UNIVERSITY OF LEEDS

Measuring the Severity of Depression from Text using Graph Representation Learning

Simin Hong

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

Faculty of Engineering

School of Computing

February 2023

Abstract

The common practice of psychology in measuring the severity of a patient's depressive symptoms is based on an interactive conversation between a clinician and the patient. In this dissertation, we focus on predicting a score representing the severity of depression from such a text. We first present a generic graph neural network (GNN) to automatically rate severity using patient transcripts. We also test a few sequence-based deep models in the same task. We then propose a novel form for node attributes within a GNN-based model that captures node-specific embedding for every word in the vocabulary. This provides a global representation of each node, coupled with node-level updates according to associations between words in a transcript. Furthermore, we evaluate the performance of our GNN-based model on a Twitter sentiment dataset to classify three different sentiments and on Alzheimer's data to differentiate Alzheimer's disease from healthy individuals respectively. In addition to applying the GNN model to learn a prediction model from the text, we provide post-hoc explanations of the model's decisions for all three tasks using the model's gradients.

Intellectual Property

The candidate confirms that the work submitted is his/her own and that appropriate credit has been given where reference has been made to the work of others.

Signed

Acknowledgements

I am indebted and grateful to my supervisors Prof. Cohn and Prof. Hogg for their thoughtful comments and recommendations on this dissertation.

Prof. Cohn, I would like to thank you for providing me with such a great opportunity to enter the PhD program and teaching me the methodology to carry out the research and to present the research works as clearly as possible.

My sincere gratitude also goes to Prof. Hogg for providing invaluable advice through research, offering insights in terms of machine learning and always encouraging and supporting me.

I gratefully acknowledge my funding from University of Leeds. Thanks to the School of Computing department for providing a friendly environment through my PhD study.

Prof. Aletras and Prof. Sharoff, your attention to detail and thoughtful feedback have been greatly appreciated. I am grateful for the time and effort you invested in reviewing my work and providing valuable insights. Your guidance and support have been instrumental in helping me improve my thesis. Thank you again for your invaluable contributions.

I am extremely grateful to my parents for their endless love, support and encouragement throughout my journey. Also I express my thanks to my brother for his warm support and advice, which inspired me during my study. Their belief in me has kept my spirits and motivation high during this process.

Words cannot express my gratitude to Steve Jobs. His words of wisdom, ”stay hungry, stay foolish,” drive me forward and inspire me to live a life full of curiosity, determination, and courage.

List of Abbreviations

GNN	Graph Neural Network
MDD	Major Depressive Disorder
ML	Machine Learning
NLP	Natural Language Processing
MDE	Major Depression Episode
PHQ	Patient Health Questionnaire
AUs	Action Units
FACS	Facial Action Coding System
MFCCs	Mel-frequency Cepstral Coefficients
AD	Alzheimer's Disease
LDA	Latent Dirichlet Allocation
PRIME-MD	Primary Care Evaluation of Mental Disorders
POS	Linguistic Patterns of Part-of-speech
BOW	Bag-of-words
LIWC	Linguistic Inquiry and Word Count
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
1D-CNN	One-dimensional Convolutional Neural Network
LSTM	Long Short-term Memory
GRU	Gated Recurrent Unit
ELMo	Embeddings from Language Models
BERT	Bidirectional Encoder Representations from Transformers
GCNs	Graph Convolutional Neural Networks
Text GCN	Text Graph Convolutional Neural Network
PTSD	Post-traumatic Stress Disorder
BiLSTM	Bidirectional Long Short-Term Memory
t-SNE	T-distributed Stochastic Neighbor Embedding
AVEC	Audio/Video Emotion Challenge

FC	Fully Connected
DL	Deep Learning
T-CNN	Temporal Causal Neural Network
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
MLP	Multi-layer Perceptron
GAT	Graph attention network
MPM	Message Passing Mechanism
SGNN	Schema-based Graph Neural Network
SGD	Stochastic Gradient Descent
CT	Control
PHI	Protected Health Information
BDI-II	Beck Depression Inventory–Second Edition
DAIC	Distress Analysis Interview Corpus
DAIC-WOZ	Distress Analysis Interview Corpus

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The present chapter describes the motivation of the work. The importance of designing Graph Neural Network (GNN) models whose outputs can be understood by human experts is discussed. The general clinical measurement of depression is introduced. Depression indicators which are widely used in automated methods for assessing depression severity are summarized, including both verbal signs and non-verbal signs of depression. In particular, symptoms of depression that can be identified from the verbal modality are discussed. A brief summary of appropriate depression datasets that researchers select for their depression detection models.

1.1 Research Motivation

According to a 2021 World Health Organization fact sheet, Major Depressive Disorder (MDD) affects 264 million people globally (Islam et al., 2021; World Health Organization, 2017). It is likely that the COVID-19 pandemic has raised heightened concerns about mental well-being, especially with respect to MDD (Bakioğlu et al., 2021). In addition to the high prevalence of major depressive disorder, there is a high number of undiagnosed depressive episodes. Overall, 85% of people suffering from depression are underdiagnosed (Falagas et al., 2007). Relevant research also showed that about 30% of patients suffering from an episode of major depression do not seek treatment, with only 10% of the 30% being adequately treated. There is a pressing need to find a convenient and automated method to assess depression severity. The motivation to pursue the development of a methodology could enhance accessibility to mental healthcare by overcoming traditional barriers. Current technological means can provide the infrastructure for monitoring psycho-emotional state in high-risk individuals as part of early detection. Given the complexity of diagnosis, medical personnel should quicken intervention offering people help promptly, particularly for those who are unaware of which depressed state they are experiencing.

In general, research which aims to make a diagnostic prediction of the severity of depression from patient data uses machine learning regression methods. Most research implements multiple modalities including visual, audio and linguistic features relating to clinical symptoms of depression for MDD detection (Al Hanai et al., 2018; Dham et al., 2017). Such automated joint feature analyses indeed have improved diagnostic accuracy. However, in most clinical situations, we lack access to audio-video data, where Machine Learning (ML) algorithms designed to learn clinical transcripts are relevant and important for developing automatic diagnosis system. In addition, Williamson et al. (2016) argues that some clinical texts such as dialogue transcriptions provide the most informative and effective indicator for predicting depression compared to any other source (audio or video). Some studies using Natural Language Processing (NLP) with ML models have been very successful in the mental health application (**pennebaker2015development**; Lin et al., 2020; Morales & Levitan, 2016; Williamson et al., 2016).

With the rise of social media, online blog posts and sites such as Twitter, Reddit and Facebook provide an interesting domain to investigate depression (De Choudhury et al., 2013a; Nguyen et al., 2014; Tsugawa et al., 2015). Not only does the NLP technique extend the performance of automatic diagnosis approaches for depression to non-clinical settings, but can also be very effective to generalize to another field such as analyzing text sentiment or to assess different kinds of mental illness.

Implementing diagnostic tools to predict who may suffer from depression requires very little human involvement of physicians. Those patients who are predicted to have depression could potentially be referred straight to mental health professionals in their area or who accept their health care coverage. However, current approaches which assess depression as a binary problem have limitations since they can only make a coarse assessment of psychological state, rather than a more fine-grained and nuanced one. The evaluations of the work that models depression as a numerical rating scale rather than a binary prediction are still rare in the literature.

It is much more challenging to examine depression in a more fine-grained way than a binary diagnostic threshold such as labeling an individual as depressed or not. Published methods reviewed in this research tend to use a binary categorization: people with symptoms of depression that did not meet the diagnostic threshold are labeled as non-depressed. The diagnostic categorization enables information to be processed rapidly (Andrews et al., 2007; Lewinsohn et al., 2000). However, this mental health categorization blurs the intricacies of a mental illness phenomenon and makes a diagnosis less reliable. I aim to develop an innovative way of modeling depression precisely by deep learning on graphs. I leverage inductive bias in deep models to assess the severity of depressive symptoms. Some researchers (van Borkulo et al., 2015; Wichers et al., 2016) suggest that depression can be treated as a latent disease pattern strengthening the symptom chain, and various states of depression can therefore be measured on the basis of a network of symptoms of MDD. I learn to capture specific symptoms of depression and then I predict depression states at different levels based on the detected symptoms. This approach to detecting fine-grained depression may help us improve the validity of the mental illness experience and the reliability of the diagnosis, and provide clinical significance suggesting clues for treatment.

A graph deep learning model may be a promising way of learning high-dimensional semantic features that implicitly convey the clinical significance of depression. I design a GNN model that takes into account high-level patterns in the language, where I propose to learn these patterns to capture a specific schema or a structure of depression.

Since most work on predicting depression levels using deep models is developed without any explanation of their outcomes, deep learning algorithms are treated as black-boxes. Without reasoning about the mechanisms behind the predictions, clinical experts cannot understand the decisions a deep model makes. My motivation is to develop a way of interpreting a deep model in order to obtain a human understanding of the outcomes of the model. The results of the GNN model can be explained by visualizing a group of latent layers of the model. This research work demonstrates that the model makes a decision based on what it has learned. In addition, I measure the importance of psychological variables in the context of depression using the proposed GNN model, which can help human experts explore the potential values of clinical data. For instance, aligning the model evidence with the empirical evidence found in cognitive psychology.

1.2 The Severity of Depressive Disorder Episode

In clinical settings, the diagnosis of a Major Depression Episode (MDE) requires five or more symptoms to be present within a 2-week period (McDermott & Ebmeier, 2009). The symptoms should include at least a depressed mood or anhedonia (loss of interest or pleasure-L1). The secondary symptoms of MDE are appetite or weight changes, sleep difficulties (i.e., insomnia or hypersomnia), psychomotor agitation or retardation, fatigue or loss of energy, diminished ability to think or concentrate, feelings of worthless or excessive guilt, and suicidality. Previous investigations have reported that cognitive dysfunction, age, unemployment, and suicidal ideation are associated with depressive severity (Johanson & Bejerholm, 2017).

Some research proposes that major depression symptoms are best represented by somatic and non-somatic factors (Van Loo et al., 2012). The somatic items include sleep difficulties, appetite or weight changes, poor concentration, fatigue, and psychomotor agitation or retardation. The non-somatic items involve depressed mood, anhedonia, feelings of worthless and thoughts of death. However, to our knowledge, there is a lack of a systematic study of the relationship between depressive symptoms and depression severity. There is no consensus if the number of symptoms is indicative of depression severity or if the degree of each symptom can be used as an index to classify depression to more specific degrees: moderately depressed, severely depressed and so on. Therefore, the severity of depression is commonly assessed with the aid of medical instruments rating a continuous mental state of a patient based on his/her utterance, such as self-reporting Patient Health Questionnaire (PHQ) which is used to screen for depression.

