_{\pm.04}$	$0.89_{\pm.04}$	$0.71_{\pm.02}$	$0.71_{\pm.04}$	$0.35_{\pm.01}$	$4.71_{\pm.23}$
+SDD	$0.69_{\pm.07}$	$0.72_{\pm.08}$	$0.89_{\pm.07}$	$0.92_{\pm.02}$	$0.93_{\pm.07}$	$0.85_{\pm.07}$	$0.78_{\pm.06}$	$0.34_{\pm.00}$	$5.07_{\pm.38}$
+AFINN	$0.48_{\pm.04}$	$0.62_{\pm.02}$	$0.71_{\pm.05}$	$0.78_{\pm.04}$	$0.79_{\pm.03}$	$0.70_{\pm.03}$	$0.74_{\pm.03}$	$0.34_{\pm.00}$	$4.27_{\pm.22}$
+NRC	$0.60_{\pm.05}$	$0.68_{\pm.03}$	$0.71_{\pm.05}$	$0.78_{\pm.04}$	$0.80_{\pm.08}$	$0.74_{\pm.02}$	$0.71_{\pm.05}$	$0.34_{\pm.00}$	$4.35_{\pm.26}$
+ALL	$\mathbf{0.44_{\pm.06}}$	$\mathbf{0.55_{\pm.04}}$	$\mathbf{0.63_{\pm.06}}$	$\mathbf{0.72_{\pm.07}}$	$\mathbf{0.69_{\pm.03}}$	$\mathbf{0.67_{\pm.04}}$	$\mathbf{0.67_{\pm.03}}$	$0.34_{\pm.00}$	$\mathbf{3.59_{\pm.31}}$
SOTA	$0.53_{\pm.05}$	$\mathbf{0.55_{\pm.03}}$	$0.75_{\pm.07}$	$\mathbf{0.64_{\pm.03}}$	$0.81_{\pm.05}$	$\mathbf{0.62_{\pm.02}}$	$0.83_{\pm.04}$	$0.44_{\pm.02}$	$3.78_{\pm.13}$

Table 3: Results for the DAIC-WOZ test set. The mean MAE and standard deviation are reported for five runs. The best MAE for each symptom is **in bold**. SOTA means current state-of-the-art results in the literature (Milintsevich et al., 2023).

All utterance representations are then concatenated into one sequence, later processed by an utterance-level encoder. In the end, the classification head produces a real number in the range from 0 to 3 for each symptom. Several changes are made to the original architecture to gain training efficiency. First, the BiLSTM utterance-level encoder is replaced with a randomly initialized 4-layer 12-head transformer encoder. Second, we change the way the input data is represented. In the original model, each utterance of the interview is encoded separately by a word-level encoder. This is far from optimal since most of the utterances are short (<10 tokens), thus, a lot of computation is wasted on padding tokens. Instead, the utterances are concatenated into one input text separated by the `[SEP]` special token. This way, the number of passes through the encoder is reduced from the number of utterances K to $\bar{K}$, defined as in Equation 1, where $|U_i|$ is the number of tokens in an utterance and m is the maximum input length of the word-level encoder.

$$\bar{K} = \left\lceil \frac{\sum (|U_i| + 1)}{m} \right\rceil \quad (1)$$

In practice, it reduces the number of word-level encoder passes by $\sim$40 times for each input. After, we perform the *Mean* `[SEP]` *pooling* on the tokens representing each utterance to get the final utterance representation. The overview of the model architecture is presented in Figure 1.

Lexicons. To incorporate the external knowledge into the model, we use three lexicons: AFINN (Nielsen, 2011), NRC (Mohammad and Turney, 2013), and SDD (Yazdavar et al., 2017). AFINN is a sentiment lexicon that includes a list of 2,477 terms manually rated for the sentiment valence with a value between -5 (negative) and $+5$ (positive). Nielsen (2011) used Twitter postings together with different word lists as a source for the lexicon. NRC is a word-emotion association lexicon that is a list of 14,182 words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). Mohammad and Turney (2013) compiled terms from Macquarie Thesaurus (Bernard, 1986), WordNet Affect Lexicon (Strapparava and Valitutti, 2004), and General Inquirer (Stone et al., 1966) and labeled them with the help of crowd-sourced workers. SDD is a part of the Social-media Depression Detector and is a lexicon of more than 1,620 depression-related words and phrases created in collaboration with a psychologist clinician.

