Analysing Relevance of Discourse Structure for Improved Mental Health Estimation

Navneet Agarwal[1] and **Gaël Dias**[1] and **Sonia Dollfus**[2]
[1]Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France.
[2]CHU de Caen, Service de Psychiatrie; Normandie Univ, UNICAEN, ISTS,
GIP Cyceron; Normandie Univ, UNICAEN, UFR de Médecine, 14000 Caen, France.

Abstract

Automated depression estimation has received significant research attention in recent years as a result of its growing impact on the global community. Within the context of studies based on patient-therapist interview transcripts, most researchers treat the dyadic discourse as a sequence of unstructured sentences, thus ignoring the discourse structure within the learning process. In this paper we propose Multi-view architectures that divide the input transcript into patient and therapist views based on sentence type in an attempt to utilize symmetric discourse structure for improved model performance. Experiments on DAIC-WOZ dataset for binary classification task within depression estimation show advantages of Multi-view architecture over sequential input representations. Our model also outperforms the current state-of-the-art results and provide new SOTA performance on test set of DAIC-WOZ dataset.

1 Introduction

In recent years, automated depression estimation has attracted significant research initiatives which is unsurprising given the widespread impact and heavy toll of depression. Within the context of depression estimation based on text, two major categories of input exist: (1) social media posts (twitter and reddit) of self-declared patients and (2) clinical interviews between patients and therapist. Detection of depression is a challenging problem with patient-therapist interviews being the common practice to analyse a patient's mental health within clinical setting. Within such dialogues, therapists look for indicative symptoms such as loss of interest, sadness, exhaustion, sleeping and eating disorders, etc. within patient's responses and base their evaluation on this information. Complementary to these interviews, different self-assessment screening tools have also been defined such as the Personal Health Questionnaire depression scale, with PHQ-8 being considered a valid diagnosis and severity measure for depressive disorders (Kroenke, 2012). Throughout the literature, different strategies have been proposed for automatic estimation of depression, which consists of inferring the screening tool score based on the interview transcript. Multi-modal models combine inputs from different modalities (Ray et al., 2019; Qureshi et al., 2019; Niu et al., 2021). Multi-task 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; Oureshi 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 mental health lexicons (Xezonaki et al., 2020). Feature-based solutions compute multiple multi-modal characteristics (Dai et al., 2021). Graph-based systems aim to study complex structures within interview transcripts (Hong et al., 2022; Niu et al., 2021). 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 this extensive list of research initiatives, ways to express the structure of an input transcript remains a relatively unexplored research direction. Indeed, most related works treat the overall transcript as a sequence of sentences taking into account the information contained in therapist questions and patient responses. These models disregard interview structure and consider it to be an unstructured list of sentences, forcing the model to learn inter-dependencies within the discourse. In this paper we argue that discourse structure combined with sentence type can improve models learning ability by reducing the number of noisy transactions within the data. In order to validate our hy-

Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 127–132

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 symptom severity category in the DAIC-WOZ dataset, distributed by train, validation and test sets.

pothesis, we design Multi-view architectures that separate a dialogue stream based on sentence type into two different views, i.e. the therapist view and the patient view. As such, the interview structure is taken into account by learning interactions (1) within the views i.e. interactions between questions only and answers only, and (2) between the two views i.e. interactions between the corresponding questions and answers. This allows the models to focus on specific structures of the transcript as well as control the discourse symmetry. Experiments over the DAIC-WOZ dataset show improvements in model performance with multi-view architecture and provide new state of the art results on the test set of DAIC-WOZ dataset.

2 Related work

Different architectures and strategies have been used throughout literature to train automated models for depression estimation based on patient-therapist interviews. 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 different layers to combine several low-level and mid-level features from audio, visual and textual modalities of the participants' inputs. 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. 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, Oureshi 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. Along the same line, Bailey and Plumbley (2021) study gender bias from audio features and 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. Within the same context, Hong et al. (2022) use graphical representation of the input that encodes word level interactions within each transcript. They propose Schema-based Graph Neural Networks (SGNN) and use multiple passes of the message passing mechanism (MPM) (Gilmer et al., 2017; Xu et al., 2019) to update the schema at each node of the text graph.