1.3 Depression Evaluation Instruments

Quantification of severity of depressive symptoms is often aided by rating scales completed by a trained mental health professional. Most studies use labels such as the PHQ, which are calculated based on a clinical Patient Health Questionnaire metric (Kroenke & Spitzer, 2002). The PHQ is a self-administered version of the Primary Care Evaluation of Mental Disorders (PRIME-MD) diagnostic instrument for common mental disorders. Large clinical studies use PHQ as a valid measure of depression severity (Thombs et al., 2014).

Most often diagnostic scales are generated by various versions of Patient Health Questionnaire (PHQ)-2/8/9, comprised of 2, 8 or 9 items respectively (Thombs et al., 2014). In particular, an 8-item version (PHQ-8)(Burnard, 1991) is commonly used as an abbreviated and validated version of PHQ. However, PHQ-9 (Kroenke & Spitzer, 2002) is also widely used, which includes a ninth item related to suicidal ideation. In this thesis, PHQ-8 labels are used and treated as objective truth for assessment purposes. In brief, this PHQ-8 generalizes the symptoms following a regular catalog of eight issues:

- Tiredness and Lethargy
- Depressed Mood
- Trouble Sleeping
- Feelings of Failure or Worthlessness
- Lack of Interest or Ability to Take Pleasure
- Changes in Appetite
- Trouble Concentration
- Psychomotor Impairment

The PHQ-8 metric is defined by summarising these eight items (Kroenke et al., 2009) into a single numeric score. Possible PHQ-8 scores can vary from 0 to 24. Additionally, a total score of 0 to 4 represents no significant depressive symptoms. A total score of 5 to 9 represents mild depressive symptoms; 10 to 14, moderate; 15 to 19, moderately severe; and 20 to 24, severe. A cutpoint PHQ-8 score ≥ 10 representing clinically significant depression was applied to assess subjects who may have a major depressive disorder.

1.3.1 Indicators of Depression

Machine learning tools analyze human dimensions, including facial expression, voice and speech, and language to assess depression. I refer the depression assessment to the process of detecting the presence of depression or evaluating the severity of signs of depression. This section provides a review of each modality highlighting markers including face and gesture, voice and speech, and language and social factors.

1.3.1.1 Non-verbal Indicators

There are two major non-verbal signs of depression that have also been extensively reviewed in the evaluation of depression: 1 visual indicators; 2 acoustic indicators.

Visual indicators including information related to facial, head, body and eye movements could provide key clues with depressive symptoms (Dham et al., 2017). For instance, depressed people tend to avoid eye contact. Joshi et al. (2013) demonstrated that body expressions, gestures, eye and head movements can be significant cues for depression detection. Some research (Cummins et al., 2013; Girard et al., 2014; Joshi et al., 2013; Scherer et al., 2013) has explored relationship that exists between nonverbal behavior and depressive symptoms. Girard et al. (2014) investigated facial features using facial Action Units (AUs) and showed that high-level depressed individuals made fewer affiliative facial expressions and more non-affiliative facial expressions and decreased head motion (i.e., amplitude and velocity). Scherer et al. (2013) employed Facial Action Coding System (FACS) (Ekman & Rosenberg, 1997) and concluded that visual signals detected from nonverbal behaviors can be strong predictors of depression. For instance, their investigations showed that depressed people perform a more downward angle of gaze, less intense smiles, shorter average duration of smile, longer self-touches and fidget with both hands (e.g. rubbing, stroking) and legs (e.g. tapping, shaking).

The properties of acoustic speech can be used as possible cues to detect depression. Some studies investigated that cognitive and physical changes associated with depression can result in differences in speech (Cummins et al., 2015; Dham et al., 2017; Williamson et al., 2016). This idea has driven research into using speech as an important marker for depression. Some research (Cummins et al., 2015; Mundt et al., 2012; Stolar et al., 2015; Trevino et al., 2011; Williamson et al., 2013) investigated a number of feature sets for detecting depression, including monotone pitch, reduced articulation rate, lower speaking volumes and loudness variation from speech. For example, some research revealed that depressed patients tend to speak lower, flatter and softer and can be perceived in a series of prosodic features (of speech) including pitch, loudness and speaking rate (Stolar et al., 2015). Some research (Quatieri et al., 2015; Williamson et al., 2013) applied spectral coordination measures, particularly Mel-frequency Cepstral Coefficients (MFCCs) to estimate the severity of depression. Alghowinem et al. (2013) found that loudness and intensity provide indications of a trend towards spontaneous speech for depressed subjects.

1.3.1.2 Linguistic & Social Indicators

Depression with multiple symptoms can be identified through a person's spoken language (Al-Mosaiwi, 2018). Depression can be considered in multiple dimensions and can be detected from the language, such as depressed mood, depression history, and the amount of cognitive impairment caused by the episode; for example, Durkheim (2005) points out that a person expressing his or her mind not to be integrated into social life has a tendency to suicide – this is also associated with the self-perception of the depressed person. I conclude that depression could be driving the relationship between depressive symptoms and a particular lexicon. Moreover, compared to visual and vocal features, such as those recorded in videos and audio, text-based semantic features are often the most informative indicators obtained by analyzing the patient's textual records (Williamson et al., 2016).

Medical research (De Choudhury et al., 2013b; Poulin et al., 2014; Stirman & Pennebaker, 2001; Tsakalidis et al., 2018) shows that text sentiment analysis methods can be effective in detecting depression. Shen et al. (2017) found that sentiment words with valence, arousal, and dominance, as features, are predictive of depression. These indicators are composed of a class of typical linguistic markers and their occurrences interact with other word entities within the context of a depressed mind.

Topic modeling infers the emotional state of people living with different kinds of cognitive disorders such as depression, Alzheimer's Disease, etc. Topics are designed or selected from a set of effective questions (i.e., a question "have you ever been diagnosed with depression") that can better reveal the conditions of patients (Arseniev-Koehler et al., 2018). Hand-crafted topic features are very informative to identify symptoms from the text data. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) topic models are frequently used to select topic features. However, LDA tools are less adaptive to other mental domain areas and this feature engineering work may not be efficient to scale larger dataset (Lin et al., 2020). Hand-crafted topic features are most informative to identify symptoms from the text data. Moreover, topic modeling can be difficult to apply to assess new patient conditions because patients may refuse to respond to some of the topics from the same feature set.

This research is motivated to learn to represent this type of high-level model available in the language of a depressed person. To achieve this, I propose a novel form for node attributes within a GNN based model that captures a node-specific embedding of every word in the vocabulary.

I learn representations of each word which are shared globally and can be updated according to associations among words in a transcript. I summarize the representations of all the words in the transcript to predict depression states. The research focuses on modeling depression as a continuous phenomenon where facts are aggregated on a transcript record until ideally all possible major depressive features have been discovered.

Some research (Arseniev-Koehler et al., 2018; Morales et al., 2017) also argues that depression should be modeled as a continuous phenomenon rather than simply a binary outcome. I provide a fine-grained prediction of the severity of depressive disorders with a scale between 0 and 24 rather than simply predicting the presence of depression (i.e., depressed or non-depressed). Moreover, this novel approach that contextualizes token embedding with graphs enhances the predictive power of the model in the depression prediction task.

1.4 Thesis Outline

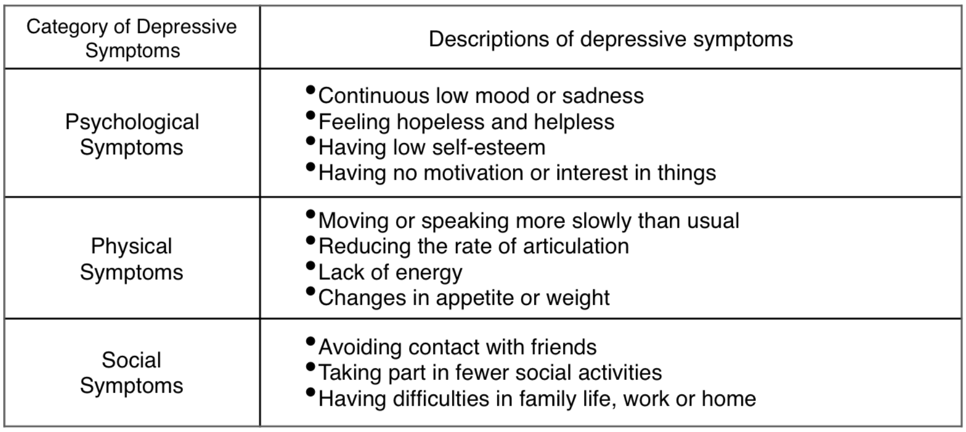
This thesis is organized into seven chapters. Chapter 2 exhaustively reviews the state-of-the-art, presenting relevant machine learning methods employed in the literature. In this chapter, I also discuss and analyze state-of-the-art word embedding feature representations applied in NLP. Besides, I investigate existing post-hoc techniques that explain the predictions of deep models. Chapter 3 introduces a novel paradigm of representation learning methods on graphs, including how to convert text to graphs and how to generate graph-based representations for words. The methodology of generating 2-dimensional word embeddings using graphs developed in this PhD is described in chapter 4 with the main experiment conducted as part of the current study. Chapter 5 presents several post-hoc analyses of explaining outputs generated by the designed deep model. Chapter 6 evaluates the designed deep model on two different text datasets, with a discussion of experimental results and model visualization. Finally, chapter 7 presents the conclusions of this work.

Chapter 2

Literature Review

2.1 Psychological Background of Depression

Depression, in general, is characterized by low mood, a lack of interest, cognitive and psychomotor impairment and suicidal ideation (American Psychiatric Association, DSM-5 Task Force, 2013). For clinical psychologists, depressive symptoms can be summarized into the integration of psychological, physical, and social perspectives (see Figure2.1). These perspectives suggest strong depressive evidence which has in turn provided support for depression diagnosis with validity. Furthermore, a patient who has these symptoms persist for two weeks is considered to have a major depressive disorder (American Psychiatric Association, DSM-5 Task Force, 2013).



Category of Depressive Symptoms	Descriptions of depressive symptoms
Psychological Symptoms	• Continuous low mood or sadness • Feeling hopeless and helpless • Having low self-esteem • Having no motivation or interest in things
Physical Symptoms	• Moving or speaking more slowly than usual • Reducing the rate of articulation • Lack of energy • Changes in appetite or weight
Social Symptoms	• Avoiding contact with friends • Taking part in fewer social activities • Having difficulties in family life, work or home

Figure 2.1: Some examples describing depressive symptoms with three different perspectives.

