INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Node-weighted Graph Convolutional Network for Depression Detection in Transcribed Clinical Interviews

Sergio Burdisso[⋆,1], *Esaú Villatoro-Tello*[⋆,1], *Srikanth Madikeri*[1], *Petr Motlicek*[1,2]

[1]Idiap Research Institute, Switzerland
[2] Faculty of Information Technology, Brno University of Technology, Czech Republic
{sergio.burdisso, esau.villatoro, srikanth.madikeri, petr.motlicek}@idiap.ch

Abstract

We propose a simple approach for weighting self-connecting edges in a Graph Convolutional Network (GCN) and show its impact on depression detection from transcribed clinical interviews. To this end, we use a GCN for modeling non-consecutive and long-distance semantics to classify the transcriptions into depressed or control subjects. The proposed method aims to mitigate the limiting assumptions of locality and the equal importance of self-connections vs. edges to neighboring nodes in GCNs, while preserving attractive features such as low computational cost, data agnostic, and interpretability capabilities. We perform an exhaustive evaluation in two benchmark datasets. Results show that our approach consistently outperforms the vanilla GCN model as well as previously reported results, achieving an F1=0.84% on both datasets. Finally, a qualitative analysis illustrates the interpretability capabilities of the proposed approach and its alignment with previous findings in psychology.

Index Terms: depression detection, graph neural networks, node weighted graphs, limited training data, interpretability.

1. Introduction

According to the World Health Organization (WHO), an estimated 970 million people in the world are living with a type of mental disorder, being depressive and anxiety disorders the most prevalent [1]. Traditionally, the diagnosis and assessment for depression are done using semi-structured interviews and a Patient Health Questionnaire (PHQ) [2] as main tools, and it is generally based on the judgment of general practitioners. However, practitioners may fail to recognize as many as half of all patients with depression [3]. Therefore, there is an acknowledged necessity for digital solutions for *(i)* assisting practitioners in reducing misdiagnosis, and *(ii)* addressing the burden of mental illness diagnosis and treatment [4, 5, 6].

Previous research has shown that language is a powerful indicator of our personality, social, or emotional status, and mental health [7, 8]. As a result, many work exists at the intersection of artificial intelligence (AI), speech and natural language processing, psycholinguistics, and clinical psychology, showing that screening interviews, projective techniques, and essays writing provide valuable insights into the cognitive and behavioral functioning of subjects [9, 10, 11, 12]. Existing work on depression detection, via the use of textual transcriptions from psychotherapy sessions, varies from sentiment-based approaches [13], going through methods designed to identify relevant vocabulary [10, 14], to various neural network architectures to best model the interviews, including bidirectional LSTM [15], hierarchical attention-based networks [16, 17], and deep neural graph structures [18]. Other studies have experimented with multi-target hierarchical regression models to predict individual depression symptoms, aiming to improve performance by simultaneously predicting both binary diagnostic and depression severity regression scores [19]. Finally, some works have explored the utility of enriching the models with additional (domain-specific) data [17, 20], e.g., incorporating external linguistic knowledge to enforce higher values for attention weights corresponding to salient affective words. Contrary to previous work, our proposed approach has the following salient features: does not require any external resource (data agnostic), does not depend on large pre-trained language models to learn embeddings (low computational cost), and has interpretability capabilities by design, a must in AI-supported diagnosis.

In particular, we propose to use a Graph Convolutional Network (GCN) to classify the transcribed sessions between a therapist and a subject seeking medical attention. Overall, the main contributions of this paper are: (1) a novel weighting approach for self-connection nodes to address the limiting assumptions of locality and the equal importance of self-connections vs. edges to neighboring nodes in GCNs; (2) to the best of our knowledge, we evaluate for the first time an inductive implementation of GCNs in the task of depression detection from transcribed interviews, outperforming previously published results on two benchmark datasets; and (3) we demonstrate the interpretability potential of the proposed model, a key characteristic in AI-supported diagnosis, showing that what the model learned aligns with findings in psychology research.[1]

