Graph Representation Learning for Predicting the Severity of Depressive Symptoms Using Text

Simin Hong, Anthony G. Cohn, David C. Hogg
School of Computing
University of Leeds
{scsho, a.g.cohn, d.c.hogg} @leeds.ac.uk

Abstract

Major depressive disorder (MDD) is a growing problem, which may result in suicide, particularly in the current situation that the world is suffering the COVID-19. Most research relies on utilizing multiple modalities including visual, audio and linguistic features of an interview ranging between 7-33 minutes to make a score-level depression prediction. Since averaging features over an entire interview will lose most temporal information, the system fails to explain why the model made certain predictions. Instead, we assume that there are existing high-level cognitive bias features due to depression underlying an individual's opinions that can be derived from context-aware semantics in a subject's transcript, and these cognitive biases may capture and preserve useful temporal information from text. To test our hypothesis, we propose to first convert an unstructured raw transcript to a text level graph, and then implement a graph neural network (GNN) to learn a complex structure representing cognitive bias features that preserve determinant temporal details of depressive symptoms at different levels over the transcript. We aim to utilize graph representation for learning a mapping encoding high dimensional probability distributions over a text-level graph. Our novel graph-based deep learning model is capable of evaluating depression level for each subject in an end-to-end manner. We showed our results on the DAIC-WOZ benchmark demonstrating the effectiveness of the proposed approach and its superiority over other state-of-the-art methods.

1 Introduction

1.1 Background

Over 300 million people worldwide have been affected by depression which may cause suicide [26], and mental health is becoming more important, particularly under the serious circumstance of COVID-19. Thus, there is a pressing need to find a convenient and automated method to assess depression severity. Medical personnel could thus quicken intervention offering people help promptly, particularly for those who are unaware of which depressed state they are experiencing. Most existing research on automated methods implements multi-modal machine learning methods to capture all potential vocal and facial expression-based features relating to clinical symptoms of depression for MDD detection [3,8,22,24]. Such automated joint feature analyses indeed have improved diagnostic accuracy.

Automatic depression recognition tools for predicting the state of mental health of individual basis can capture verbal behaviors from clinical textual records in a robust, accurate and automatic way. Compared to visual and vocal features, such as those recorded in videos and audios, text-based semantic features are often the most informative indicators obtained by analyzing the patient's textual records. Moreover, some medical research[32] shows that natural language processing techniques can be effective to make inferences about peoples' mental states. One well-known theory posited by Beck [2] points out that depression is instituted by one's view of oneself and can be identified by cognitive biases in one's thoughts. In other words, Beck highlights that depressive symptoms can be captured effectively by cognitive bias features existing in people's minds rather than in people's observable behaviors, i.e., visual behaviors. As a result, these "implicit and complicated depression features" cannot be easily monitored and captured by a sensing tool. Also, some heuristic researches in psychological literature found that Beck's theory can be used to learn from small experiment samples with good confidence [2,3,18,23].

1.2 Challenges and Contributions

Challenges: Both the 6_{th} and 7_{th} International Audio/Video Emotion Challenge (AVEC)[8] provided the same data containing video-based facial actions, audio and the conversation transcribed to text for each sample interview arranging from 7 to 33 minutes, and only one decision is provided for each entire interview. Thus the length of decision unit is much longer than for traditional emotion detection tasks, where their databases usually provide labels for short-term recordings [32]. To solve the challenge of processing and evaluating large amounts of data, how to discover, capture and preserve detailed temporal information over an entire interview are significant. These short-term details within the interview are the most informative when predicting the state of depression of an individual. However, using statistical functions (e.g., max, min, max, etc.,) on short-term features over an entire interviews may lose useful temporal information such as short-term signs in regret, despair and anxiety.

Analyzing a large data volume is typically beneficial for the accuracy, since its contextual information conveys the most relevant evidence for determining depressions at different levels, such as mentioning previous depression diagnoses and ongoing therapy, having sleeping issues and repetitive anxious mood states, etc.. Therefore, it is important to map the whole interview to a high-level feature vector capturing both short-term details and context at a low dimension.