2.1.1 Linguistic Features in Depression

Psychological research (Al-Mosaiwi & Johnstone, 2018; Schoene & Dethlefs, 2016; Trifan et al., 2020) has shown that people's spoken and written language reflects their mental states. Thus a collection of psycholinguistic evidence, such as pronouns, tense, and lexicon about depression, is widely used to differentiate between depression and non-depression. For example, certain lexical items, including the use of words such as "depressed", "hopeless", and "exhausted" are often used by people who are diagnosed with depression. People with depressive symptoms tend to use markers of linguistic style based on the depression lexicon which is quite different from other people who do not exhibit depressive symptoms.

There are two major types of depression lexicon which can be categorized as follows:

1. Use of first person singular: Some clinical findings report that depressed patients express their thoughts conveying significantly more first person singular pronouns, such as "me" and "I", fewer first person plural pronouns such as "we" and "us", and fewer second and third person pronouns, such as "they", "them" or "she" (Zimmermann et al., 2017). 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. Use of negatively valenced words: There is an existing certain style of language which can be utilized to identify depression. Some research (Morales & Levitan, 2016; Nguyen et al., 2014) has found that depressed people prefer to use more negatively valenced words and fewer positive emotion words, which showed a good predictive validity in depression classification between depressed and control groups from language. For example, depressed patients presumably have more black and white views of the world and this would manifest in their style of language (Holtzman et al., 2017). They have a tendency of using more "absolutist words" (Adam-Troian & Arciszewski, 2020), such as "always", "never", "nothing" or "completely".

2.1.2 Cognitive Biases

These two types of features mentioned above are mainly used to capture differences between depressive and non-depressive patients. The features of cognitive biases are less widely involved and used in the prediction of depression. Some empirical psychological results demonstrate that utterances of depressed people directly and explicitly manifest cognitive biases presenting in their depressed thoughts (Al-Mosaiwi, 2018; Pennebaker et al., 2003). The study of the use of cognitive biases to predict depression symptoms remains less focused.

Cognition is a non-specific term that refers to mental processes associated with thinking, learning and memory. Cognitive bias can be treated as a systematic error in thinking that affects a person's behavior. The Diagnostic and Statistical Manual of Mental Disorders-Fifth Edition (DSM-5) suggests that cognitive impairment is a major indicator of an MDE. Self-reported measures of diminished concentration and attention are frequently observed in individuals presenting with an MDE as part of MDD. Cognitive deficits in MDD are consistent, replicable, non-specific, and clinically significant. The magnitude of cognitive deficits has been demonstrated to be proportionate to the frequency of depressive episodes and duration of illness (Gorwood et al., 2008). For example, individuals with greater depressive symptom severity are more likely to present with cognitive impairments as compared to those with milder illness severity.

Some studies show that assessing the original euthymic subjects, such as recognizing the core belief of a depressed individual, appears to be the best method of investigating the severity of depressive symptoms (Gladstone et al., 2001; Korobkin et al., 1998). According to Beck's cognitive theory (Beck, 1979), depressed people report a negative spin of thoughts involving pessimistic ideas about the self, the world, and the future. Some studies (Parker et al., 1998; Young, 1999a) referring to the hypothesis of "lock and key" assume that there is an existing salient pattern in depression. Such a lock and key hypothesis emphasizes the developmental construction of mistaken beliefs based on the interaction between the self and the environment. Young (1999a) identified a set of early maladaptive schema referring to a stable and constant theme that emerges during childhood. Thus, these schema-driven features, which are implicitly taken for granted by depressed individuals, play a substantial role in priori truths and could be used as a classifier to detect MDD and assess different levels of depression symptoms. Moreover, according to Pyszczynski and Greenberg (1987)'s control theory of depression, depressed individuals think a great deal about themselves, stressing the role of self-focused attention and extreme self-criticism.

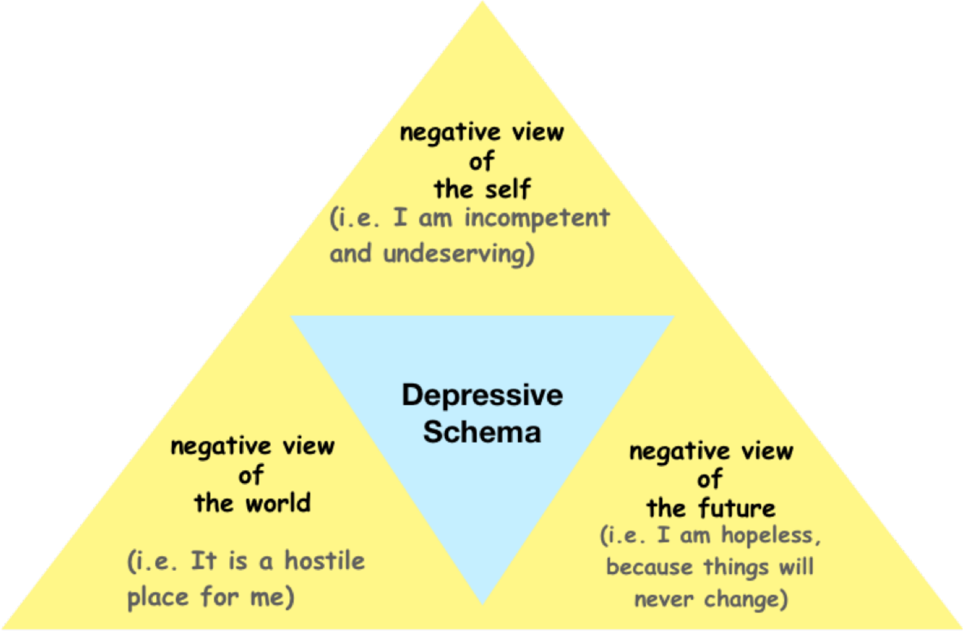


Figure 2.2: Beck's cognitive theory of depression.

2.1.3 Cognitive Schemata-based Theories of Depression

There are a variety of innovative postulates which have been researched to identify "cognitive schema" in depressed thoughts. For instance, some studies show that assessing originally euthymic subjects, such as recognizing the core belief of a depressed individual, appears to be the best method of investigating the severity of depressive symptoms (Dozois & Beck, 2008; Gladstone et al., 2001; Hammen & Zupan, 1984; Korobkin et al., 1998; Moore & Fresco, 2007; Young, 1999b; Young & Lindemann, 1992).

According to Aaron Beck's cognitive theory of depression (Beck, 2002), Beck assumes that depressed thoughts, which are driven by schema, cause severe depressed affect (Beck, 2002, 1979; Riskind et al., 1989). The schema can lead a depressed individual to negative perspectives about himself, the world, and the future (see the model of schema in Figure 2.2). We describe the schema as a 'package' of knowledge, which stores information and ideas about ourselves and the world around us.

Beck posits three mechanisms which are responsible for depression. The first mechanism is called a "negative self-schema" — in which self is associated with traits of helplessness (e.g., "I feel inferior to some people"), unlovability (e.g., "I am undesirable") and worthlessness (e.g., "I feel I have little value as a person"). According to Beck, negative self-schemas maintain a negative triad made up of three components:(1) the self, (2) the world and (3) the future

(Beck, 2002). For sufferers of depression, their thoughts describing negative and irrational views of themselves, their future and the world around them are symptomatic of depressed people. Some psychological literature (Brewin et al., 1992; Shestyuk & Deldin, 2010) in their findings empirically supports the view that negative traits are disproportionately prominent in the self-schemata of persons diagnosed with MDD. This indicates that people with MDD rate themselves as exhibiting negative traits more strongly and positive traits less strongly compared to healthy individuals.

The second mechanism is that people prone to depression possess a depressive schema, or a deep level knowledge structure. This structural model states that depression forms a systematic negativity pervading the cognitive processes — sufficiently produces a systematic bias in an abstraction of interpretation, short-term memory, and long-term memory (American Psychiatric Association, DSM-5 Task Force, 2013). For instance, a specific stressor, such as a huge financial change and social isolation, can trigger these schemas, causing an episode of depression.

Negative schemas seem to then lead to the third and final mechanism: errors in logic or cognitive biases. Beck argued that people who prone to depression possess cognitive schemas leading them to perceive the event in a negative way such that they may exaggerate a minor setback and believe that it is a complete disaster. Cognitive schemas enhance both automatic and controlled processing of schema-consistent, negative information in turn leading to core MDD symptoms such as sadness, hopelessness, worthlessness and guilt (American Psychiatric Association, DSM-5 Task Force, 2013).

Similarly, other studies (Parker et al., 1998; Young, 1999a; Young & Lindemann, 1992) referring to the hypothesis of "lock and key" assume that there is an existing salient schema in depression. Such a lock and key hypothesis emphasizes the developmental construction of mistaken beliefs based on the interaction between the self and the environment (Parker et al., 1998). Young (1999a) identified a set of early maladaptive schemas referring to a stable and constant theme that emerges during childhood. Thus, these schemas, which are implicitly taken for granted by an individual, play a substantial role in priori truths and could be operated as templates to detect MDD and assess different levels of depression (Korobkin et al., 1998; Moore & Fresco, 2007; Young, 1999b).

Clinical psychologists suggest that people with depression are more focused on themselves, and less connected with others (Holtzman et al., 2017). Moreover, depressed persons presumably have more black and white views of the world and this would manifest in depression language (Al-Mosaiwi & Johnstone, 2018). It is found that people with cognitive schemas become prone to produce depressive thoughts, focusing selectively on certain aspects of a situation while ignoring equally relevant information.

Given existing psychological theories of depression mentioned above, depression affects and influences the way individuals feel, think, and communicate. We know that language use reflects the thought processes of people and words they used can be assessed to gain insight into their thought processes. Undoubtedly, the words we use in our daily life can express our mental state, mood and emotion (Pennebaker et al., 2003). Both psychologists and linguists have investigated how psychological theories could manifest in language (De Choudhury et al., 2013b; Poulin et al., 2014; Stirman & Pennebaker, 2001). As a result, language analysis to identify and monitor human mental health issues has been regarded as an appropriate means of modeling mental health.

2.2 Depression Datasets

This section provides the details of a set of depression datasets that are widely employed in depression assessment systems.