2. Graph neural network architecture

A Graph Convolutional Network (GCN) is a multilayer neural network that operates directly on a graph and induces embedding vectors of nodes based on the properties of their neighbors [21, 22] (Figure 1). Formally, considering a graph $G = (V, E, A)$, where $V(|V| = n)$ represents the set of nodes, E is the set of edges, and $A \in \mathcal{R}^{n \times n}$ an adjacency matrix representing the edge values between nodes. The propagation rule for learning the new k-dimensional node feature matrix $H^{(l)} \in \mathcal{R}^{n \times k}$ is computed as:

$$H^{(l+1)} = f(H^{(l)}, A) = \sigma(\tilde{A}H^{(l)}W^{(l)}) \quad (1)$$

where $\tilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ represents the normalized symmetric adjacency matrix, $D_{ii} = \sum_j A_{ij}$ is the degree matrix of adjacency matrix A, $W^{(l)}$ depicts the weight to be learned in the l_{th} layer, and σ is an activation function, e.g., ReLU:

This work was supported by Idiap internal funds. ⋆*Corresponding authors*.

[1]Our code is available at https://github.com/idiap/Node_weighted_GCN_for_depression_detection

10.21437/Interspeech.2023-1923

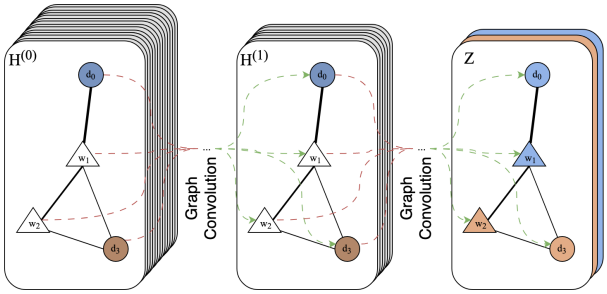


Figure 1: *A two-layer GCN with nodes represented at three levels: initial (e.g. one-hot), $H^{(0)}$, intermediate/hidden, $H^{(1)}$, and output, Z, representations with the probability of each output label. Circles → document nodes & triangles → word nodes.*

$\sigma(x) = \max(0, x)$. In order to use GCNs for text classification [21], we generate a large and heterogeneous text graph that contains word nodes (V_{words}) and *training* document nodes (V_{tr_docs}) so that global word co-occurrences can be explicitly modeled. Accordingly, the entire set of nodes is composed as $V = \{V_{tr\_docs}, V_{words}\}$, i.e. the number of training documents (corpus size) plus the number of unique words (vocabulary size) of the corpus. Particularly, in this work, we use a two-layer GCN defined as:

$$H^{(1)} = \sigma(\tilde{A}H^{(0)}W^{(0)}) \tag{2}$$

$$Z = \text{softmax}(\tilde{A}H^{(1)}W^{(1)}) \tag{3}$$

where $W^{(0)}$ is the learned word embeddings lookup table, and $W^{(1)}$ represents the learned weight matrix in the second layer. Loss is computed by means of the cross-entropy function between Z_i and $Y_i, \forall i \in V_{tr\_docs}$. Intuitively the first layer learns the intermediate representation of the nodes (words and documents) while the second one learns the output representation, as illustrated in Figure 1. Note that in the output representation, label information from the documents has been propagated to the word nodes as output probabilities, allowing the model to learn the relation between words and output labels (e.g. depression or control labels), a key aspect favoring the interpretability of the model (see Section 4.1).

In order to make a fair comparison of the GCN's performance against other classification approaches, in this work we use the inductive version of GCNs as described in [23] instead of the original transductive one [21]. Thus, the initial node feature matrix $H^{(0)}$ is generated such that word node vectors are represented as one-hot vectors, i.e., $H_i^{(0)} = \{0, 1\}^m, \forall i \in V_{words}$, where m is the vocabulary size of the training documents. And, for the representation of document node vectors $H_i^{(0)}, \forall i \in V_{tr\_docs}$ the *term-frequency-inverse document frequency* (TF-IDF) values of the corresponding word in that specific document is used, i.e., $H_{ij}^{(0)} = \text{TF-IDF}(i, j), \forall i, j$ where i and j are a document and a word, respectively.