Input marking. In particular, we employ the technique proposed by Zhou and Chen (Zhou and Chen, 2022) to identify and annotate the lexicon words in the input text. It involves marking a lexicon word using the "@" token on either side (see Table 1 for examples). We chose the "@" token for marking since it is not present in the data but included in the model's vocabulary. This way, the pre-trained model's architecture remains unchanged[1]. The proportion of marked words within the DAIC-WOZ is illustrated in Table 2, where the statistical test is Student's t-test with p-value < 0.05.

[1]Typed marking strategies that include emotion and sentiment values have also been tested and provided no additional insights compared to the simple input marking.

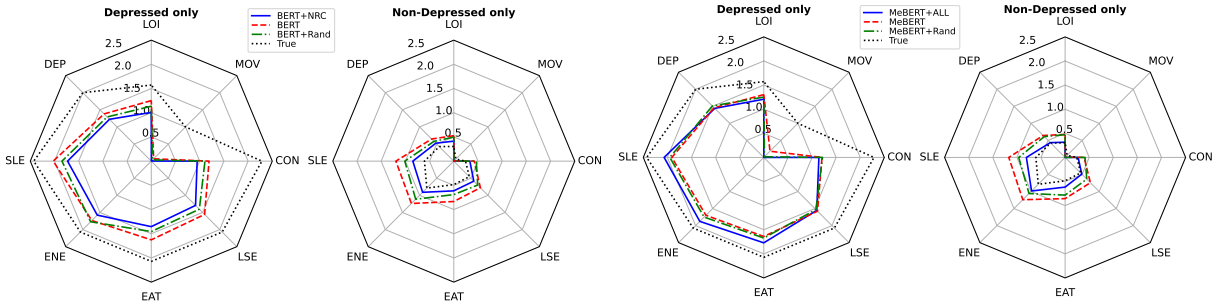

Figure 2: Average predicted values for depressed and non-depressed patients of the DAIC-WOZ test set.

Model	LOI	DEP	SLE	ENE	EAT	LSE	CON	MOV	SUI
BERT	**0.59±.03**	**0.65±.03**	0.81±.01	0.62±.02	0.75±.06	0.60±.02	**0.65±.01**	0.81±.01	0.82±.01
+SDD	0.58±.03	0.62±.02	0.81±.01	**0.64±.03**	0.74±.03	**0.63±.03**	0.63±.03	**0.82±.02**	0.82±.01
+AFINN	0.57±.03	0.60±.03	0.80±.02	0.62±.02	0.76±.02	0.59±.03	0.64±.01	0.81±.02	**0.83±.01**
+NRC	0.55±.04	0.62±.04	**0.82±.01**	0.60±.02	0.79±.04	0.59±.03	0.61±.04	0.80±.01	0.82±.02
+ALL	0.56±.05	0.63±.02	0.79±.02	0.61±.02	**0.80±.02**	0.58±.03	0.61±.01	**0.82±.01**	0.82±.02
MeBERT	**0.58±.03**	0.58±.02	0.82±.02	0.62±.01	0.78±.03	0.60±.04	0.62±.03	**0.82±.01**	0.84±.01
+SDD	0.53±.04	**0.60±.02**	**0.83±.01**	0.62±.02	0.79±.01	0.60±.02	0.61±.03	0.81±.02	**0.86±.01**
+AFINN	0.57±.03	0.55±.04	**0.83±.01**	0.62±.02	0.79±.01	**0.63±.02**	0.58±.02	0.81±.02	0.85±.02
+NRC	0.57±.03	0.58±.03	0.82±.02	**0.63±.03**	0.79±.02	**0.63±.01**	0.61±.03	0.80±.02	0.85±.01
+ALL	0.56±.03	0.59±.04	0.80±.02	0.62±.02	**0.80±.02**	0.61±.01	**0.63±.02**	**0.82±.02**	0.84±.01

Table 4: Results for the PRIMATE test set. The mean macro-F1 score is reported for five runs. The best macro-F1 for each symptom is **in bold**. As standard splits are not provided, we cannot present SOTA results. As standard splits are not provided, we cannot present SOTA results.