In addition, both the 6_{th} and 7_{th} AVEC challenges contain 107 samples in the training set and 35 samples in a development set, and the database is unevenly distributed (e.g., the number of depression samples in the training set is only 30). With such a small sample size, the number of features should be small to avoid the problems of dimensionality and overfitting. However, the dimensions of audio and video features are very large and therefore, sparse parametrization needs to be taken into account when training machine learning models for learning a joint multi-modal feature vector.

Contributions: In order to overcome the mentioned challenges, we propose a graph-based deep learning model to solve this domain task of quantifying depressive states by using only text. We show some advantages of using graph representation, which facilitates learning text-based semantic features derived from word entities and relations. Since each interview contains hundreds of utterances, extracting short-term details according to utterances (in the form of text) which are not context-oriented may lead to the issue of dimensional explosion and overfitting. For example, both subject 1 and subject 2 mentioned the same word of "hopeless" at an utterance, although it is a strong short-term signal indicating depressive states, weighting their words of "hopeless" at the similar level may cause an error due to different contexts (e.g. ["I am hopeless"] vs. ["he is hopeless"]). In contrast, using a graph representation may allow us to exploit an intuitive and compact data structure for learning to represent typical correlations between both adjacent words and between non-adjacent words by using this occurrence of words (e.g., the "hopeless" is either following the word "I" or "he"). Because of the way in which the graph structure encodes information of the importance among words within the context, exploring the deep learning algorithm on the graph which concentrates on learning to efficiently represent these high-dimensional probability distributions require a very small number of parameters. To our best knowledge, we are the first to present a graph-based deep learning model to perform the task of predicting different depression levels.

1.3 Related Work

Research in measuring the severity of depressive symptoms aims to train a regression model in an automatic way, making use of relevant features extracted from various modalities to predict depression scores [3,8,22,24]. However, applying a feature-fusion approach is vulnerable to bias, since this approach easily fails to perform consistently better results on diagnosing users in the test set[25]. One potential reason would be that this existing approach learns to mapping low-level features from multi-modalities as determinants for estimating the severity of depressive disorders. As a result, this may lead to learn wrong representation encoding inter-subject variability unrelated to depression[12,13,27]. For instance, happiness could be recognized by detecting the feature of smiling in the context of discussing family relationships, exciting experiences, etc. However, it is hard to distinguish such smiling from smiling which takes place in conventional situations such as in a case of greeting someone.

Moreover, current multi-modal machine learning methods[8,10,20,24] aim at learning the "all-for-good representation" by capturing context-unaware features extracted from all modalities of audio,video and text. Their algorithms are designed to follow a clinical psychological finding [6,7,8], which is, all depression-related indicators/markers (e.g. monotone pitch, reduced articulation rate, lower speaking volumes, etc.) across several personal dimensions should be collected to assess the severity of depressive symptoms. Imaging that those "indicators" are represented as "features" (e.g. points, edges, curves, etc.) for detecting a "cat" object in a picture, and then training a deep learning model to capture all those possible low-level features, which are pixels, requires a great amount of training data, because the model has to depend on learning superficial statistical regularities from the cheap data [1]. However, in reality, clinical data is never cheap and always at a premium.

2 Graph-based deep learning algorithm for depression prediction

2.1 text-based semantic features

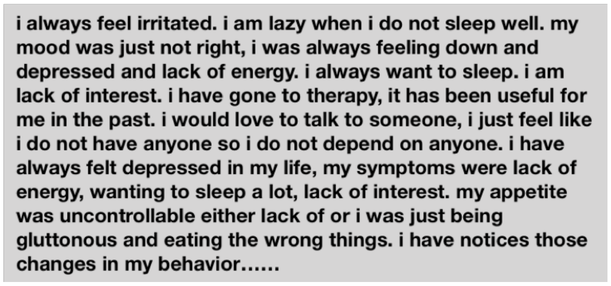
semantic context features: Inspired by the observation that all self-reported text transcriptions involving depression related personal opinions are in the fashion of context- orientation. Furthermore, there are some empirical psychological findings [20] which indicate utterances of depressed people directly and explicitly manifest cognitive biases presenting in their depressed thoughts. These indicators obtaining cognitive bias features are composed of a class of typical linguistic markers and their occurrences interact with other word entities within the context of a depressed mind. Two major types of depression related cognitive biases could be efficiently identified:

1. Self-oriented cognitive bias: One finding emphasizes a person's expression conveying significantly more first person singular pronouns, such as "me", "myself" and "I", and fewer second and third person pronouns, such as "they", "them" or "she" [28]. From the perspective of a clinical psychologist, people with depression repeat this pattern of pronoun usage because they are more focused on themselves, and less connected with others, whereas people who do not exhibit depressive symptoms do not display this preference.

2. Black-and-white cognitive bias: Another highlights a certain style of language [11] which can be utilized to identify depression. Research has found that "absolutist words", such as "always", "never", "nothing" or "completely", are

more effective markers for depression recognition, as depressed patients presumably have more black and white views of the world and this could be manifestly found in their style of language.

According to the above psychological studies [3,4,9,10], the co-occurrences of cognition-based signal words in the utterances of a patient are strongly and explicitly related to that patient's depression level. Our motivation is to learn these key words and correlations between both adjacent words and between non-adjacent words by using this occurrence of these key words, thus we could capture context-based semantic behaviors keeping the most differentiated information relating to depression levels. For example, the occurrence of first person singular pronouns, i.e., "I" and "my", and the occurrence of absolutist word indicators, i.e., "never", "always", enable to efficiently connect their adjacent words within some certain contexts regarding to sleeping problems, fatigue problems, anxiety problems and others. Examples are shown in Figure 1.



i always feel irritated. i am lazy when i do not sleep well. my mood was just not right, i was always feeling down and depressed and lack of energy. i always want to sleep. i am lack of interest. i have gone to therapy, it has been useful for me in the past. i would love to talk to someone, i just feel like i do not have anyone so i do not depend on anyone. i have always felt depressed in my life, my symptoms were lack of energy, wanting to sleep a lot, lack of interest. my appetite was uncontrollable either lack of or i was just being gluttonous and eating the wrong things. i have notices those changes in my behavior......

Figure 1: Example of extraction of a raw transcript

We found that we can encode this kind of information using graphs [21], because graphs are particularly capable of representing strong relational inductive biases, which could perform efficient reasoning by exploiting the graphical structures within text. As a result, we propose to exploit the power of graphs to develop a deep neural network operating over graphs [9,15,21], for the purpose of learning a flexible graph representation for depression prediction task. In our experiments, our approach realized better generalization on limited and unevenly distributed dataset by using only text.

2.2 Method

We first build a text level graph for each individual's self-reported transcript. All parameters for the text level graph are taken from some global-sharing matrices. Then we present the detail of implementing a graph neural network to learn the mapping for predicting depression scores from the context. Our designed architecture is shown in Figure 2.

Building Text Graphs: We regard all the unique word appearing in a transcript as the nodes of a graph. We consider a transcript-level graph has potential word node attributes as:

$$H = (h_1 \ldots, h_l)^T \in \mathbb{R}^{l \times d}$$

where h_i represents a feature vector of the word and l denotes the total number of word nodes of a graph. h_i is initialized

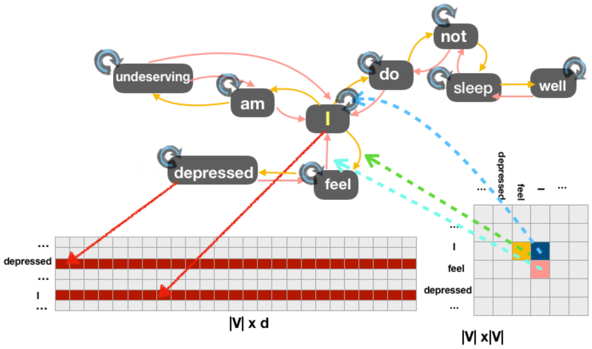


Figure 2: An example structure of a graph for a single sentence, "i am undeserving, i do not sleep well,i feel depressed.", extracted from a transcript of a subject. All the parameters come from the global shared representation matrices, which are shown at the bottom of the figure. Every unique word has a trainable attribute with a d (=200) dimensional vector which is known as word2vec embedding;$|V|$ denotes the total number of unique words in a transcript.

with the pre-trained word2vec embedding (d=200) and can be updated by training.

