USING GRAPH REPRESENTATION LEARNING WITH SCHEMA ENCODERS TO MEASURE THE SEVERITY OF DEPRESSIVE SYMPTOMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph neural networks (GNNs) are widely used in regression and classification problems applied to text, in areas such as sentiment analysis and medical decision-making processes. We propose a novel form for node attributes within a GNN based model that capture node-specific embeddings for every word in the vocabulary. This provides a global representation at each node, coupled with node-level updates according to associations among words in a transcript. We demonstrate the efficacy of the approach by augmenting the accuracy of measuring major depressive disorder (MDD). Prior research has sought to make a diagnostic prediction of depression levels from patient data using several modalities, including audio, video, and text. On the DAIC-WOZ benchmark, our method outperforms state-of-art methods by a substantial margin, including those using multiple modalities.

1 INTRODUCTION

Nowadays deep learning techniques have been frequently used in sentiment analysis (Tan et al., 2013; Mukhtar & Khan, 2018; Liao et al., 2021), particularly in movie recommendations and product reviews and ratings. In the healthcare domain, sentiment analysis with deep learning provides many benefits such as utilizing medical information to achieve the best result to increase healthcare quality (Abirami & Askarunisa, 2017; Bi et al., 2020). We focus our work on applications related to mental health such as using deep learning to make a diagnostic prediction.

There is a pressing need to find a convenient and automated method to assess depression severity. Worldwide, more than 300 million people are suffering depression (Organization et al., 2017). Apart from the high prevalence of Major Depressive Disorder, overall 85% of depressed individuals are underdiagnosed (Falagas et al., 2007). Research has shown that about 30% of patients suffering from an episode of major depression do not seek treatment, with only 10% of them being adequately treated (Falagas et al., 2007). We validate our approach based on its application to the diagnosis of MDD. We demonstrate that our approach can result in more accurate predictions of the severity of depression, which can enhance automated health care decision making.

Medical research (Tsakalidis et al., 2018) shows that text sentiment analysis method can be effective in making inferences about people's mental states, such as MDD. These may not be reflected in a person's visual appearance, i.e., visual behaviors. For example, depressed people use more first person pronouns than non-depressed people because they are more focused on themselves and less connected with others. Visual data alone will therefore be insufficient for capturing depressive symptoms that are only manifested in a person's verbal utterances.

Mental disorders are difficult to detect. In practice, clinicians first measure the severity of depressive symptoms in patients and then identify depression in them. During in-person clinical interviews, clinicians widely use a structured Patient Health Questionnaire (PHQ) — a clinically validated tool determining the severity of depressive symptoms across several personal dimensions (Kroenke & Spitzer, 2002).

Inspired by the observation of the provided data set, we formulated the hypothesis that the context of words in a transcript can be used to generate PHQ scores (Kroenke et al., 2009). This PHQ metric for

depressive disorder forms a context constructed with eight items, such as “sleep problem”, “anxiety problem”, “fatigue problem”, “depression problem”, and ”no motivation or interest in things”. We observed that each text covers information relating to at least one of these eight topics (see figure 1). This motivates use of a predictive model that represents these eight PHQ topics, thereby introducing an inductive bias into the learning process. Thus the contextual information about these eight PHQ topics should be encoded to determine depression states.

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: An extract from a raw transcript.

We motivate generating graph-level representations for transcripts (as inputs) to encode contextual information from these transcripts (we assume a transcript encodes facts representing depressive symptoms). To achieve this, we propose a novel form for node attributes within a GNN based model that capture node-specific embeddings for every word in the vocabulary. The representations of the same words are shared globally and can be updated according to associations among words in a transcript. We summarize the representations of all the words in the transcript to predict depression states.

Intuitively, the generated node-level embeddings maintain records linking certain existing facts that indicate the known symptoms of a subject. When subsequent facts are discovered which indicate additional depressive symptoms, records will be updated by aggregating both old and new facts. In other words, those records will be updated via a message passing mechanism over the transcript context until ideally all major depressive features (characterizing depressive symptoms) have been discovered.