For the definition of the edge types in A, we consider *(i)* word-to-word, *(ii)* word-to-document, similar to [21, 23]. Our key contribution here is the addition of a new edge type for *(iii)* self-connections, acting as a trade-off parameter in the definition of $\tilde{A}$. Formally, this is expressed as follows:

$$A_{ij} = \begin{cases} \text{PMI}(i,j) & \text{if } i,j \text{ are words \& PMI}(i,j) > 0 \\ \text{PR}(i,j) & \text{if } i,j \text{ are words \& } i = j \\ \text{TF-IDF}_{i,j} & \text{if } i \text{ is document \& } j \text{ is word} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Table 1: *Composition of the DAIC-WOZ and E-DAIC datasets for depressed (D) and control (C) participants. Column 'Category' depicts the number of participants for each class, 'Vocabulary' represents the vocabulary size for each partition, 'LR' indicates the average lexical richness per instance, and 'Duration' indicates the length (hrs:mins:secs) values.*

Dataset		Category	Vocabulary	LR	Duration
DAIC-WOZ	*train*	[D] 30 (*28%*) [C] 77 (*72%*)	$m = 5858$ ($\bar{x}$=621.11)	0.48	26h53m ($\bar{x}$=15m04s)
	dev	[D] 12 (*34%*) [C] 23 (*66%*)	$m = 3268$ ($\bar{x}$=664.22)	0.47	10h01m ($\bar{x}$=17m09s)
E-DAIC	*train*	[D] 37 (*23%*) [C] 126 (*77%*)	$m = 7991$ ($\bar{x}$=576.20)	0.55	43h29m ($\bar{x}$=16m04s)
	dev	[D] 12 (*21%*) [C] 44 (*79%*)	$m = 4201$ ($\bar{x}$= 488.05)	0.58	14h47m ($\bar{x}$=15m50s)
	test	[D] 17 (*30%*) [C] 39 (*70%*)	$m = 4183$ ($\bar{x}$=447.87)	0.63	15h14m ($\bar{x}$=16m19s)

where PMI is the Point-wise Mutual Information and PR stands for the *PageRank* algorithm [24], which given a graph computes the importance of each node in relation to the role it plays on the overall structure of the graph. Intuitively, high PMI values will strongly link word nodes with high semantic correlation, high TF-IDF values will strongly link word nodes to specific document nodes, and high PageRank values will strongly link a node to itself proportionally to its global structural relevance; this last modification aims to mitigate the assumption of locality and equal importance of self-loops, a known limitation in the vanilla GCN [22]. We will refer to this modification as ω-GCN.

Finally, it is worth mentioning that GCNs allow to easily optimize the model efficiency by means of applying simple feature selection techniques to reduce the vocabulary size (i.e. number of word nodes), prior to the graph construction, which has a direct impact on both the number of trainable parameters and model's interpretability (see section 3.2 and 4.1).

3. Experimental setup

3.1. Datasets

For the experiments, we use the Distress Analysis Interview Corpus - wizard of Oz (DAIC-WOZ) dataset [25] and the Extended Distress Analysis Interview Corpus (E-DAIC) [26]. Both datasets contain semi-structured clinical interviews in North American English, performed by an animated virtual interviewer,[2] designed to support the diagnosis of different psychological distress conditions. Datasets are multimodal corpora, composed by audio and video recordings, transcribed text from the interviews, and the Patient Health Questionnaire (PHQ-8 [2]) scores. During our experiments, we only used the speech transcripts from the subjects's responses.

Table 1 shows the composition of the datasets. Observe that the vocabulary size of the DAIC-WOZ is smaller than the E-DAIC vocabulary; suggesting a lesser variation of terminology in the provided answers, also reflected in a lower lexical richness (LR), an indicator of the E-DAIC complexity.

[2]For DAIC-WOZ the virtual interviewer is human-controlled, while for the E-DAIC the virtual interviewer is fully automatic. A portion of the DAIC-WOZ transcriptions were generated using the ELAN tool from the Max Planck Institute for Psycholinguistics [27], while the E-DAIC transcripts were obtained using Google Cloud's ASR service.

3.2. Implementation details

As baseline models, we used different BERT-based models as well as simple models. More precisely, we used six pre-trained transformer-based models (*bert-base-cased*, *bert-base-uncased*, *bert-large-cased*, *bert-large-uncased*, *roberta-base*, *roberta-large*) to which a final linear layer was added to classify the input using, as usual, the *[CLS]* classification special token. In addition, to make the baselines as standard and simple as possible we made use of the *Transformers* Python package [28] *AutoModelForSequenceClassification* class so that the size and number of linear layers are automatically selected according to each model. For each model, we also evaluated two versions, one enabling fine tuning of the base model and another not fine tuning the base model as part of the training process. Regarding simple and classic models, we used a Support Vector Machine (SVM) with linear kernel and Logistic Regression (LR) model, both using TF-IDF-weighted words as features.