We then build edges between the nodes and use a reasonably small size sliding window on each text document to collect word co-occurrence statistics for setting edge weight values. We thus define the edge set E, thus:

$$E = \left\{ e_{ij} | i \in [1, l]; j \in [i - p, i + p] \right\}$$

where p denotes the number of adjacent words connected to each word in the graph. Each edge in E has an associated trainable attribute, a real-valued weight.

Let $G = (V, E)$ denote a text level graph. For graph regression modeling, we consider a given a set of graphs $\{G_1, \cdots, G_N\} \subseteq \mathcal{G}$ and their labels $\{y_1, \cdots, y_N\} \subseteq \mathcal{Y}$, where N is the total number of transcriptions (or the population size), we aim to learn a representation vector h_G that helps predict the label of an entire graph, $y_G = g(h_G)$.

Graph Neural Network(GNN): A GNN uses the graph structure and node features h_v to learn a representation vector of an entire graph, h_G. We use the Message Passing Mechanism (**MPM**) [7,14,21] which is employed for convolution. The graph performs message passing between the nodes in order to learn the representation of each node which captures the structural information within its network neighborhood. MPM is defined as:

$$m_v^{(k+1)} = \sum_{u \in N(v)} M^{(k)} \left(h_v^{(k)}, h_u^{(k)}, E \right) \tag{1a}$$

The message function $M^{(k)}$ [14, 21] is applied to push message from the surrounding nodes around the node v and its passing phase runs for $k(= 2)$ iterations through the directed edges. $N(v)$ is a set of nodes adjacent to v. $h_v^{(k)}$is the feature vector of node v at the k_{th} iteration and u is a set of nodes adjacent to v.

$$h_v^{(k+1)} = U^{(k)}\left(m_v^{(k+1)}, h_v^{(k)}\right) \tag{2a}$$

The updated function $U^{(k)}$ [14, 21] is applied to node v to compute an updated node attribute. In this step, we get the new embedding of the node v which is updated by holding messages that encode correlations between itself and its neighbors.

Finally, we obtain the final set of embeddings for each node in the convolution unit at the final layer. We apply the READOUT function which aggregates node features from the final iteration and sum them up together to get the vector h_G:

$$h_G = \text{READOUT}\left(\left\{h_v^{(K)} | v \in G\right\}\right) \tag{3a}$$

This graph representation h_G represents the whole graph in order to learn an end-to-end mapping. The READOUT function performs as a fully connected output layer, where its output is a single depression score.

The representations of the nodes and weights of edges are shared globally and can be updated in the text level graphs through a MPM approach.

Our GNN demonstrates a strong representation learning capability to learn more flexible and strong context-centered semantic features. Because of the way in which the model learns the mapping of high-dimensional feature vectors (e.g., represent cognitive semantic features, details are in the section 2.1) for predicting depression levels, we could improve the performance of the challenge task of depression prediction.

Data: Both the 6_{th} and 7_{th} International Audio/Video Emotion Challenge (AVEC)[8] provided the same data containing video-based facial actions, audio and the conversation transcribed to text for each participant and the corresponding. The depressed state of subjects is based on the PHQ-8 metric [17] ranging from 0 to 24. We utilized the text transcriptions of 142 individuals and within the dataset, 43 out of 142 subjects (30%) were labeled as depressed.

The provided dataset has been split into a training set having 107 patients and a development set containing 35 patients. In line with prior work [3,8,22,24] and to ensure comparable results, we test on the "development set" from the original competitions [8], since the actual test set was not in the public domain.