In this way, we can represent the most relevant contextual information which is universal across all transcripts – highly possible involving the context of eight PHQ topics we discussed above. Using graph structures to capture context-level features is an innovative pathway which we hypothesise can be used to measure different levels of depression. In our experiments, our approach realized a good generalization on a limited, incomplete, and unevenly distributed dataset (in figure 3).

1.1 Related Work

Research in measuring the severity of depressive symptoms aims to train a regression model to predict depression scores (Valstar et al., 2016; Alhanai et al., 2018; Tsakalidis et al., 2018). Some prior work (Ringeval et al., 2017; Alhanai et al., 2018; Haque et al., 2018) applies a sequence-level deep learning model to capture implicit depressive signals. Their model, in general, uses a multi-modal sentence embedding, rather than a mapping of a whole interview, to predict a PHQ score. Other prior work (Valstar et al., 2016; Cummins et al., 2017) implements statistical functions (e.g., max, min) on short-term features over an entire interview, but this may fail to preserve useful temporal information across an entire interview, such as some short-term signs of regret, anxiety, etc.

Current machine learning methods (Valstar et al., 2016; Alhanai et al., 2018; Song et al., 2018; Du et al., 2019) concentrate on learning features from multi-modalities — as determinants for estimating the severity of depressive disorders. In principle, we should be able to improve performance further using all available modalities, however feature-fusion method may lead to learn a wrong representation encoding inter-subject variability unrelated to depression (Williamson et al., 2016).

The feature-fusion approach to learning from small data may cause poor generalization: with a small sample size, the number of features should also be small to avoid the problems of dimensionality and overfitting. However, given that the dimensions of audio and video features are very large, sparse parametrization needs to be taken into account when training machine learning models for learning a joint multi-modal feature vector (Williamson et al., 2016).

2 METHOD

We generate a node-level embedding matrix for each word of a transcript. These node attributes represent an underlying relationship between the current word and every other word in the transcript. In this way, a graph neural network (GNN) schematizes underlying associations among words via their internal representations. We propose to aggregate their internal representations to encode information from the context of a transcript generalizing all major depressive symptoms[1].

Building a Text Graph: For a given transcript, we build a text graph. We regard all unique words appearing in a transcript as the nodes of the graph. Each edge starts from a word in the transcript and ends with an adjacent word.

Let $G = (V, E)$ denote a text level graph. V is a set of nodes representing all the unique words in a given text, and E is a set of undirected edges between pairs of these nodes, each represented by a set of the two nodes at either end of the edge. Each node has an attribute which is a matrix. To emphasise this representational structure of the attribute matrix, we refer to it as a 'schema' $U_i \in \mathbb{R}^{n \times d}$. The j_{th} row of U_i is a vector of length d containing the representation that node v_i has of v_j; and n denotes the total number of unique words (the vocabulary size) in a corpus.

2.1 SCHEMAS

We generate a schema for each word node which performs a role of recording a global context. This global context retains information from interactions between the current word and every other word. In this way, each word node maintains "a dynamic record" (in the form of a schema) of the context from the given transcript. The schemas are progressively updated by a GNN model. This resulting model produces final embeddings of the words in a transcript in relation to all words in the vocabulary, including the other words that make up the transcript.

Schemas preserve structures that represent relationships between the identifier word and every other word. Thus we can exploit these schemas to capture context-level features in an explicit way by learning a GNN model. Our innovation is to represent word proximity through the graph structure and co-occurrence within the same transcript within the schema at each node.

2.2 SCHEMA-BASED GRAPH NEURAL NETWORK (SGNN)

We use multiple passes (layers) of the message passing mechanism (MPM) (Gilmer et al., 2017; Xu et al., 2018) to update the schema at each node of the text graph.

Initialization: We first initialise an $n \times d_1$ matrix (in our experiments, we set $d_1 = 32$) as the schema $U_i^{(1)}$ at each node $v_i \in V$. The schema is all zeros apart from the row corresponding to the word associated with this node, which is a random d-dimensional vector using a linear transformation.

Message passing layer: In this work, the operation of message passing is split into two steps to update the schema at each node.