2.2.1 AVEC (2013 - 2014)

The Audio/Video Emotion Challenge (AVEC) 2013 uses a depression corpus that includes 340 video recordings of 292 subjects performing a human-computer interaction task (Valstar et al., 2013). The video files each contain a range of vocal exercises, including free and read speech tasks. The level of depression is labeled with a solitary score for each recording utilizing the Beck Depression Inventory–Second Edition (BDI-II). Recording lengths fall between 20-50 minutes with a 25-minute mean value. The AVEC 2014 corpus (Valstar et al., 2014) is a subset of the AVEC 2013 corpus. The AVEC 2014 corpus includes 300 videos, with duration ranging from 6 seconds to 4 minutes. The files have a read speech passage (The North Wind and the Sun) and an answer to one of a number of questions.

2.2.2 AVEC (2016 - 2017)

The AVEC16 and AVEC17 focused on categorical assessment, and encouraged participants to address prediction of self-reported scores on the PHQ-8 scale ranging from 0 to 24, employed by a Wizard-of-Oz (DAIC-WOZ corpus). This is part of a larger corpus, the Distress Analysis Interview Corpus (DAIC), that contains clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. The provided dataset has been split into a training set having 107 patients and a development set containing 35 patients. For each patient in the training and development sets, a PHQ-8 score and binary depression decision are provided.

2.2.3 AVEC 2019

The AVEC 2019 (Ringeval et al., 2019) proposes Detecting Depression with AI Sub-Challenge (DDS). The level of depression (PHQ-8 questionnaire) was assessed from audiovisual recordings of US Army veterans' clinical interviews conducted by a virtual agent driven by a human as a Wizard-of-Oz (DAIC-WOZ corpus). The DAIC corpus contains new recordings with the virtual agent being, this time, fully driven by artificial intelligence, i.e., without any human intervention. The AVEC 2019 corpus includes interviews of 275 subjects for a total duration of more than 73 hours. Thus, besides the automatic assessment of the severity of depression, DDS also seeks to understand how the absence of a human controlling the virtual agent influences this automatic assessment.

2.2.4 DAIC-WOZ

Distress Analysis Interview Corpus (DAIC-WOZ) (Gratch et al., 2014) dataset is provided by the University of Southern California. A total of 50 hours of data was collected from 189 folders (clinical interviews) from 142 patients. The dataset contains video-based facial actions, audio and the conversation transcribed to text for each participant. This corpus is created from semi-structured clinical interviews where the participant speaks to a remote-controlled digital avatar Ellie. The clinician, through the digital avatar, asks a series of questions specifically aimed at identifying depressive symptoms. The agent prompts each patient, with queries that included questions and conversational feedback. a PHQ-8 score ranging from 0 to 24 and a binary depression decision are provided for each participant. D'mello and Kory (2015) and Kroenke et al. (2009) defines cut-points at [0,5,10,15,20] for minimal depression, mild depression, moderate depression, moderately severe depression, and severe depression, respectively.

2.2.5 DementiaBank

The DementiaBank Database (Becker et al., 1994) represents data collected between 1983 and 1988 as a significant aspect of the Alzheimer Research Program at the University of Pittsburgh. DementiaBank is a shared database of combined media communications for the study of correspondence in dementia. A subset of the participants from the dataset also has HAM-D depression scores.

2.3 Depression and NLP Application for Text-based Diagnosis of Mental Illness

Over 300 million people worldwide have been affected by depression which may cause suicide (World Health Organization, 2017). The impact of depression can be exacerbated by other societal and environmental factors such as COVID-19 (Santomauro et al., 2021). Thus there is a pressing need to find a convenient and automated method to assess depression severity. Moreover, the automatic depression diagnostic system can provide effective support to psychologists in the diagnostic process.

Textual data, including transcripts of clinical interviews or notes describing patients' mental states and non-clinical text such as social media posts, provide a wealth of information that expresses the emotional state and mental health of the authors of their texts. Natural language processing methods demonstrate promising improvements to enhance proactive mental health care and facilitate early diagnosis of symptoms of major depression or moderate-to-severe depressive symptoms (Haque et al., 2018; Lin et al., 2020; Valstar et al., 2016).

Detecting mental illness from text can be cast as a text classification or sentiment analysis task, where we can leverage NLP techniques to automatically identify emotional indicators in mental illness. In recent years, sentiment analysis — a subfield of NLP — has been applied to automatically classify or detect disease-related emotional polarities in texts (Balani & De Choudhury, 2015; Delahunty et al., 2018; Deshpande & Rao, 2017). The sentiment analysis approach can be effective in detecting the level of depression by exploring how the depression level relates to the emotions that people recall when asked to report their recent feelings (Li et al., 2020).

Depression, a worldwide mental illness, is the most likely negative emotion associated with a psychopathological consequence (Blanco & Joormann, 2017; Lovibond & Lovibond, 1995; Rottenberg, 2017). Depression is associated with profound dissatisfaction in emotional experiences, such as hopelessness, lack of interest and etc (Yang et al., 2012). In order to capture complex associations between emotional dimensions and depression levels, sentiment analysis approach is used to monitor and analyze an individual's mental well-being or psychological conditions from a wide variety of textual data (Jackson et al., 2017; Mukherjee et al., 2020). Consequently, sentiment analysis shows rapid growth in the domain of health and well-being.

Automatic depression detection based on sentiment-aware NLP techniques is an area of ongoing research in the fine-grained classification of depression, such as recognizing a state of depression in a range of non-depressive, mild depressive, moderate depressive and severe depressive individuals (**de2021profile**; Burdisso et al., 2019). Individuals prefer to use their own words to express their mental states, moods, and feelings. The link between language and the psychological state of people has led to the exploration of data from textual sources (Bathina et al., 2021). Emotions which were proved to be veritable risk indicators for the development of depressive disorder enhance the detection of signs of depression (Deshpande & Rao, 2017). However, many clinical analyses, such as the assessment of emotional consequences of illness in depression or pattern classification of valence in depression, make NLP challenging in this area. For instance, a patient who may incorrectly report his mental state unconsciously or intentionally can mislead the diagnosis.

2.4 NLP Techniques with Machine Learning

ML-based models have been used for NLP downstream tasks relying heavily on feature engineering and feature extraction. Traditional machine learning algorithms used for mental illness detection are based on feature selection procedures. The most frequently used features are based on Linguistic Patterns of Part-of-speech (POS) (Birjali et al., 2017), Bag-of-words (BOW) (Lin et al., 2017), Linguistic Inquiry and Word Count (LIWC) (Islam et al., 2018; Pennebaker et al., 2001) and statistics such as n-gram (Shickel et al., 2020). However, these methods are inapplicable due to limitations of quantitative performance evaluation (Bengio et al., 2000).

Traditional machine learning models are designed to learn patterns from text in terms of a combination of various extracted features. Designing these inflexible hand-engineered features can be extremely time-consuming, expensive and difficult to be applied to dynamic and flexible situations (Hamilton et al., 2017). However, deep learning techniques allow models to automatically capture valuable features without feature engineering, contributing to significant improvements. For NLP applied to detecting mental illness from text, deep learning techniques have recently attracted more attention and have shown better performance than traditional machine learning methods (Collobert et al., 2011; Naseem et al., 2020; Yenduri et al., 2021; Zhang, Wang, et al., 2020).

Deep learning-based frameworks mainly consist of two layers: 1. An encoder layer; 2. A decoder layer:

The encoder layer generates embeddings by transforming inputs which are sparse one-hot encoded vectors into dense vectors. These embeddings can preserve semantic and syntactic information such that deep learning models can be better trained. Recent embedding techniques such as language modeling Embeddings from Language Models (ELMo) (Peters et al., 2018) and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) shift the paradigm from initializing the encoder layer (or the first layer) of deep models to pretaining the entire model with hierarchical representations. These popular language modeling methods enable to generate different embeddings for a word that capture the context of the word. Other embedding techniques such as GloVe word embeddings (Pennington et al., 2014) and Word2vec (Mikolov et al., 2013) boost two to three percentage points on most tasks with limited training data (Kim, 2014). Word2vec and GloVe, as an approximation to language modeling, are widely applied as a transfer learning baseline processing multiple NLP related downstream accelerating convergence and, in the meantime, avoiding the overfitting issue (Socher et al., 2011; Turney & Pantel, 2010).

For the decoder layer, we denote it as a classifier or a regressor performing a downstream task such as mental illness detection or prediction. The most popular deep learning-based methods used for downstream tasks are Convolutional Neural Network (CNN)-based methods, Recurrent Neural Network (RNN)-based methods, transformer-based methods and hybrid-based methods. Some studies utilizing CNN involve One-dimensional Convolutional Neural Network (1D-CNN) used in the NLP field. A 1D-CNN obtains context information by imitating n-grams. In the first layer of 1D-CNN, the size of the filter region can be considered as n in the n-grams. The work of Chen et al. (Chen, 2015) applied a 1D-CNN model with pre-trained word embeddings for text classification. The RNN and its variants such as Long Short-term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Cho et al., 2014) are effective to capture contextual information of the text. In LSTMs, the input gate controls how much new information from the current input is allowed into the cell state. The forget gate determines how much information will be forgotten from the previous cell state. The output gate controls how much the current output depends on the current cell state. GRUs are a variant of LSTMs. GRUs have two gates, a reset gate and an update gate, that control how much information from the past is forgotten or retained. Some studies based on LSTM or GRU exploited an attention mechanism to find significant word information from text (Ahmed et al., 2021; Luong et al., 2015).

Moreover, there are many other deep learning models such as transformer-based methods and hybrid-based methods. Transformer architectures can capture long-range dependencies using attention and recurrence (Zhang, Wang, et al., 2020). A Transformer consists of an encoder and a decoder. An encoder block is mainly composed of a multi-head self-attention module and a position-wise feed-forward network. Compared to the encoder block, a decoder block additionally inserts an attention module between the multi-head self-attention module and the position-wise feed-forward network that helps the decoder focus on relevant parts of the input sentence. The Transformer relies entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution.

Some research studies exploit pre-trained language models such as BERT or ELMo for NLP tasks including text classification and sentiment analysis (Naseem et al., 2020; Zhang et al., 2021; Zhang, Wang, et al., 2020). BERT is a transformer-based language representation model that utilizes the masked language model to predict random words that are masked in a sequence and to subsequently learn bidirectional representations. ELMo trains a bi-directional LSTM and concatenates the results to produce word embedding. ELMo considers different aspects of words including their usage in a specific context. The usage and development of large-scale pre-training models achieved state-of-the-art on widely used NLP tasks including question answering, text classification and other applications (Naseem et al., 2020; Terechshenko et al., 2020). Hybrid-based methods (Naseem et al., 2019; Tadesse et al., 2019) that combine several neural networks for text classification have been used. Tadesse et al. (2019) employed an LSTM-CNN combined model to extract local features and sequence features. Their model outperformed the individual CNN or LSTM classifiers.