Privacy: This data[1] does not contain protected health information(PHI). Personal names, specific dates and locations were removed from the audio recording and transcription by the dataset curators.

3 Experiments

Our experiments consist of two parts. First, we predict the PHQ score for each participant. We then compare our method to other existing works in terms of measuring the severity of depression symptoms (Table 1). Second, we compare our GNN with other baseline models to show the effect of using a graph-level embedding captured from just text modality. Our approach is superior to state-of-the-art methods in both Table 1 and Table 2.

Table 1: Comparison of deep learning(DL) based approaches for measuring depression symptoms severity on development set using mean absolute error (MAE). The task is evaluated : PHQ score regression. Modalities: A: audio, V: visual, L: linguistic(text), A+V+L: combination.

Regression: PHQ score DL Methods	Features	PHQ score
Sentence-based DL model	context-free	MAE
Baseline Challenge [89]	A + V	5.52
Haque et al. [49]	A + V + L	5.18
Alhanai et al. [24]	A + L	5.1
Alhanai et al. [24]	L	5.2
Haque et al. [49]	V	5.01
Song et al. [33]	V	5.15
Du et al. [10]	V	4.65
Graph-based DL model	context-aware	MAE
Our Approach	L	**4.2**

In Table 1, we compare our method to prior work on measuring depressive symptom severity under the same condition of utilizing deep learning algorithms. There are two major differences between our method and prior work:

1. Our method concentrates on learning context-aware semantic features. We convert the text to the graph level to achieve the goal of learning a mapping of high dimensional probability distributions of correlations between both adjacent words and between non-adjacent words through an entire transcript. On the other hand, prior work performs a context-free modeling to capture sensor-based features from audio, visual, and (or) text. Their feature fusion models, in general, learn a mapping of sentence-level embedding, which relies on capturing content-based semantic features across an interview.
2. We introduce a novel way of using graph representation mapping the whole interview to a high-level feature vector with the purpose of predicting depression states. On the other hand, prior work can just, at the best extent, learn a multi-modal sentence-level embedding. For example, both the work of Alhanai et al.[24] and the work of Haque et al.[49] apply a sequence-based model of LSTM[24]. Du's work[10] implements a CNN[4] model for learning a sentence-level embedding.

3.1 Baselines

Since the proposed method and some other state-of-work methods have a lot of differences in terms of multiple feature set combinations (e.g., visual+audio, audio+text, etc.). It is hard to check the effectiveness and advantage of using

[1] https://dcapswoz.ict.usc.edu/

a graph-based deep learning model to perform the task of depression prediction. Therefore, we only use text modality to test the performance of different models. We compare our method with the following baseline models for measuring depressive symptom severity (in Table 2). The difference is learning a target mapping based on either sentence level(context-free) or the graph level(context-aware). We only convert the text to the graph when learning a graph-level embedding.

Row 1 uses pre-trained word embedding for learning just semantic content features which is computed via simple average [30].

Row 2 and 3 implement sequence-based modeling, where it uses the last hidden state as the sentence-level representation of the text. Row 4 uses dilated convolution [4] and max pooling operating on word embeddings to get the sentence-level representation of text. These two models all used 300-dimensional GloVe word embeddings.

Row 5 shows our work of using graph neural network to learn a representation of an entire text graph. In comparison with state of the art results on the same dataset, our approach based on a GNN achieves the best performance.

Table 2: Baseline comparisons. Row 1 is a pre-trained embedding. Row 2-4 are learned a sentence-level embedding. Row 5 is our learned a graph-level emebdding

Regression: PHQ score		
# Model	Feature	MAE
1 D2V [29]	L	5.81
2 LSTM [24]	L	5.13
3 GRU [24]	L	5.49
4 T-CNN [4]	L	6.177
5 **GNN**	L	**4.2**

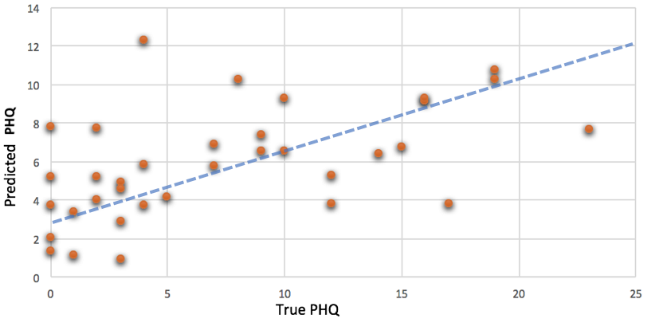


Figure 3: Results on development set for our graph-level PHQ prediction system, with predicted PHQ plotted as a function of true PHQ.