We first modify the schema at each node v_i:

$$\hat{U_i}^{(k)} = U_i^{(k)} W_1^{(k)} + \frac{1}{n} \mathbf{1}\mathbf{1}^T U_i^{(k)} W_2^{(k)} + \frac{1}{n} \mathbf{1}_i \mathbf{1}^T U_i^{(k)} W_3^{(k)} \tag{1}$$

[1] our code links to: `https://github.com/Clio-dl/Using-SGNN-for-Depression-Estimate.git/`

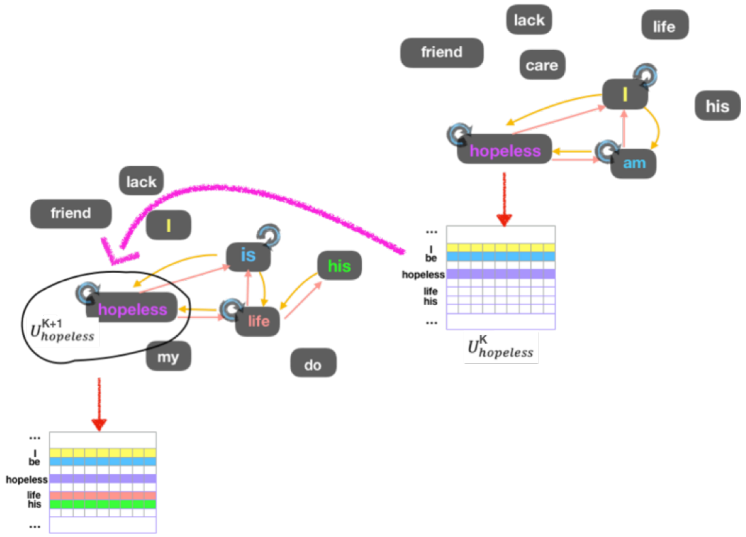

Figure 2: For one of layers of SGNN model k, the upper figure showing an output is a schema generated for the word node "hopeless". This schema can be treated as an "inner record" of the node "hopeless", which is formed as $U^k_{hopeless} \in n \times d_k$, where the generated d_k-dimensional representation of a node presents as a row. The blank rows correspond to word nodes that have not been encountered with this identifier node "hopeless" in the current text context. The figure on the bottom shows an example of a modified schema of the same word node "hopeless" after learning. Our model using schemas can encode information about associations between the node "hopeless" and its new neighbors, such as a node "his" (colored in green). As a result, the internal representation of this identifier node "hopeless" is updated by the model. Moreover, its original record containing its other already existing neighbors has been explicitly preserved while learning. ***Note:*** The words "am", "are" and "is" are all stemmed to the same word, "be", and are thus represented by a single node.

where $\mathbf{1} \in \mathbb{R}^n$ is a vector of ones.

The first term updates each row independently. The second term operates on the sum of the columns, replicated in each row. The third term operates on the row corresponding to the current word, replicated in each row. These terms are a subset of equivariant linear functions which are computed by (Maron et al., 2018).

Second we compute the message function, which is defined as:

$$M^{(k)}\left(\hat{U_i}^{(k)}, \hat{U_j}^{(k)}\right) = \hat{U_j}^{(k)} + RELU\left(\left[\hat{U_i}^{(k)} || \hat{U_j}^{(k)}\right] W_4^{(k)}\right) W_5^{(k)} \tag{2}$$

where $||$ denotes concatenation along the second axis. $\left(W_m^{(k)}\right)_{1 \leq m \leq 5}$ are learnable parameters. This is essentially a two-layer neural network.

In the next step, each node's schema is updated as:

$$U_i^{(k+1)} = \sum_{j \in N(i)} M^{(k)}\left(\hat{U_i}^{(k)}, \hat{U_j}^{(k)}\right) \in \mathbb{R}^{n \times d_{k+1}} \tag{3}$$

We apply equation (3) as the sum aggregator over the k-th layer of the SGNN.

After all K message-passing layers have been applied, the schemas are pooled to get representations of all nodes in the graph. We apply an elementwise max-pooling operater to extract node features. We use the READOUT function which aggregates node features by summing them together:

$$h_G = \sum_{v_i \in V} U_i^{(K)} \tag{4}$$

We use h_G to predict a PHQ score for each transcript. In our experiments, we apply a 2-layer multi-layer perceptron (MLP).