Convolutional neural models find it difficult to learn the dependencies between distant positions (Battaglia et al., 2018). In comparison, sequential models can connect contextual memory and store more long-term global information. Therefore sequential models, in general, achieve good performances in text analysis (Battaglia et al., 2018; Liao et al., 2021). However, these methods mainly focus on consecutive word sequences, they do not encode word co-occurrence information in an explicit way, and their complex model structure is like a black-box (Castelvecchi, 2016; Sarker, 2021; Zhang, Cui, et al., 2020).

GNNs belong to an emerging area that has also made a tremendous impact across technological domains. Graphs, due to their unique structural properties, inherently capture relationships between entities and are thus potentially very useful to encode relational information between variables. Graph-based deep learning approaches apply message passing to learn feature representations for each feature in a node, in which nodes iteratively aggregate feature vectors from their neighborhood to compute a new feature vector at the next hidden layer in the network (Kipf & Welling, 2016a). Different GNN variants use different aggregators to gather information from each node's neighbours and use varied methods to update the hidden states of nodes. Convolutional GNNs, such as Graph Convolutional Neural Networks (GCNs) stack layers of learned first-order spectral filters followed by a nonlinear activation function to learn graph representations. There are two major types of GCNs: spatial convolutional networks (Defferrard et al., 2016; Kipf & Welling, 2016a) and spectral convolutional networks (Hamilton et al., 2017; Niepert et al., 2016). The Graph attention network (GAT) (Veličković et al., 2017) applies the idea of self-attention to graph representation learning. It computes the hidden representations of each node on graph by assigning different importance weights to nodes of the same neighborhood.

Recently, GNNs have been increasingly utilized for many NLP applications including text classification, relational reasoning, conversation generation, and question answering. High performance in some text classification tasks has demonstrated the potential value of graph-based deep learning models in NLP applications (Liang et al., 2022; Liao et al., 2021; Yao et al., 2019; Yasunaga et al., 2021). Yao et al. (2019) showed a pioneering way of converting a text classification problem to a node classification problem. They used a Text Graph Convolutional Neural Network (Text GCN) to achieve strong classification performances with a small proportion of labeled documents without any external pre-trained word embeddings such as Word2vec and GloVe. Peng et al. (2018) proposed a graph-CNN-based deep model to first convert text to graph-of-words and then used graph convolution operations to convolve the word graph. Wang and Li (2022) applied GAT model to update the representation of a node by assigning different weights to its neighbours and then fused all the nodes in the graph together into the document embedding. There remain challenges in applying graph-based deep learning methods such as in selecting the appropriate types of graph structures and to what extent these graph structures help to improve the performance of the domain task (Wang et al., 2021).

2.5 Automatic Depression Detection with Machine Learning

In general, automatic depression detection techniques first extract different types of features from interview facial expressions/audio of patients who are asked a set of carefully crafted questions from different topics. Models are trained using the extracted features, the indicators include visual, speech and linguistics, to measure the severity of depressive symptoms or generate a prediction of the presence of depression (Alhanai et al., 2018; Cummins et al., 2017; Gong & Poellabauer, 2017; Haque et al., 2018; Lin et al., 2020; Sun et al., 2017; Tsakalidis et al., 2018; Valstar et al., 2016; Williamson et al., 2016).

Early studies of automatic detection of depression have made great efforts to extract effective features from highly correlated interview questions. Gong and Poellabauer (2017) proposed a topic modeling to preserve important information in long interviews. They utilized a context-aware analysis to enhance the performances of both depression detection and depression severity prediction. Some work (Cummins et al., 2017; Valstar et al., 2016) implemented statistical functions (e.g., max, min) on short-term features over a long interview transcript, but they fail to preserve useful temporal information (i.e., some short-term signs of regret, anxiety, etc.) across the long interview. Williamson et al. (2016) proposed a Gaussian Staircase Model by analyzing the semantic context to obtain coarse depressive descriptors. Sun et al. (2017) extracted text features from the topic-related questions such as “Have you been diagnosed with depression/Post-traumatic Stress Disorder (PTSD)?” and applied a random forest to detect depression.

More recently, there has been an emergence of deep learning techniques that learn representations automatically without feature engineering, which helps make great improvements in depression detection systems.

Some work (Alhanai et al., 2018; Haque et al., 2018; Ringeval et al., 2017) applied a sequence-level deep learning model to capture implicit depressive signals. Such models, in general, use a multi-modal sentence embedding to predict the severity of depressive symptoms. Alhanai et al. (2018) proposed a deep model which was trained jointly with the acoustic and linguistic features. Haque et al. (2018)introduced a new way of learning a multi-model sentence embedding to predict depressive symptom severity. By building an innovative causal neural network, they aimed to transform all multi-modal features to a sentence-level embedding, and then they made a prediction based on this sentence-level embedding. Lam et al. (2019) proposed a deep learning model using multi-head attention modules to extract contextual information from clinical text and a 1D-CNN to extract audio features. They used text features and audio features that were highly correlated with depression severity. Lin et al. (2020) utilized a Bidirectional Long Short-Term Memory (BiLSTM) with an attention layer to deal with linguistic content and a 1D-CNN to deal with acoustic features. They implemented a fully connected layer to summarize all embeddings from both audio and text and make a prediction for the severity of depression.

There are existing studies on utilizing deep learning models that leverage pre-trained language modeling (i.e., ELMo or Bert) or pre-trained word embeddings (i.e., GloVe) to extract global features from text (Mallol-Ragolta et al., 2019; Ray et al., 2019; Solieman & Pustozerov, 2021; Zhang, Wang, et al., 2020). Their results demonstrate the effectiveness of training downstream deep learning models with pre-trained language representations.

Although promising results have been obtained using both machine or deep learning methods, great difficulties still exist in practice. For instance, videos of clinical interviews may not be available due to the privacy problem. Since language functions play an important role in the detection of cognitive impairment cross different levels of MDD, speech transcripts can assist in the early detection of the disease. Hence techniques at the nexus of natural language processing and deep learning offer an inexpensive solution to this early detection problem (Al-Mosaiwi, 2018; Croisile et al., 1996; Rude et al., 2004).

Furthermore, for methods (Lam et al., 2019; Lin et al., 2020; Ray et al., 2019; Williamson et al., 2016; Zhang, Wang, et al., 2020) utilizing only one type of features, methods based on text features perform better than those based on either audio or visual features in both binary classification of depression and prediction of severity of depressive symptoms. As a consequence, research has achieved high performance in depression detection by utilizing text modality alone. It was found that text features are often the most informative indicators obtained by analyzing the patients' utterances (Williamson et al., 2016), and human language alone can be a very good predictor of depression among those multimodalities (De Choudhury et al., 2013b).

2.6 Interpretability and Explainability in Machine Learning

Molnar (Molnar, 2020) defines interpretability as the degree to which a human can understand the cause of a decision or the degree to which a human can always predict the results of an ML model, and explainability is a method for understanding why particular decisions were made by the model.

Deep learning-based methods achieve good performance by utilizing feature extraction and complex neural network structures for illness detection. Although they relieve much of burden of hand designing features, it pays the cost of interpretability (Castelvecchi, 2016; Hamilton et al., 2017; Zhang, Cui, et al., 2020). Augmenting existing clinical methods by improving predictive accuracy is not enough for future research, as this high-level task may require an in-depth study of understanding why the model made certain predictions. It is important, when guiding clinicians, that they understand not only what has been extracted from text but the reasoning underlying the predictions.

With the growing success of the usage of deep learning techniques applied in mental health, the interpretation and explanation of model behavior has become important to boost the detection performance, to empower decision-making, and to avoid severe misdiagnosis consequences. Recently, explainability of deep models on images and texts has achieved significant progress. The development of post-hoc techniques to explain the predictions gives a rise to explanation techniques of deep models. For instance, explaining a black box by using an approximation model, derivatives, variable importance measures, or other statistics.

In general, the area of explainability of deep models focuses on studying the underlying relationships behind the predictions of deep models. Some studies interpret the model using input-dependent explanations, studying the important scores for input features, or a high-level understanding of the general behaviors of deep models (Dabkowski & Gal, 2017; Du et al., 2018; Selvaraju et al., 2017; Simonyan et al., 2013; Zhou et al., 2016). Some existing approaches apply dimensionality reduction techniques to project high-dimensional data to a lower-dimensional and human-comprehensible space (Du et al., 2018; Selvaraju et al., 2017; Zhou et al., 2016). Oh et al. (2019) designed a gradient-based visualization method to highlight the important input features. Wang et al. (2020) applied t-distributed stochastic neighbor embedding T-distributed Stochastic Neighbor Embedding (t-SNE) to identify most important linguistic biomarkers which help to detect a certain disease. Wang et al. (2018) provided an explanation on attention-based networks with integrated gradients to analyze the sensitivity between the input features and the predictions. Furthermore, Clough et al. (2019) explored the meaning of hidden neurons with the called concept activation vector technique to understand the whole classification procedures. Olah et al. (2017) interpreted the model using an input-independent explanation. They study the input patterns by maximizing the predicted score of a certain class.

2.7 Conclusion

Literature review shows that NLP techniques can be effective at making inferences about a person's mental state. Psychological findings have shown that depressive symptoms can be effectively captured by certain existing patterns in the language behaviors of depressed individuals. The link between language use and depression is important and could lead to opportunities to automatically detect people at risk of depression from textual data. The next chapter describes how to use a GNN model to measure depression severity from patients' clinical text.

Chapter 3

Predicting the Severity of Depression Using a Graph Neural Network

3.1 Introduction

This chapter aims to demonstrate a depression prediction algorithm that learns patients' transcripts using graphs. The work implements an end-to-end graph representation learning to measure the severity of depression indicated by the PHQ-8 score ranging from 0-24.

The dataset contains the conversation transcribed to text for each sample interview arranging from 7 to 33 minutes. 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 (Busso et al., 2008). The challenges 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, mean, etc.) on short-term features over an entire interview may lose useful temporal information such as short-term signs in regret, despair and anxiety.

Analyzing a large data volume is typically beneficial for accuracy, since its contextual information conveys the most relevant evidence for determining depression 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 space where we can obtain contextual features of depression.