Experimental setup. We used two pre-trained models in the word-level encoder of our architecture: BERT-Base model (Devlin et al., 2018) and MentalBERT (Ji et al., 2022). We refer to them as **BERT** and **MeBERT** further on. Both models share the same architecture; however, BERT was pre-trained on general domain data, while MeBERT used mental health-related data, mostly based on Reddit. Each model is finetuned with the same hyperparameters (mostly following Mosbach et al., 2020) and different input markings. For example, the BERT+SDD model uses BERT as a pre-trained model and SDD lexicon for input marking. +ALL models use a union of all three lexicons. All models are trained with a mini-batch size of 16, PyTorch realization of AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of $2 \cdot 10^{-5}$ and linear scheduler with a warm-up ratio of 0.1. For the word-level PLMs, only their attention layers are finetuned. The utterance-level encoder is randomly initialized based on the transformer encoder architecture with the following hyperparameters: 4 layers, 12 attention heads, hidden dimensions of encoder and pooler layers of 768, intermediate hidden dimension of 1536. The rest of the hyperparameters follow the default `BertConfig` from the HuggingFace Transformers library (Wolf et al., 2020). For the DAIC-WOZ dataset, results are evaluated with micro-averaged mean absolute error (MAE). Symptom-based errors are calculated for each symptom individually. PHQ-8 score is obtained by summing the eight symptom scores, and MAE for PHQ-8 is calculated on this summation. We evaluate results on the PRIMATE dataset with a macro-averaged F1 score.

3 Results and Discussion

Table 3 shows the results for the DAIC-WOZ test set. For the BERT model, the lexicon-based input marking brings slight overall improvement when AFINN or NRC lexicons are introduced. Most notably, the NRC input marking shows improved or equal MAE for all symptom scores except DEP. The combination of all lexicons is marginally beneficial overall, and results have deteriorated with the exclusive introduction of the SDD lexicon. On the other hand, for the MeBERT model, the combination of all the lexicons produces the best results overall, both symptom-wise and for the global PHQ-8 score. Furthermore, both AFINN and NRC

lexicons improve the prediction for the MeBERT model, similar to the BERT model. Also, when only the SDD lexicon is used for input marking, the model shows worse performance than the baseline setting.

Figure 2 depicts a more detailed overview of the best-performing models: BERT+NRC and MeBERT+ALL. Additionally, we finetune the +Rand version of both BERT and MeBERT to verify if the improvement comes only from the input marking by randomly marking 8% of the words in each interview. From the results, the improvement for the BERT+NRC model comes from the non-depressed population. MeBERT+All model, however, improves for both depressed and non-depressed populations and is less sensitive to the marking bias. Interestingly, +Rand models show some improvement for the non-depressed population, suggesting that input markings alone act as a regularizer.

Table 4 shows the results for the PRIMATE test set. Contrary to the results from Table 3, introducing external knowledge does not clearly improve performances. The models that use the lexicon input marking show signs of improvement for some symptoms, but it is largely inconsistent. Unlike for the DAIC-WOZ, the SDD-based input marking provides the best F1 score for three symptoms, both for BERT and MentalBERT models, while the benefits of AFINN and NRC are limited or absent and spread over symptoms.

The results from the DAIC-WOZ show that PLMs can indeed benefit from the introduction of external knowledge about the sentiment and emotional value of the words. Surprisingly, the introduction of the depression-specific lexicon had the opposite effect. We hypothesize that two reasons could cause it. First, as seen in Table 2, SDD covers less than 0.5% of words in the interview, almost 15 times less than AFINN and NRC. Thus, the introduced signal might be too weak for the model to learn. Second, the SDD lexicon was based on Twitter data, while DAIC-WOZ contains transcripts of real conversations. From our observations, the people describe their problems more explicitly in their social media posts. At the same time, DAIC-WOZ conversations are more generally themed, and the PHQ-8 scores are based on the person's self-assessment test rather than the conversations themselves. This brings us back to the conceptual difference between the DAIC-WOZ and PRIMATE datasets. While the first one aims at establishing the link between the underlying person's mental condition and their speech, the latter one sets a goal of detecting whether a particular symptom is mentioned in the text. In addition, the PRIMATE dataset is annotated by layman crowd workers, and the labels are not consistent and contain inevitable mistakes (Milintsevich et al., 2024). This might explain the reason behind the greater impact of the AFINN and NRC lexicons for modeling the DAIC-WOZ dataset.

4 Conclusion

This paper targets lexicon incorporation in transformer-based models for symptom-based depression estimation. The external information is supplied through a marking strategy, which avoids any modification to the model's architecture. The set of endeavoured experiments shows that introducing sentimental, emotional and/or domain-specific lexicons can correlate with overall performance improvement if adapted to the targeted task[2].

Limitations

The main limitation in automated clinical mental health assessment with natural language processing is the difficulty of acquiring and accessing large quantities of data. DAIC-WOZ and PRIMATE are rare exceptions as it is publicly available and clinically verified. However, DAIC-WOZ, in particular, suffers from a small number of data points that makes it hard to train and validate hypotheses, as both validation and test sets are particularly small. As a consequence, this piece of research requires further validation on a larger body of clinical data.