Burdisso et al. (2023) define a more complex input graph structure that models the interactions between transcripts and a global word graph. They use an inductive version of GCN (Wang et al., 2022) and define w-GCN that mitigates the assumptions of locality and equal importance of self-loops within GCN. 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. 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 mental health domain and propose the use of prefix-tuning as a parameter-efficient way of fine-tuning language models for mental health.

3 Dataset

For our experiments we use the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) dataset which is part of a larger corpus, the Distress Analysis Interview Corpus (DAIC)(Gratch et al., 2014). The dataset contains clinical interviews aimed towards psychological evaluation of participants for detecting conditions such as anxiety, depression and post-traumatic stress disorder. These interviews were collected with the goal of developing a computer agent that interviews participants to identify verbal and non-verbal signs for mental illness(DeVault et al., 2014). In particular, we use Wizard-of-Oz interviews from the dataset which were conducted by virtual agent Ellie, controlled by a human interviewer from another room. These interviews have been transcribed and annotated for a variety of verbal and non-verbal features. Along with the transcripts, the dataset also contains corresponding visual and audio features extracted from the interview recordings. Depression severity is accessed based on PHQ-8 depression scale, and score of 10 is used as threshold to differentiate between depressed and non-depressed participants. The dataset is divided into training, development and test sets containing 107, 35 and 47 interviews respectively. The dataset is biased towards lower PHQ-8 scores with almost 70% data points belonging to negative class in case of binary classification (PHQ-8 score < 10) and only 4 instances with severe depression (PHQ-8 score > 20). Refer table 1 for more details.

4 Methodology

Studies have shown that questions asked by the therapist during an interview contain relevant information and provide context to patient responses. Although Xezonaki et al.(Xezonaki et al., 2020) validate the importance of therapist questions for depression estimation, they represent the input as an unstructured sequence of sentences. Within this paper we emphasise on the importance of discourse structure for better understanding the input text. To take into account both patient and therapist information, while maintaining discourse symmetry and structure, we propose Multi-view architecture that utilize sentence types to divide the interview into different views. Our aim is to use this view based division of the transcript to control the number of noisy interactions, between unrelated questions and answers, learned by sequential models, allowing more efficient training of neural network models.

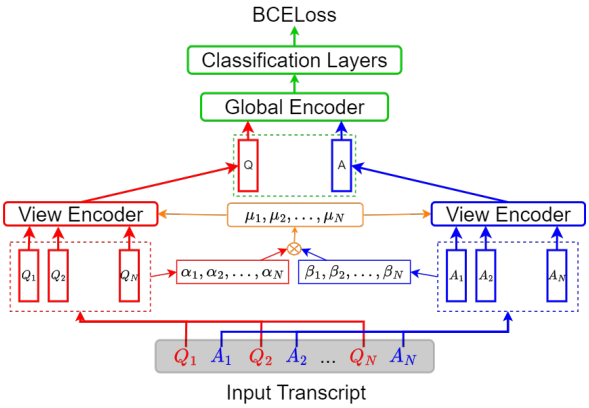

Figure 1: Multi-view architecture based on sentence transformer based text encoding. View specific information in highlighted in red and blue with orange highlighting cross attention and green the global network.

4.1 Multi-view Strategy

Figure 1 illustrates the proposed Multi-view architecture. The underlying idea is to learn transcript level representation of the two views separately before fusing them using *Global encoder* layer to generate transcript level representation of the interview containing information from both questions and answers. In particular, dedicated sub-networks, patient network and therapist network, are defined for processing corresponding view inputs ($Q_1, Q_2, .., Q_N$ and $A_1, A_2, .., A_N$). These sub-networks use multihead attention mechanism in order to combine sentence level text encodings and learn interview level representations, Q and A, of the views. *View encoders* defined within this model also use cross attention for a co-dependent learning of individual views. The coherent structure of a dialogue plays an essential role in global understanding of the message conveyed by the patient. Patient responses often rely on therapist questions in order to contextualize their meaning. This is particularly true for one word responses like "yes", which don't hold much relevance by themselves. As a consequence, tackling the co-dependency between questions and answers[1] is of the

[1]Note also that a question that might not seem to be important, but for which the answer is meaningful, should definitely be highlighted by the learning model.