Since each interview contains hundreds of spoken words, 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 or overfitting. For example, both subject 1 and subject 2 mentioned the same word, "hopeless", in an utterance; although it is a strong short-term signal indicating depressive states, weighting both instances of "hopeless" at a 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 the representation of a word using connection information between this word and its neighbors. For instance, given a specific context, the "hopeless" is following the word "I" instead of the word "he", we can learn to represent the word "hopeless" which depends on the correlation between ("i", "hopeless"). Because of the way in which the graph structure encodes information about the importance among words in a context, deep learning on graphs can efficiently represent the regression using a small number of parameters.

In order to overcome these challenges, I introduce an automatic depression detection method based on the GNN model to measure the severity of depressive symptoms ranging from 0 to 24 using only text. I investigate the effectiveness of graph representation learning for depression prediction task by comparing to general deep learning models, such as CNN, LSTM and etc. This chapter shows two advantages of using a graph representation learning: 1. It facilitates learning context-level semantic features derived from word entities and relations. 2. It can learn to efficiently represent high-dimensional probability distributions while requiring a very small number of parameters.

3.2 Text-based Depression Indicators

Semantic context features: Inspired by the observation of a dataset provided with text transcripts, there are a variety of words generated by patients as they describe their opinion and depressive symptoms. Depressed language is common in the context of depression and can be used to measure various levels of depression. Some empirical psychological findings (Al-Mosaiwi, 2018) also demonstrate that utterances of depressed people directly and explicitly manifest cognitive biases presenting in their depressed thoughts. Depressed language reflecting cognitive bias lives in depressed minds. It is composed of a class of typical linguistic markers and their connections with other word entities. According to Beck's cognitive theory, depressed language can form a "depressive schema" storing information about depression.

There are two typical types of depression-related cognitive biases that can be effectively 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" (Zimmermann et al., 2017). 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 finding highlights a certain style of language (Holtzman et al., 2017) 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 (Bai et al., 2018; Beckham et al., 1986; Du et al., 2019; Hamilton et al., 2017), I found that the associations between cognition-based word indicators can strongly and explicitly signify depression in utterances. My motivation is to learn underlying correlations between these key words and their adjacent and non-adjacent words by using co-occurrence information among words; thus I could capture semantic context features (i.e., depressive features) and utilize them to generalize depression levels. For example, associations between first person singular pronouns and their adjacent words, i.e., ("I", "am, hopeless") efficiently create a context of PHQ-8 criteria for depression. Examples are shown in Figure 3.1. This kind of information uses graphs (Battaglia et al., 2018) can be encoded because graphs are particularly capable of representing strong relational inductive biases, which could perform efficient reasoning by exploiting the graphical structures within text.

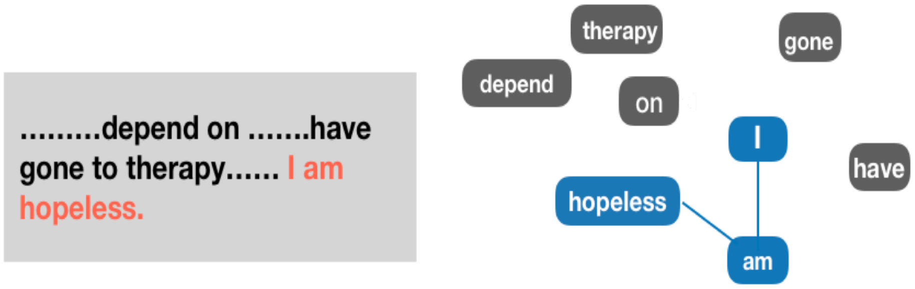

.........depend on ....have gone to therapy...... I am hopeless.
therapy
gone
depend
on
I
have
hopeless
am

Figure 3.1: The figure on the left shows an example of a single text "I am hopeless" extracted from a raw transcript. On the right-hand side, I build a graph for this single text. An edge exists in the graph for all words which are in the neighborhood of that word. The size of the neighborhood is defined by a parameter p. For the convenience of the display, I set $p = 1$ for displaying associated edges among the nodes (colored in blue).

As a result, I propose to exploit the power of graphs to develop a deep neural network operating over graphs (Battaglia et al., 2018; Hamilton et al., 2017; Kipf & Welling, 2016b), for the purpose of learning a flexible graph representation for depression score prediction.

3.3 Proposed Model and Method

In this section, I present the model structure of a graph neural network in detail. I first show how to construct the graph structure for each patient transcript. Then I introduce how to use graphs to update node representations. Finally, the output and loss function of the model applied for the domain task are introduced.

3.3.1 Graph Construction

I constructed individual graphs for each text. I represented words as nodes and the co-occurrence relationship between words are edges, denoted as $G = (V, E, H)$, where 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. The embeddings of nodes in each text graph were initialized with pre-trained GloVe embedding, denoted as $H \in \mathbb{R}^{|V| \times d}$, where $|V|$ denotes the total number of unique words in a transcript, and d is the dimension of a word vector.

I first preprocessed the text, including cleaning and tokenizing (Adel & Shi, 2021) to obtain the word sequence. Then I applied a fix-sized sliding window (I set $p = 4$ at default) to detect edges according to word co-occurrence on word sequence. I define the edge set E like this:

$$E = \Big\{ e_{ij} | i \in \big[1, |V|\big]; j \in \big[max(1, i-p), min(|V|, i+p)\big] \Big\}$$

For example, given a sentence sequence S1 (in figure 3.2), if I set $p = 1$, there is an edge between word 'i' and word 'am', word 'am' and word 'i', a self-edge of word 'i', a self-edge of word 'am', and so on. An example of constructing a text graph is shown in Figure 3.2.

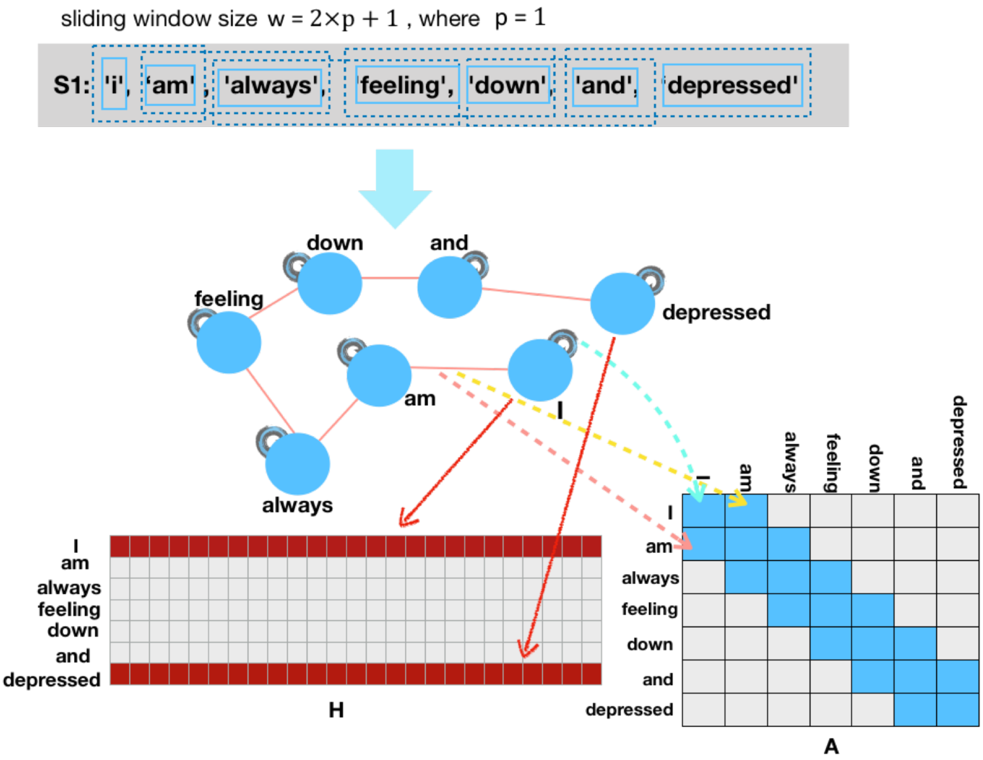

sliding window size w = 2×p + 1 , where p = 1
S1: 'i', 'am', 'always', 'feeling', 'down', 'and', 'depressed'
down
and
feeling
depressed
am
I
always
I
am
always
feeling
down
and
depressed
H
I
am
always
feeling
down
and
depressed
A

Figure 3.2: An illustration of constructing a text graph for a single sentence extracted from a transcript of a subject. S1 represents the word sequence after preprocessing. I set $p = 1$ in the figure for displaying associated edges among the nodes. $A \in \mathbb{R}^{|V| \times |V|}$ represents the adjacency matrix. H represents the global shared feature matrix.

I consider a set of graphs $\{G_1...G_Z\}$, and their labels $\{y_1, \cdots, y_Z\}$[1]. T is the total number of transcripts (or the sample size). I aim to learn a representation vector h_{G_i} that can be used to predict the label of a new unlabeled graph, G_q.

3.3.2 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. I use the Message Passing Mechanism (MPM) (Battaglia et al., 2018; Gilmer

[1] For input graphs in the training process the labels are integers in the range $0 - 24$.

et al., 2017; Xu et al., 2018) which is a message function applied to push messages from the surrounding nodes around the node v.

The graph performs message passing between nodes in order to learn the representation of each node which captures the structural information within its network neighborhood. MPM is defined as:

$$m_v^{k+1} = \sum_{u \in N(v)} h_u^k \tag{3.1}$$

This message function is where I aggregate all messages coming from neighbors of node v shown by $N(v)$ and its message passing phase runs for K iterations. m_v^{k+1} is a message vector of v at $(k+1)$-th iteration.

Next, node embedding is updated by message passing using a node update function:

$$h_v^{k+1} = \sigma^k\Big(m_v^{k+1}, h_v^k\Big) \tag{3.2}$$

where $\sigma^k(.)$ is a 1-layer Multi-layer Perceptron (MLP).

In this step, I get the new embedding of the node v which is updated by holding messages that encode correlations between itself and its neighbors.

Finally, I obtain the final set of embeddings of each node in the convolution unit at the final layer. I apply a READOUT function which aggregates node features from the final iteration and sums them up together to get the graph-level embedding h_{G_i}:

$$h_{G_i} = \text{READOUT}\Big(\Big\{(h_v^K | v \in G_i\Big\}\Big) \tag{3.3}$$

Finally I apply a linear function, performing as a Fully Connected (FC) output layer. The output of the FC layer is y_i which is a PHQ score: $y_i = g(h_{G_i}, w)$ where w are the weights of the FC layer.