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

Regression: PHQ score		
Methods	Modalities	PHQ score MAE
Baseline Challenge (Valstar et al., 2016)	A+V	5.52
Gaussian Staircase Regression (Williamson et al., 2016)	A+V+L	4.18
LSTM (Haque et al., 2018)	A+V+L	5.18
LSTM (Alhanai et al., 2018)	A+L	5.1
Support Vector Regression (Song et al., 2018)	V	5.15
DepArt-Net (Du et al., 2019)	V	4.65
LSTM (Alhanai et al., 2018)	L	5.2
C-CNN (Haque et al., 2018)	L	6.14
GNN (Gilmer et al., 2017)	L	4.24
Our Approach	L	**3.76**

Table 2: Results of ablation studies. We applied 10-fold stratified cross-validation and give mean results. MAE: mean absolute error. RMSE: root mean squared error.

Metric	MAE		RMSE	
Setting	Train	Test	Train	Test
Original(schema-GNN)	3.85	3.52	4.48	4.32
i) Fast(schema-GNN)	4.28	3.86	5.14	4.51
ii) Without equivariant linear layers (Maron et al., 2018)	4.14	3.95	5.60	4.49
iii) Mean Reduction	4.32	5.16	4.45	5.12

2.3 DATA

Both the 6th and 7th International Audio/Video Emotion Challenge (AVEC) (Valstar et al., 2016) used the same DAIC-WOZ dataset containing video-based facial actions, audio and the conversation transcribed to text for each participant. We utilized only the text transcripts of all 142 individuals 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 (Valstar et al., 2016; Alhanai et al., 2018; Cummins et al., 2017; Haque et al., 2018) and to ensure comparable results, we test on the "development set" from the original competitions (Valstar et al., 2016), since the actual test set is not in the public domain.

Privacy: This data[2]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

In this section, we show the results of our method on the DAIC-WOZ benchmark (details can be seen in part of describing the data in section 3.3). We then compare our method to the state-of-art works centering on feature fusion learning algorithms.

In this experiment, we predict the PHQ score for each participant. The loss objective is mean absolute error (MAE) for regression.

We compare our method to prior work on measuring depressive symptom severity. The performance of our method and nine other methods, including the state of the art method, is set out in Table 1. Our method outperforms all other methods, despite using only the textual modality.

We found that our model performs better than the standard GNN (Gilmer et al., 2017). The standard GNN utilizes node representations which are initialized by pre-trained 300-dimensional GloVe word

[2]`https://dcapswoz.ict.usc.edu/` last accessed on 18/09/20

embeddings (Mikolov et al., 2013). However, our method constructs schemas which are initialized with random vectors. Moreover the standard GNN propagates node features in terms of vectors while our model uses 2-D node attributes. This change increases the expressive power of the MPM, resulting in better results.

The machine learning algorithms of (Valstar et al., 2016; Williamson et al., 2016) perform modeling statistics using handcrafted features derived from audio, text and (or) visual inputs. However, our method uses raw text.

We also note that our model performs better than other prior work (Alhanai et al., 2018; Song et al., 2018; Haque et al., 2018; Du et al., 2019). That is likely due to the difference of representation learning. Their work uses a multi-modal sentence-level embedding to predict a PHQ score while our model is trained to learn a graph-level embedding of each transcript.

3.1 Ablation Study

To further analyze our model, we perform three ablation studies (see Table 2). To give more reliable measures of performance, we perform 10-fold cross validation on the combined training and development dataset. We concatenate the training and development set as one set and then divide it into 10 folds in a stratified manner. Each time one fold is used for testing and the other 9 folds are used for training.

In (i), we give equations 1-4 (in section 2.2) for 'point-wise' multiplication. According to the results in Table 2, we can see that using convolution layers can better model the relations between words compared with point-wise multiplication layers.

In (ii), we remove the second and third (equivariant) terms in (1). There was a significant reduction in performance (Table 2). As a result, the representation power of our model can increase when parametrized with these layers, thereby improving the model's performance on the data set.