3.2 Implementation details

We set the size of the word node embedding as 200 and initialized with GloVe [19]. p (discussed in section 2.2) is set to 8. We set the learning rate as 10^{-3}, $L2$ weight decay as 10^{-4}, and dropout rate as 0.5. The batch size of our model is 32. The loss objective is mean absolute error (MAE) for regression. We trained GNN for a maximum of 500 epochs using the ADAM optimizer [17] and stopped training if the validation loss does not decrease for 10 consecutive epochs.

Qualitative Analysis: Some incorrect results predicted by our model may be caused by the unequal distribution of dataset. In figure 3, the model made more biased predictions on those patients who are diagnosed with the most severe depressive symptoms. For example, the model predicts a depression score of around 4 to the patient whose ground truth score is 17, and the patient who has the depression score of 23 has been wrongly predicted with a score of around 8. Mitigating biased prediction may require more subject samples particularly having higher depression scores (i.e., a PHQ score$\geq$10) in order to help the model learn more general features of depression related cognitive biases. Thus it could improve the generalizability of the model to perform consistently better results on predicting depression scores for observations who have severe depressive symptoms.

4 Conclusions

Our work shows a new and potential way of improving the performance of quantifying depression levels by learning a graph representation from text. Our novel solution efficiently integrates some heuristics findings from psychology with making use of observations from the provided data. Hopefully, our work may shed light on integrating the intelligence from another domain to build an applicable structure of using that domain knowledge to design a task-oriented deep learning architecture. Moreover, applying deep learning on graphs in the application area of mental health diagnosis might facilitate automatic screening.

Our work may help other researchers who are standing at the middle of “machine learning” and “other challenging areas” open their minds contributing to developing and applying deep learning tools to their professional arenas, especially they may face the similar challenge as we do, for instance, their data is not cheap in the real world application. Our work would, in some extent, weaken their concerns of taking a risk of promoting this new and growing technology in their communities.

References

[1] Aaron T. Beck, Robert A. Steer & Margery G. Carbin. 1988. Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review***8**(1), pages 77–100.

[2] Aaron T. Beck. 2002. Cognitive models of depression. *Clinical advances in cognitive psychotherapy: Theory and application, 14(1),* pages 29-61.

[3] Beckham Ernest Edward, William R. Leber, John T. Watkins, Jenny L. Boyer & Jacque B. Cook. 1986. Development of an instrument to measure Beck's cognitive triad: The Cognitive Triad Inventory. *Journal of consulting and clinical psychology,*54 (4):566.

[4] Bai Shaojie, Kolter J. Zico & Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling.*arXiv preprint arXiv:1803.01271*

[5] Cummins Nicholas, Vidhyasaharan Sethu, Julien Epps, James R. Williamson, Thomas F. Quatieri & Jarek Krajewski. 2017. Generalized Two-Stage Rank Regression Framework for Depression Score Prediction from Speech. *IEEE Transactions on Affective Computing.*

[6] Cho Kyunghyun, Van Merriënboer Bart, Gulcehre Caglar, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk & Bengio Yoshua. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation.*arXiv preprint arXiv:1406.1078.*

[7] Gilmer Justin, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals & George E. Dahl. 2017. Neural message passing for quantum chemistry. *In Proceedings of the* 34_{th} *International Conference on Machine Learning-Volume 70,* pages 1263-1272. JMLR.org.