3.3.3 Loss

I minimize a loss function by updating the parameters through backpropagation of gradients. For assessing the severity of depressive symptoms, Root Mean Squared Error (RMSE) loss, as defined in Equation 3.4, is chosen to be the criterion function.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i^{pred} - y_i^{true})^2} \tag{3.4}$$

where y_i^{pred} denotes the ith predicted value, y_i^{true} the actual score. Note that, the RMSE metric has the same units as the corresponding score, and lower values of RMSE indicate better performances.

As we shall see later in section 3.4 which presents the experiments performed using this GNN model. In those experiments, the GNN model demonstrates a strong representation learning capability to learn more flexible and strong contextual semantic features. Because of the way in which the model learns a mapping between a graph and a depression level, in the experiments the proposed model improved the accuracy of the prediction compared to the state-of-the-art methods.

3.3.4 Overall Pipeline

I present a pipeline of applying a GNN model to learn a mapping between a long text (i.e., a clinical transcript) and a domain task purpose (i.e., predict a depression score representing the severity of a patient's depression). The diagrams (Figure 3.3 + Figure 3.4) illustrate an instance of using a graph representation learning method to examine a depressed state from a patient transcript.

I first create a graph from a raw patient transcript. Next, I implement a GNN model to learn representations of each node by aggregating information from their adjacency. Running the message aggregation mechanism, the GNN model functions as a graph algorithm to update each node representation of a graph (note that each graph represents a separate transcript). Figure 3.3 displays these two steps. Finally, I concatenated all node embeddings for making a graph-level prediction. The output of the model is a PHQ score (shown in 3.4).

3.4 Evaluation

The experiments consist of two parts. First, I predict the PHQ score for each participant. I then compare (in section 3.4.3) the performance of the model to state-of-the-art deep models in

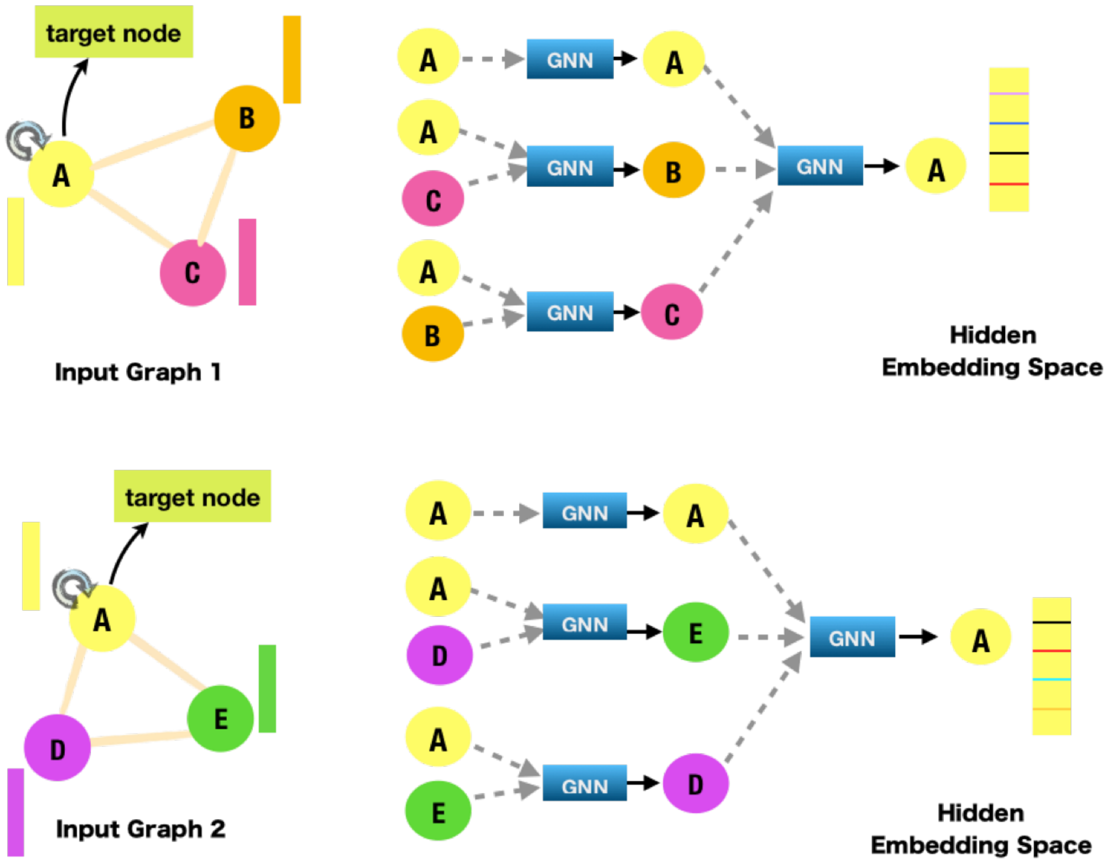


Figure 3.3: The figure showing a pipeline of an embedding of node A (in yellow) in an input graph generated by the proposed GNN model. The embeddings of the rest of the nodes are generated in a similar way. In each graph input, nodes represent unique words in a transcript and lines between nodes represent word-word relations. The information of a word is collected by its adjacent nodes and is used to update the state of the node (represented by a colored rectangle attached to the corresponding node).

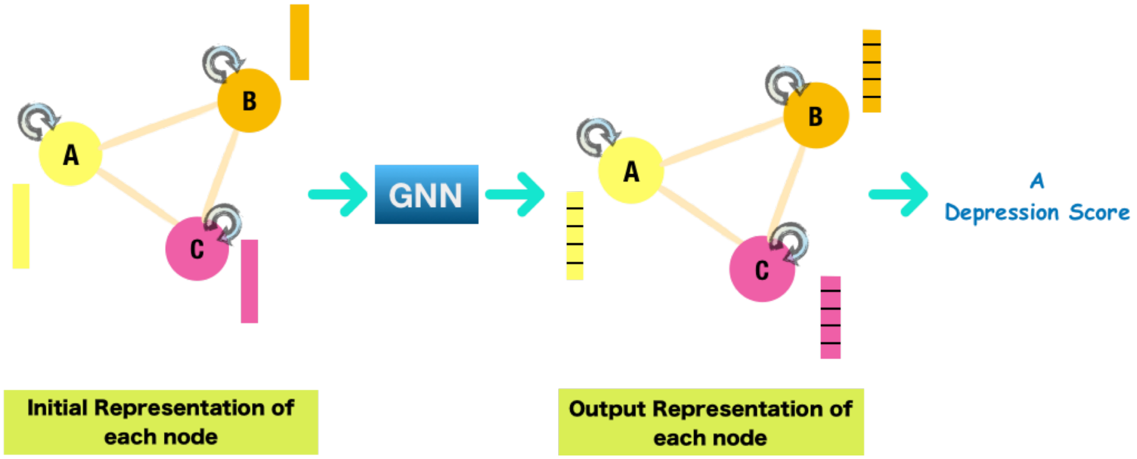
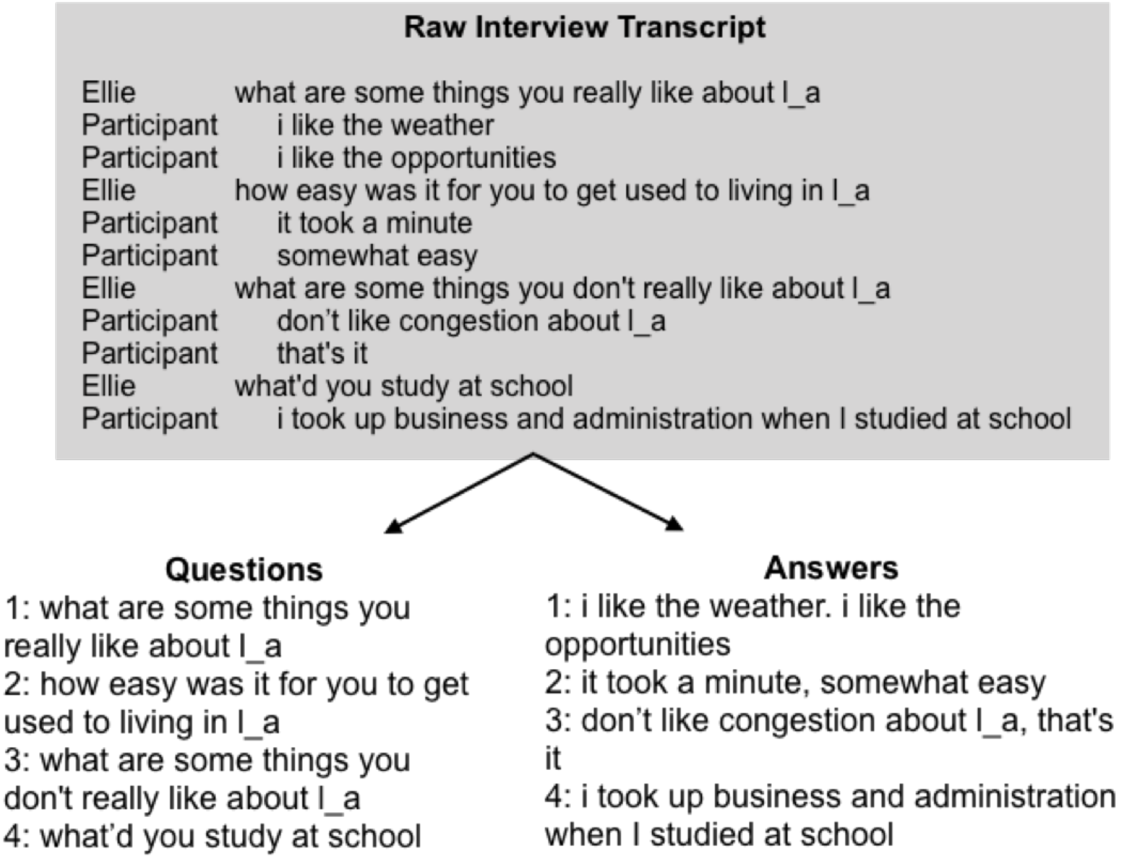

A
B
C
GNN
A
B
C
A
Depression Score
Initial Representation of each node
Output Representation of each node

Figure 3.4: An illustration of a training phase of the proposed GNN model. Each node represents a unique word in a transcript. I initialized each node with a pre-trained word embedding and all nodes were updated during training. Overall, the feature extraction process in test phase is similar to the process in this training phase. I learned a regression model and used it to predict the depression severity score of the patient.