For GCN models, the size of nodes' intermediate representation was set to 64, i.e. we set $k = 64$ for the k-dimensional feature matrix $H^{(1)} \in \mathcal{R}^{n \times k}$. We performed a preliminary evaluation varying $k \in \{32, 64, 128, 256, 300\}$ from which 64 showed to consistently be the best performing one. In addition, since GCN models allow us to control the vocabulary size (i.e. number of word nodes), we trained different GCNs using different vocabulary sizes, as with SVM and LR models. Namely, we applied the following feature selection techniques to build the vocabulary: (a) automatic selection based on term weights learned using LR; (b) top-k best selection based on *ANOVA F-value* between words and labels with $k \in \{100, 250, 500, 1000, 1500\}$; and (c) full vocabulary. Trying different sizes allowed to control the complexity of the final model; GCNs with smaller vocabularies have smaller graphs, making them simpler and easier to interpret.

Finally, all neural-based models were implemented using PyTorch while non-neural ones using Scikit-learn. Additionally, for a fair comparison, all the models were optimized on each dataset using *Optuna* [29] with 100 trials for hyperparameter search maximizing the macro averaged F1 score. For all neural-based models AdamW [30] optimizer (β_1=0.9, β_2=0.999, ϵ=1e$-$8) was used with *learning rate* and number of epochs n searched in $\gamma \in [1e{-}7, 1e{-}3]$ and $n \in [1, 10]$, respectively. On the other hand, for non-neural baselines, search was performed varying the regularization parameter $C \in [1e{-}3, 10]$, the class weight (balanced, none) and the penalty norm (L2, L1, L2 + L1, or none). As a result, a total of 40 optimized models were obtained.[3]

4. Results

Table 2 summarizes our results for the experiments on the *dev* partition DAIC-WOZ and on the *dev* and *test* partitions of E-DAIC.[4] For each partition, we divide the table into non-GCN models (i.e., classic and BERT-based baselines and previous research) and GCN models (vanilla GCN and our proposed ω-GCN). In addition to the results, we also report the total number of trainable parameters ('#Params') and the vocabulary size ('Vocab size'). Dashes indicate the corresponding metric is not reported in the original paper, while results marked with $*$ are not directly comparable as the model uses external domain-specific resources. Finally, for each dataset, we only report the best-performing models among all 40 optimized models (see Section 3.2).

Table 2: *Results for dev and test partitions for DAIC-WOZ and E-DAIC datasets respectively. Performance is reported in terms of the F score ($F1$) for both control (C) and depression (D) classes, and the Macro-F for the overall problem (Avg.).*