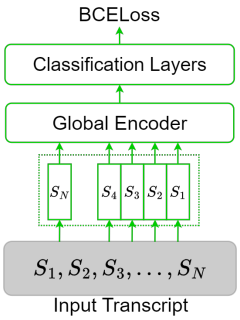

Figure 2: Baseline configuration based on unstructured sequential interview representation.

utmost importance for the learning process. As a consequence, we propose to design a multi-view architecture with inter-view attention (shown with orange color in Figure 1) that transfer attention scores from one view to another, following the cross-attention paradigm (Sood et al., 2020). Formally, attention scores $\mu_1, \mu_2, ..., \mu_M$ are shared between the two *view encoders*, and are the result of function $\mu_i = f(\alpha_i, \beta_i)$ that combines the individual view attention scores α_i and β_i.

Baseline: We define a baseline configuration that uses comparable architecture for a fair comparison. Within this configuration, interviews are treated as a sequence of unstructured sentences and passed through an encoder layer to learn interview level representation which in-turn is passed through classification layers to get final prediction (Figure 2).

5 Experimental Setup

We use sentence-transformers (Reimers and Gurevych, 2019), all-mpnet-base-v2 in particular, for generating sentence level text encodings used within our experiments. Adam optimizer with weighted binary cross entropy loss (BCELoss) is used during training to account for class imbalance in data. Learning rate is treated as a hyperparameter and tuned during training. Both encoders, *global encoder* and *view encoder*, are defined using transformer based Multihead Attention Networks (Vaswani et al., 2017). Cross-attention at *view encoder* level is also defined using multi-head attention mechanism with inputs from both views playing corresponding roles within query, key and value. Various definitions of function $f(\alpha_i, \beta_i)$ were experimented with and f was finally defined as a mean operation. Pytorch framework is used for network definition and training of the models.

6 Results and Analysis

Experiments were conducted on the DAIC-WOZ dataset (Gratch et al., 2014) and the best model is chosen based on macro F1 over the development set and evaluated based on performance on test set. Table 2 compares performance of multi-view model (*Multi-view model*) against the sequential configuration *Sequential model* considered in our work. In particular, the multi-view model evidences better performance compared to sequential input configuration for both evaluation metrics considered in our study. Improvements of 6.6% on macro F1 score and 10.6% on Unweighted Average Recall (UAR) are obtained over the baseline. From the results we can assess that multi-view architectures are a better alternative to process question-answer based interviews, thus highlighting the significance of retaining structural information of a dialogue. In particular, multi-view architectures utilize the interview semantic structure to limit the amount of noisy interactions learned by the model and allowing more efficient learning.

During our experiments with different definitions of cross-attention function $f(\alpha, \beta)$, we observed that results obtained with non-balanced attention functions (i.e. only patient attention, only therapist attention, max) are lower compared to the balanced architectures (i.e. Mean, Learnable). Within non-balanced functions, attention scores are transferred from one view to the other based on hypothesis that only one of the views drives the learning process. Our results confirm that both views, questions and answers, are relevant, and selecting either one as the sole criteria for importance can be counterproductive. $Mean$ function evidenced best performance within our experiments.

Table 2 also shows that our multi-view model provides new state-of-the-art results over the test set of DAIC-WOZ dataset, successfully outperforming recent initiatives with comparable setups (HAN(Xezonaki et al., 2020), HCAN(Mallol-Ragolta et al., 2019)) as well as those relying on external knowledge (HAN+L(Xezonaki et al., 2020)) or different modalities (SVM:m-M&S(Dai et al., 2021)). Note that the reported results are taken directly from the original papers, and that some related work surprisingly do not evidence results over the test split, such as HCAG and HCAG+T (Niu et al., 2021), although they highly perform on the development set.