Ethical Considerations

We acknowledge the potential ethical aspects of the work that studies the methods to unobtrusively detect someone's mental health status. Here, we are using publicly available datasets collected for research purposes. Also, the lexicons we use are publicly available and have not been composed based on private confidential material. If such a system that could predict the presence of depression symptoms based on actual clinical interviews would be deployed in practice, it would require the informed consent of all participants involved

[2]Source code is available here: https://github.com/501Good/dialogue-classifier.

as well as the understanding of the validity boundaries of such systems, meaning that the predictions of such systems cannot replace the assessment of trained clinicians, but rather assist them in their activities.

Acknowledgements

This research was supported by the Estonian Research Council Grant PSG721 and 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).

References

Navneet Agarwal, Gaël Dias, and Sonia Dollfus. 2022. Agent-based splitting of patient-therapist interviews for depression estimation. In *PAI4MH @ 36th Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, USA.

Jiangang Bai, Yujing Wang, Hong Sun, Ruonan Wu, Tianmeng Yang, Pengfei Tang, Defu Cao, Mingliang Zhang1, Yunhai Tong, Yaming Yang, Jing Bai, Ruofei Zhang, Hao Sun, and Wei Shen. 2022. Enhancing self-attention with knowledge-assisted attention maps. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 107–115, Seattle, United States. Association for Computational Linguistics.

Aaron T Beck, Robert A Steer, and Margery G Carbin. 1988. Psychometric properties of the beck depression inventory: Twenty-five years of evaluation. *Clinical psychology review*, 8(1):77–100.

J.R.L. Bernard. 1986. *The MacQuarrie Thesaurus: The Book of Words*. Macquarie Library.

Sergio Burdisso, Esaú Villatoro-Tello, Srikanth Madikeri, and Petr Motlicek. 2023. Node-weighted graph convolutional network for depression detection in transcribed clinical interviews. In *INTERSPEECH*.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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 *Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Shrey Gupta, Anmol Agarwal, Manas Gaur, Kaushik Roy, Vignesh Narayanan, Ponnurangam Kumaraguru, and Amit Sheth. 2022. Learning to automate follow-up question generation using process knowledge for depression triage on Reddit posts. In *Eighth Workshop on Computational Linguistics and Clinical Psychology (CLPSY)*, pages 137–147, Seattle, USA. Association for Computational Linguistics.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of LREC*.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B.W. Williams, Joyce T. Berry, and Ali H. Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1-3):163–173.

Mingzheng Li, Xiao Sun, and Meng Wang. 2023. Detecting depression with heterogeneous graph neural network in clinical interview transcript. *IEEE Transactions on Computational Social Systems*.

Zhongli Li, Qingyu Zhou, Chao Li, Ke Xu, and Yunbo Cao. 2021. Improving BERT with syntax-aware local attention. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 645–653. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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

Kirill Milintsevich, Kairit Sirts, and Gaël Dias. 2024. Your model is not predicting depression well and that is why: A case study of PRIMATE dataset. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 166–171, St. Julians, Malta. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

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

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Anbu Savekar, Shashikanta Tarai, and Moksha Singh. 2023. Structural and functional markers of language signify the symptomatic effect of depression: A systematic literature review. *European Journal of Applied Linguistics*, 11(1):190–224.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

University of Tartu. 2018. UT rocket.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 1405–1418. Association for Computational Linguistics.

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*.

Ping-Cheng Wei, Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. 2022. Multi-modal depression estimation based on sub-attentional fusion. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 623–639.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 38–45, Online. Association for Computational Linguistics.

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*, 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, Barcelona, Spain.

Xiaoxu Yao, Guang Yu, Jingyun Tang, and Jialing Zhang. 2021. Extracting depressive symptoms and their associations from an online depression community. *Computers in human behavior*, 120:106734.

Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 1191–1198.

Tianlin Zhang, Kailai Yang, Hassan Alhuzali, Boyang Liu, and Sophia Ananiadou. 2023a. Phq-aware depressive symptoms identification with similarity contrastive learning on social media. *Information Processing & Management*, 60(5):103417.

Tianlin Zhang, Kailai Yang, Shaoxiong Ji, and Sophia Ananiadou. 2023b. Emotion fusion for mental illness detection from social media: A survey. *Information Fusion*, 92:231–246.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 50–61. Association for Computational Linguistics.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, pages 161–168. Association for Computational Linguistics.