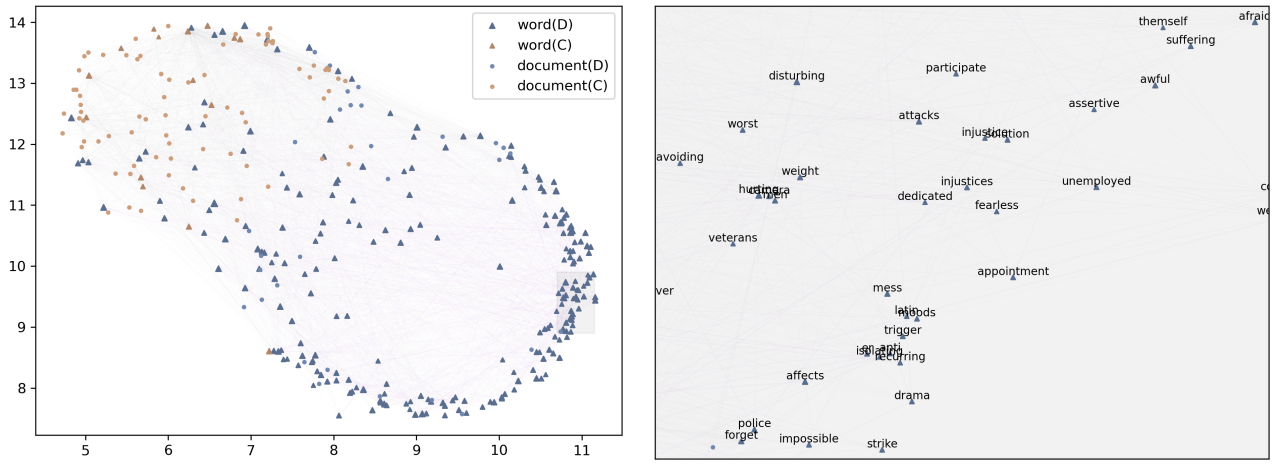
Method	#Params	Vocab size	F1 score		
			Avg.	*D*	*C*
DAIC-WOZ – (dev)					
SVM	1952	1952	0.65	0.50	0.80
LR	250	250	0.60	0.45	0.75
BERT	335M	30522	0.68	0.58	0.78
BERT+FT	335M	30522	0.59	0.53	0.65
HCAG [16]	-	-	0.77	-	-
HAN-L [17]	-	-	0.69	-	-
Symptom-based [19]	-	-	0.75	-	-
IDLV [10]	-	100	0.64	0.52	0.77
vanilla-GCN	375K	5858	0.75	0.67	0.83
ω-GCN	375K	5858	0.76	0.67	0.86
vanilla-GCN	125K	1952	0.68	0.67	0.70
ω-GCN	125K	1952	**0.79**	**0.76**	**0.83**
ω-GCN†	16K	250	**0.84**	**0.80**	**0.89**
E-DAIC – (dev)					
SVM	7991	7991	0.69	0.47	0.91
LR	7991	7991	0.71	0.53	0.90
BERT	108M	28996	0.61	0.46	0.75
BERT+FT	108M	28996	0.70	0.54	0.86
IDLV [10]	-	1000	0.64	0.38	0.90
PV-DM [20]*	-	-	0.90	-	-
vanilla-GCN	511K	7991	0.71	0.50	0.92
ω-GCN	511K	7991	**0.80**	**0.67**	**0.94**
vanilla-GCN	172K	2689	0.58	0.33	0.82
ω-GCN	172K	2698	0.70	0.54	0.86
ω-GCN†	16K	250	0.64	0.43	0.85
E-DAIC – (test)					
SVM	250	250	0.69	0.60	0.78
LR	250	250	0.72	0.63	0.81
BERT	108M	28996	0.49	0.29	0.696
BERT+FT	108M	28996	0.75	0.65	0.85
VADER [13]	-	-	-	0.72	0.85
vanilla-GCN	511K	7991	0.73	0.63	0.83
ω-GCN	511K	7991	0.72	0.63	0.81
vanilla-GCN	172K	2689	0.68	0.62	0.75
ω-GCN	172K	2698	0.73	0.63	0.83
ω-GCN†	16K	250	**0.84**	**0.76**	**0.92**

Overall, we see that the ω-GCN approach consistently outperforms its vanilla version. In addition, the model can outperform baselines and previously reported works when the correct number of features is selected. For instance, on DAIC-WOZ, ω-GCN obtains a macro $F1 = 0.84$ with only top-250 words. On the E-DAIC dataset, the ω-GCN obtains the best performance among the considered methods, with a macro-$F1$ of 0.80 and 0.84 for the *dev* and *test* partitions respectively. However, unlike the DAIC-WOZ *dev* results, reducing the vocabulary size leads to unstable performance between *dev* and *test* sets suggesting models are sensitive to the (reduced) vocabulary discrepancy between the training and evaluation sets, a similar phenomenon as the one reported in [10], where authors argue is due to the complexity of the dataset. We leave exploring methods to mitigate this phenomenon as future work by moving from a purely word-based vocabulary to, for instance, an embedding-powered or sub-word one (e.g. as BERT with WordPiece).

Finally, GCNs have order-of-magnitude fewer parameters than BERT models and are not constrained to a maximum sequence length (e.g. 512 tokens for BERT-based models).

[3] 14 simple baselines (SVM and LR with 7 vocabulary sizes), 12 BERT-based baselines (6 models with/without fine tuning), and 14 GCN models (vanilla GCN and ω-GCN with 7 vocabulary sizes).

[4] DAIC-WOZ *test* partition is not publicly available.