In (iii), we replace the max-pooling operator by mean-pooling operator. From Table 2, we can see that the result was not good when applying mean operator of the node. Max operator highlighting features can enhance discriminating depressive features, which helps to achieve a better result.

3.2 Analysis of 2-dimension node attributes

We compared the result of performances between a standard GNN model (plotted in red) and our SGNN model (plotted in blue) using 2-dimension (2-D) node attributes. Figure 4 shows that the latter generalized the development set much better, particularly for the small group of samples in the class of having scores higher than 13. Our graph representation learning framework using schemas achieves better generalisation on the very limited dataset.

We suggest that using 2-D node attributes (i.e. schemas) can improve the expressive power of the MPM. By setting a global context (matrices) of each node (as the input), we can obtain some parameters, in an explicit way, that capture the context of a word using each node's schema. On the contrary, using simple 1D node embedding such as GloVe may result in losing contextual information when they are processed by several layers of a learning model.

Settings: We set the learning rate as 10^{-3}, $L2$ weight decay as 10^{-4}, and dropout rate as 0.5. The loss objective is mean absolute error (MAE) for regression. We trained the SGNN for a maximum of 500 epochs using the SGD optimizer (Kingma & Ba, 2014) and stopped training if the validation loss does not decrease for 10 consecutive epochs.

Qualitative Analysis: Our model has some incorrect predictions which may be caused by the unequal distribution of dataset (see figure 3). In figure 4, our SGNN model made more biased predictions on those subjects with higher PHQ depression scores (plotted in blue): For instance, the model predicted a PHQ score of around 16 to the patient with an actual PHQ score of 13, and the patient who has the PHQ score of 23 has been overpredicted with a score of around 26. To improve the learning capacity of the model, more labeled training samples with PHQ scores higher than 8 would be needed.

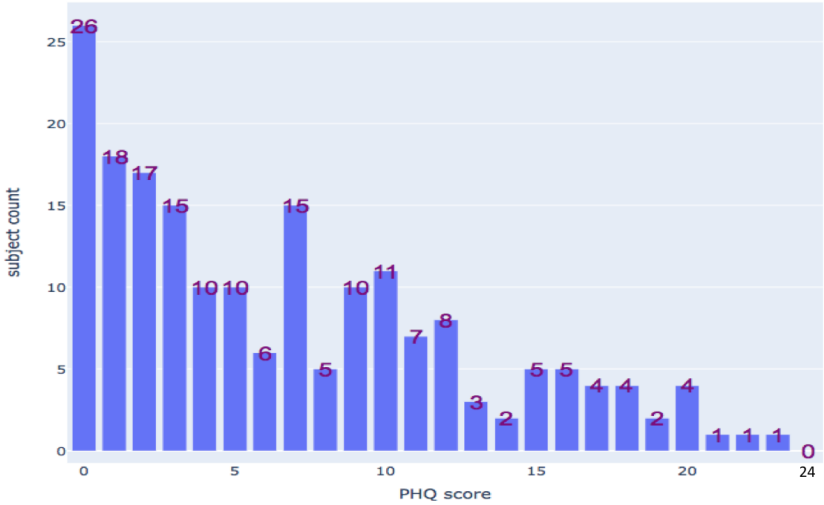

Figure 3: A histogram showing the distribution of PHQ scores on the dataset (of 189 subjects). The X-axis represents a PHQ score ranging from 0 to 24. The Y-axis represents the number of subjects for each score.