Architectures	Modality	macro F1		UAR	
		(Dev)	Test	(Dev)	Test
Raw Audio (Bailey and Plumbley, 2021)	Audio	(0.66)	-	-	-
SVM:m-M&S (Dai et al., 2021)	All	(0.96)	0.67	-	-
HCAG (Niu et al., 2021)	Text + Audio	(0.92)	-	(0.92)	-
HCAN (Mallol-Ragolta et al., 2019)	Text	(0.51)	0.63	(0.54)	0.66
HLGAN (Mallol-Ragolta et al., 2019)	Text	(0.60)	0.35	(0.60)	0.33
HAN (Xezonaki et al., 2020)	Text	(0.46)	0.62	(0.48)	0.63
HAN+L (Xezonaki et al., 2020)	Text	(0.62)	0.70	(0.63)	0.70
HCAG+T (Niu et al., 2021)	Text	(0.77)	-	(0.82)	-
Symptom prediction (Milintsevich et al., 2023)	Text	(0.80)	0.74	-	-
Sequential model	Text	(0.79)	0.75	(0.78)	0.75
Multi-view model	Text	(0.77)	**0.80**	(0.76)	**0.83**

Table 2: SOTA results on DAIC-WOZ. T, V and A stand for Text, Visual and Audio modalities.

7 Conclusion and Future Work

In this paper, we propose a multi-view architecture for automated depression estimation that treats patient-therapist interviews as a combination of two views (therapist questions and patient answers). The underlying idea it to not only use inputs from both agents within the interview (patient and therapist), but also retain the inherent structure of the discourse for improved learning. In particular, the presented multi-view approach allows to handle discourse symmetry as well as discourse structure, thus outperforming the simple encoding of the input data as a sequence of sentences. Results on the DAIC-WOZ show that the multi-view architecture steadily outperforms comparable baselines and evidences new state-of-the-art results. Based on the insightful recent research of Xezonaki et al. (Xezonaki et al., 2020), we plan to further improve our results by incorporating external knowledge from different medical resources, such as lexicon or psychiatrist manual annotation.

8 Acknowledgement

This research is supported by the FHU A^2M^2P project funded by the G4 University Hospitals of Amiens, Caen, Lille and Rouen (France).

9 Limitations

Within this paper we explore the role of interview structure on the learning ability of the neural network models. Results from our experiments show that Multi-view architectures provide a better alternate for combining patient and therapist inputs while taking into account the discourse structure. Multi-view architectures focus on using transcript structure in order to limit noisy interactions within the input. The co-dependency between the corresponding questions and answers within the interview is only modeled using shared attention weights. This limits the models ability to study patient's answers in context of associated therapist questions (and vise-versa), and requires further research into defining a complete solution.

10 Ethical Considerations

Given application in medical domain and the nature of this specific task, data privacy and protection is the biggest concern associated with the field. Depression is a condition rooted within the various aspects of a patients life, consequently, its assessment requires discussing a patient's personal and professional lives. Within our research the original data has already been anonymized and all personal information has been removed.

References

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.

Sergio Burdisso, Esaú VILLATORO-TELLO, Srikanth Madikeri, and Petr Motlicek. 2023. Node-weighted graph convolutional network for depression detection in transcribed clinical interviews. In *Proceedings of Interspeech*.

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 Lhommet, et al. 2014. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *International Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, pages 1061–1068.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *9th International Conference on Language Resources and Evaluation (LREC)*, pages 3123–3128.

Simin Hong, Anthony G. Cohn, and David Crossland Hogg. 2022. Using graph representation learning with schema encoders to measure the severity of depressive symptoms. In *10th 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.

Kurt Kroenke. 2012. Enhancing the clinical utility of depression screening. *Canadian Medical Aassociation Journal*, 184(3):281–282.

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. ISCA.

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

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, 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 on Audio/Visual Emotion Challenge and Workshop (AVEC)*, page 81–88.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 2020. Improving natural language processing tasks with human gaze-guided neural attention. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*.

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.

Kunze Wang, Soyeon Caren Han, and Josiah Poon. 2022. Induct-gcn: Inductive graph convolutional networks for text classification. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1243–1249. IEEE.

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.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How powerful are graph neural networks? In *International Conference on Learning Representations*.