(a) *Overall graph with learned node embeddings*

(b) *Zoomed-in region showing clusters of words (embeddings)*

Figure 2: *Node embbedings learned for DAIC-WOZ. As in Figure 1, circles denote documents, triangles words, and colors denote class ([D] - depression, [C] - control). The gray rectangle in (a) indicates the zoomed region (b). Graph edges are also included.*

4.1. Exploring the model's interpretability

One of the main advantages of the proposed GCN-based approach is that does not sacrifices performance for the sake of transparency. Figure 2 shows the *UMAP* [31] 2-dimensional projection of the 64-dimensional word and document embbedings learned by the best performing ω-GCN model on DAIC-WOZ. More precisely, these embeddings correspond to the intermediate representation $H^{(1)}$, with the 250 word nodes painted with the learned class in the output representation Z. The figure illustrates how the model can make use of the graph structure to learn, in the same latent space, document and word embeddings whose distance is influenced by their mutual relation and the output values. These embeddings allow to identify clusters of strongly related words with high co-occurrence and linked to similar documents in the dataset, i.e., dataset-specific "topics" that experts could potentially use for qualitative analysis. For instance, in DAIC-WOZ, interviews were conducted with war veterans and Figure 2b depicts a few examples of these word clusters —e.g. (1) about "veterans" and words like "worst", "disturbing", "avoiding" and "hurting"; (2) about "police", "strike", "drama", "moods", "trigger", "affects", "moods"; (3) about "attacks", "injustices", "solution"; and (4) about "unemployed", "suffering", "awful", "afraid".

Finally, we performed an analysis of how much of the acquired knowledge by the model fulfills known classical psychological theories/properties. For this, we used the Linguistic Inquiry and Word Count (LIWC) [32] lexical resource, composed of more than 4000 words, categorized into 64 psychological dimensions. Figure 3 shows the result of this analysis. X-axis depicts the psychological dimensions of the words learned by the model, while the Y-axis represents the normalized frequency of the respective dimensions. As shown, the model learned that depressed subjects employ higher frequency dimensions related to affective or emotional processes (affect), cognitive processes (cogmech), relativity (relativ), and negative emotions (negemo). On the contrary, control subjects use more frequently the social processes (social), biological processes (bio), positive emotions (posemo), family and body dimensions. Overall, these findings are aligned with previously reported psychological work [33].

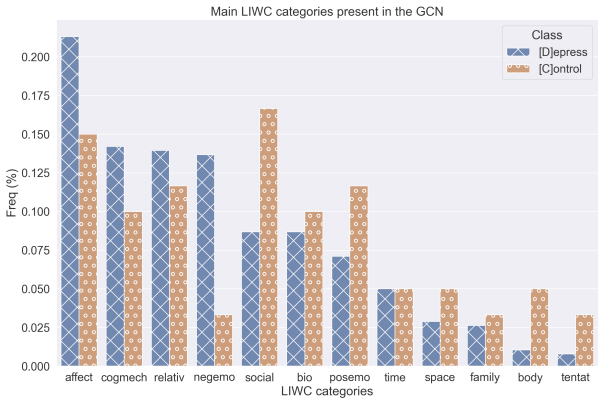


Figure 3: *Psychological dimensions present in the model.*

5. Conclusions

This paper proposes the use of Graph Convolutional Networks to detect depression from transcribed clinical interviews. The proposed approach has some attractive features, including a simple yet novel weighting approach for self-connection edges, a significantly low computational cost in terms of trainable parameters, and interpretability capabilities that help to understand the model's rationale. Evaluation results on two depression-related datasets indicate that the proposed approach is able to consistently outperform its vanilla version. Our best configurations require orders of magnitude fewer trainable parameters than transformer-based models and yet, with the right vocabulary size, are able to obtain better F1 scores than baselines and previously reported results. Finally, an exploration of the interpretability capabilities of the model showed that what it learned from raw data was, in fact, aligned with previously reported work from the psychological theory. As future work, we plan to use different nodes, from simple sub-word nodes to node hierarchies with different types. For instance, the addition of acoustic nodes, as a third type of node, would allow information transfer among acoustic, words and document embeddings.

6. References

[1] World Health Organization, *World mental health report: Transforming mental health for all*. World Health Organization, 2022. [Online]. Available: https://apps.who.int/iris/rest/bitstreams/1433523/retrieve

[2] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The phq-8 as a measure of current depression in the general population," *Journal of affective disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

[3] J. Mitchell, Alex, A. Vaze, and S. Rao, "Clinical diagnosis of depression in primary care: a meta-analysis," *The Lancet*, vol. 374, no. 9690, pp. 609–619, 2009.

[4] T. Wykes, J. Lipshitz, and S. M. Schueller, "Towards the design of ethical standards related to digital mental health and all its applications," *Current Treatment Options in Psychiatry*, vol. 6, no. 3, pp. 232–242, 2019.