[8] Valstar Michel, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie & Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition. *In Proceedings of the* 6_{th} *International Workshop on Audio/Visual Emotion Challenge,* pages 3-10

[9] Hamilton Will, Zhitao Ying & Jure Leskovec. 2017. Inductive representation learning on large graphs. *In Advances in Neural Information Processing Systems,* pages 1024- 1034.

[10] Zhengyin Du, Weixin Li, Di Huang & Yunhong Wang. 2019. Encoding visual behaviors with attentive temporal convolution for depression prediction. 14_{th} *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1-7. IEEE.

[11] Holtzman S. Nicholas. 2017. A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality,* 68, pages 63-68.

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

[13] Jayawardena Sadari, Julien Epps & Eliathamby Ambikairajah. 2019. Evaluation Measures for Depression Prediction and Affective Computing. *In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,pages 6610-6614.IEEE.

[14] Keyulu Xu, Weihua Hu, Jure Leskovec & Stefanie Jegelka 2018. How powerful are graph neural networks?. *arXiv preprint arXiv:1810.00826.*

[15] Kipf N.Thomas & Max Welling. 2016. Variational graph auto-encoders. *In NIPS Workshop on Bayesian Deep Learning.*

[16] Kroenke Kurt, Tara W. Strine, Robert L. Spitzer, Janet BW Williams, Joyce T. Berry & Ali H. Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. *Journal of affective disorders,* 114(1-3), pages163-173.

[17] Kingma Diederik P. & Jimmy Ba. 2014. Adam: A method for stochastic optimization.*arXiv preprint arXiv:1412.6980.*

[18] Michael T. Moore & David M. Fresco. 2007. Depressive realism and attributional style: Implications for individuals at risk for depression. *Behavior Therapy,* 38(2), pages 144-154.

[19] Mikolov Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *In Advances in neural information processing systems,* pages.3111-3119.

[20] Mosaiwi A. Mohammed. 2018. People with Depression use Language Differently–Here's How to Spot it. *The Conversation vom,2.*

[21] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li & Razvan Pascanu. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261.*

[22] Shubham Dham, Anirudh Sharma & Abhinav Dhall. 2017. Depression scale recognition from audio, visual and text analysis. *arXiv preprint arXiv:1709.05865.*

[23] Soygüt Gonca & Işik Savaşir. 2001. The relationship between interpersonal schemas and depressive symptomatology. *Journal of Counseling Psychology,* 48(3), pages 359-364.

[24] Tuka Alhanai, Mohammad Ghassemi & James Glass. 2018. Detecting Depression with Audio/Text Sequence Modeling of Interviews. *In Interspeech vol. 2522,* pages 1716-1720.

[25] Tsakalidis Adam, Tsakalidis A, Liakata M, Damoulas T & Cristea AI. 2018 Can we assess mental health through social media and smart devices? Addressing bias in methodology and evaluation.*Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Pages 407-423.

[26] World Health Organization. 2017. Depression and other common mental disorders: global health estimates. *No. WHO/MSD/MER/2017.2.* World Health Organization.

[27] Yuan Gong & Christian Poellabauer. 2017. Topic modeling based multi-modal depression detection. *In Proceedings of the* 7_{th} *Annual Workshop on Audio/Visual Emotion Challenge,* pages 69-76.

[28] Zimmermann Johannes, Timo Brockmeyer, Matthias Hunn, Henning Schauenburg & Markus Wolf. 2017. First-person pronoun use in spoken language as a predictor of future depressive symptoms: Preliminary evidence from a clinical sample of depressed patients.*Clinical psychology & psychotherapy 24, no. 2,* pages 384-391.

[29] Le Quoc & Tomas Mikolov. 2014. Distributed representations of sentences and documents. *In International conference on machine learning* pages 1188-1196.

[30] Cer Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John & Noah Constant. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175.*

[31] Haque Albert, Michelle Guo, Adam S. Miner & Fei-Fei Li. 2018. Measuring depression symptom severity from spoken language and 3D facial expressions. *arXiv preprint arXiv:1811.08592.*

[32] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee & Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[33] Song Siyang, Linlin Shen, & Michel Valstar. 2018. Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. *In Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 158–165.