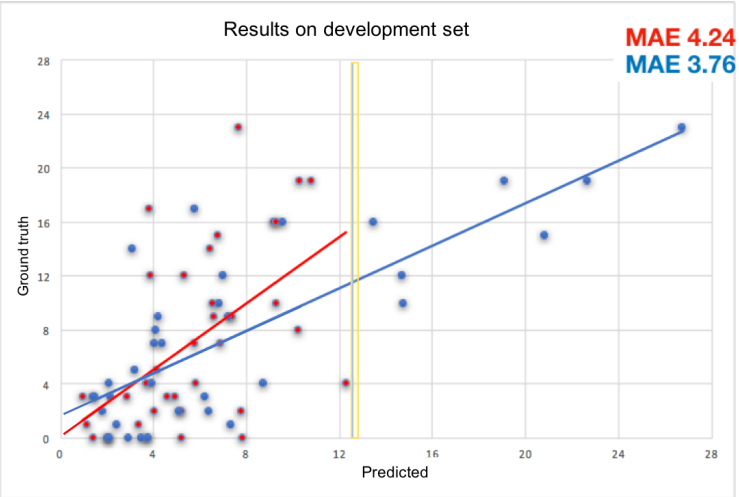

Figure 4: Results on the development set for our graph-based PHQ prediction system, with the true PHQ plotted as a function of predicted PHQ. The red color represents the performance of applying a generic GNN (Gilmer et al., 2017) algorithm, whereas the blue color represents the performance of implementing our SGNN. The yellow vertical line, as a cut-off, emphasizes a better generalization performance of a small group having scores higher than 13 (on the right of the figure).

4 Conclusions

Our work shows an innovative way of improving the performance of predicting depression states by training a deep graph learning model to learn contextual features from the text. Our results have demonstrated that it is possible to apply deep learning methods to successfully solve specific problems within the field, and not just a more general problem, even with limited data. Future work might address finding a way of explaining a GNN algorithm by demonstrating the relationship between the underlying depression features and depression scores, since understanding why the model made certain predictions could also be valuable in a clinical context.

REFERENCES

Ariyur Mahadevan Abirami and Abdulkhader Askarunisa. Sentiment analysis model to emphasize the impact of online reviews in healthcare industry. *Online Information Review*, 2017.

Tuka Alhanai, Mohammad Ghassemi, and James Glass. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pp. 1716–1720, 2018.

Qiqing Bi, Lining Shen, Richard Evans, Zhiguo Zhang, Shimin Wang, Wei Dai, Cui Liu, et al. Determining the topic evolution and sentiment polarity for albinism in a chinese online health community: machine learning and social network analysis. *JMIR medical informatics*, 8(5): e17813, 2020.

Nicholas Cummins, Vidhyasaharan Sethu, Julien Epps, James R Williamson, Thomas F Quatieri, and Jarek Krajewski. Generalized two-stage rank regression framework for depression score prediction from speech. *IEEE Transactions on Affective Computing*, 11(2):272–283, 2017.

Zhengyin Du, Weixin Li, Di Huang, and Yunhong Wang. Encoding visual behaviors with attentive temporal convolution for depression prediction. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–7. IEEE, 2019.

ME Falagas, KZ Vardakas, and PI Vergidis. Under-diagnosis of common chronic diseases: prevalence and impact on human health. *International journal of clinical practice*, 61(9):1569–1579, 2007.

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

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kurt Kroenke and Robert L Spitzer. The PHQ-9: a new depression diagnostic and severity measure, 2002.

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

Wenxiong Liao, Bi Zeng, Jianqi Liu, Pengfei Wei, Xiaochun Cheng, and Weiwen Zhang. Multi-level graph neural network for text sentiment analysis. *Computers & Electrical Engineering*, 92: 107096, 2021.

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. *arXiv preprint arXiv:1812.09902*, 2018.

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

Neelam Mukhtar and Mohammad Abid Khan. Urdu sentiment analysis using supervised machine learning approach. *International Journal of Pattern Recognition and Artificial Intelligence*, 32 (02):1851001, 2018.

World Health Organization et al. Depression and other common mental disorders: global health estimates. Technical report, World Health Organization, 2017.

Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pp. 3–9, 2017.

Siyang Song, Linlin Shen, and Michel Valstar. Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 158–165. IEEE, 2018.

Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Jiajun Bu, Chun Chen, and Xiaofei He. Interpreting the public sentiment variations on twitter. *IEEE transactions on knowledge and data engineering*, 26(5):1158–1170, 2013.

Adam Tsakalidis, Maria Liakata, Theo Damoulas, and Alexandra I Cristea. Can we assess mental health through social media and smart devices? addressing bias in methodology and evaluation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 407–423. Springer, 2018.

Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pp. 3–10, 2016.

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

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2018.