[5] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous, "Chatbots and conversational agents in mental health: a review of the psychiatric landscape," *The Canadian Journal of Psychiatry*, vol. 64, no. 7, pp. 456–464, 2019.

[6] T. Koulouri, R. D. Macredie, and D. Olakitan, "Chatbots to support young adults' mental health: An exploratory study of acceptability," *ACM Trans. Interact. Intell. Syst.*, vol. 12, no. 2, jul 2022. [Online]. Available: https://doi.org/10.1145/3485874

[7] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual review of psychology*, vol. 54, no. 1, pp. 547–577, 2003.

[8] A. M. Tackman, D. A. Sbarra, A. L. Carey, M. B. Donnellan, A. B. Horn, N. S. Holtzman, T. S. Edwards, J. W. Pennebaker, and M. R. Mehl, "Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis." *Journal of personality and social psychology*, vol. 116, no. 5, p. 817, 2019.

[9] N. Malandrakis and S. S. Narayanan, "Therapy language analysis using automatically generated psycholinguistic norms," in *Proc. Interspeech 2015*, 2015.

[10] E. Villatoro-Tello, G. Ramírez-de-la Rosa, D. Gática-Pérez, M. Magimai.-Doss, and H. Jiménez-Salazar, "Approximating the mental lexicon from clinical interviews as a support tool for depression detection," in *Proc. ICMI'21*, 2021, p. 557–566. [Online]. Available: https://doi.org/10.1145/3462244.3479896

[11] E. Villatoro-Tello, S. Parida, S. Kumar, and P. Motlicek, "Applying attention-based models for detecting cognitive processes and mental health conditions," *Cognitive Computation*, vol. 13, pp. 1154–1171, 2021.

[12] G. Ramírez-de-la Rosa, H. Jiménez-Salazar, E. Villatoro-Tello, V. Reyes-Meza, and J. Rojas-Avila, "A lexical-availability-based framework from short communications for automatic personality identification," *Cognitive Systems Research*, vol. 79, pp. 126–137, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389041723000062

[13] J. Sawalha, M. Yousefnezhad, Z. Shah, M. R. G. Brown, A. J. Greenshaw, and R. Greiner, "Detecting presence of ptsd using sentiment analysis from text data," *Frontiers in Psychiatry*, vol. 12, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyt.2021.811392

[14] S. G. Burdisso, M. Errecalde, and M. Montes-y Gómez, "A text classification framework for simple and effective early depression detection over social media streams," *Expert Systems with Applications*, vol. 133, pp. 182–197, 2019.

[15] M. Li, H. Xu, W. Liu, and J. Liu, "Bidirectional lstm and attention for depression detection on clinical interview transcripts," *2022 IEEE 10th International Conference on Information, Communication and Networks (ICICN)*, pp. 638–643, 2022.

[16] M. Niu, K. Chen, Q. Chen, and L. Yang, "Hcag: A hierarchical context-aware graph attention model for depression detection," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 4235–4239.

[17] D. Xezonaki, G. Paraskevopoulos, A. Potamianos, and S. S. Narayanan, "Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews," in *Interspeech*, 2020.

[18] S. Hong, A. Cohn, and D. C. Hogg, "Using graph representation learning with schema encoders to measure the severity of depressive symptoms," in *International Conference on Learning Representations*, 2022.

[19] K. Milintsevich, K. Sirts, and G. Dias, "Towards automatic text-based estimation of depression through symptom prediction," *Brain Informatics*, vol. 10, 2023.

[20] Z. Zhang, W. Lin, M. Liu, and M. Mahmoud, "Multimodal deep learning framework for mental disorder recognition," in *2020 15th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2020). IEEE*, 2020.

[21] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 7370–7377.

[22] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

[23] K. Wang, S. C. Han, and J. Poon, "Induct-gcn: Inductive graph convolutional networks for text classification," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 1243–1249.

[24] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[25] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3123–3128.

[26] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 2014, pp. 1061–1068.

[27] H. Brugman and A. Russel, "Annotating multi-media/multi-modal resources with ELAN," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA), May 2004.

[28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[29] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[31] L. McInnes, J. Healy, N. Saul, and L. Großberger, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018. [Online]. Available: https://doi.org/10.21105/joss.00861

[32] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[33] N. Mor and J. Winquist, "Self-focused attention and negative affect: a meta-analysis," *Psychological bulletin*, vol. 128, no. 4, 2002.