Raw Interview Transcript

Ellie what are some things you really like about l_a
Participant i like the weather
Participant i like the opportunities
Ellie how easy was it for you to get used to living in l_a
Participant it took a minute
Participant somewhat easy
Ellie what are some things you don't really like about l_a
Participant don't like congestion about l_a
Participant that's it
Ellie what'd you study at school
Participant i took up business and administration when I studied at school

Questions

1: what are some things you really like about l_a
2: how easy was it for you to get used to living in l_a
3: what are some things you don't really like about l_a
4: what'd you study at school

Answers

1: i like the weather. i like the opportunities
2: it took a minute, somewhat easy
3: don't like congestion about l_a, that's it
4: i took up business and administration when I studied at school

Figure 3.5: Example of automated extraction of questions(or queries) and responses

Table 3.1: DAIC-WOZ dataset summary with the total number of participants gathered, in addition to the number of participants categorized as depressed. Since the test set is not in the public domain, I use both training and development sets.

	Train Set	Devel. Set	Test Set	Sum
Total participants	107	35	47	189
Depressed participants	31	12	9	52

measuring the severity of depression (see Table 3.3). Second (in section 3.4.4), I compare the GNN to baseline models using the text modality alone. The experimental results demonstrates the effectiveness of learning graph-level embeddings. The proposed approach is superior to state-of-the-art methods in both Table 3.2 and Table 3.3.

3.4.1 Text Features

Data: The **DAIC-WOZ** dataset contains video-based facial actions, audio and the conversation transcribed to text for each participant. Both the 6^{th} and 7^{th} International AVEC (Valstar et al., 2016) used this dataset (see Table 3.1). I utilized only the text transcripts of all 142 individuals within the dataset. 43 out of 142 subjects (30%) were labeled as depressed. The

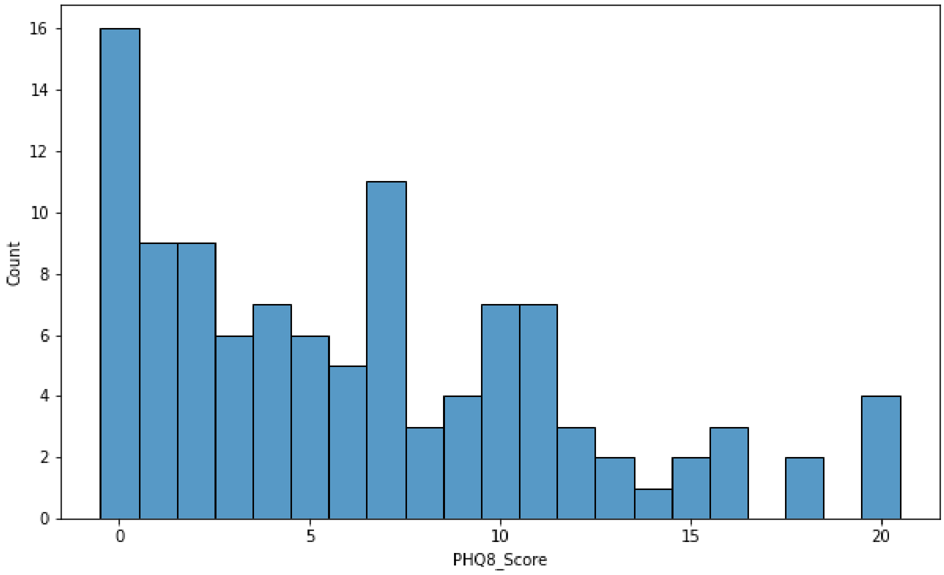


Figure 3.6: A histogram describes the distribution of ground-truth depression labels (PHQ scores) on the training set. The x-axis represents a PHQ score ranging from 0 to 24. The y-axis represents the number of subjects for each score.

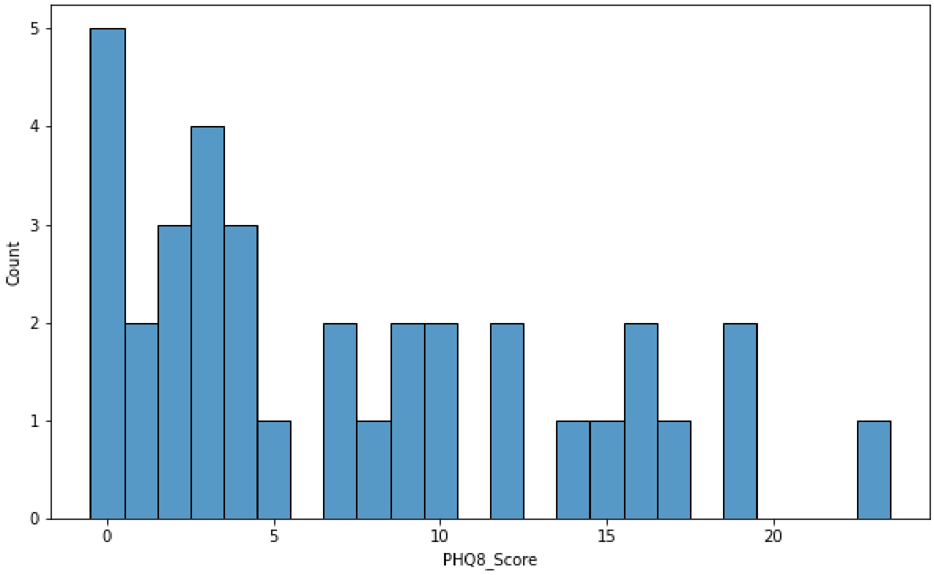


Figure 3.7: A histogram describes the distribution of ground-truth depression labels (PHQ scores) on the development set. The x-axis represents a PHQ score ranging from 0 to 24. The y-axis represents the number of subjects for each score.

provided dataset has been split into a training set having 107 patients and a development set containing 35 patients. In line with prior work (Alhanai et al., 2018; Cummins et al., 2017; Haque et al., 2018; Valstar et al., 2016) and to ensure comparable results, I 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.

Preprocessing

- Given a raw transcript, question-answer pairs were formed each time a remote-controlled digital avatar Ellie asked a new question. Examples are shown in Figure 3.5. In the experiments, the features used for depression prediction were extracted from patients' responses. To avoid relying on the clinicians' expertise and to increase the generalization capability, I adopt just patients' answers as the model's training inputs. The average length of the patient transcripts is nearly 1402 words.
- Slang words present in the transcripts are canonicalized. For example, 'lookin' was translated to 'looking', 'wanna' was translated to 'want to' and 'bout' was translated to 'about'. All text is forced to be lowercase. Lemmatization is applied to pre-process raw transcripts.
- The most frequently occurring stopwords such as 'a', 'on', 'at', and 'in', were automatically removed based on the hypothesis that these words do not carry much information.
- Word tokenization is applied before feeding into the training model.

3.4.2 Experiment Setup

I set the size of the word node embedding as 200 and initialized with GloVe (Mikolov et al., 2013). I set p as 4 in equation 3.2 giving the sliding window size of 9 (i.e., 2p+1). I set the learning rate as 10^{-3}, $L2$ weight decay as 10^{-4}, and dropout rate as 0.5. The batch size of the model is 16. The loss objective is RMSE. I trained the GNN for a maximum of 500 epochs using the ADAM optimizer (Kingma & Ba, 2014) and stopped training if the validation loss does not decrease for 10 consecutive epochs.

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

3.4.3 A Comparison Against Baselines

Deep learning models, in general, learn distributed representations of words sequentially. Two representative deep learning networks are LSTM (or GRU) and CNN. These two models are frequently used to measure the severity of depressive symptoms (a regression task) or to detect depressed or non-depressed (a classification task). However, given the challenges mentioned in section 3.1, I propose a graph-based deep learning model to estimate various depressive states.

I evaluated the performance of two types of embeddings summarized from different DL models respectively: 1. A sentence-level embedding ; 2. A graph-level embedding. The experimental result of the comparative methods evaluated on the DAIC-WOZ dataset is shown in Table 3.2. The following section gives a detailed description of these DL models.

3.4.3.1 Experimental Approach

Experiment 1 with LSTM:
LSTMs: A LSTM network has embedding vectors that are passed to the LSTM layer. The final hidden state is connected through a fully connected layer and then it is connected to a linear layer which outputs a PHQ score. The model consists of three layers, followed by a fully connected layer. The output of the connected layer is fed into a linear layer to produce a PHQ score. For the loss function, I used MAE for regression. The optimizer algorithm was the adam optimizer with $\beta 1 = 0.9$ and $\beta 2 = 0.999$ with L2 weight decay of 1e-4. Dropout was applied with a 0.5 probability of being zeroed. The initial learning rate was 1e-4. A batch size of 16 was used.

Experiment 2 with GRU:
GRUs: GRUs are similar to LSTMs but they have fewer tensor operations; therefore, they are a little speedier than LSTMs. The experiment setup used for the GRU model is as same as for the LSTM model.

Experiment 3 with T-CNN:
Temporal Causal Neural Network (T-CNN) : T-CNN (Cummins et al., 2017) is derived from a CNN model and has a causal convolution as a sequence model that convolves only on the elements from current timestamp or earlier in previous layers for prediction. I implemented a 10-layer causal convolutional network with kernel size of 5 with 128 hidden nodes per layer. Then it is sent to a fully connected layer through ReLU activation and finally it is connected

Table 3.2: Baseline comparisons. Row 1-3 are learned a sentence-level embedding. Row 4 is learned a graph-level embedding. I adopt 10-fold stratified cross-validation and report mean with standard deviation in the parentheses.

Regression: PHQ score			
Model	Feature	MAE ↓	RMSE ↓
1) LSTM	L	5.20 (0.6494)	7.49(0.6827)
2) GRU	L	4.83 (0.5994)	7.12(0.6258)
3) T-CNN	L	5.42 (0.6096)	6.32(0.7559)
4) **proposed GNN**	L	**4.16(0.4120)**	4.95(0.4107)

by a linear layer for generating a PHQ score. Dropout was applied to all non-linear layers with a 0.5 probability of being zeroed. The loss objective was RMSE for regression. The model was optimized with the Adam optimizer with $\beta 1 = 0.9$ and $\beta 2 = 0.999$ with L2 weight decay of 10^{-4}. The initial learning rate was 10^{-4} for regression. A batch size of 16 was used.

To give more reliable results, I perform a 10-fold cross-validation on the combined training and development dataset. Also, I initialized sequence inputs with 200-dimensional GloVe word embeddings. I concatenated responses of the same question (in a transcript) as a long sentence and I generated embeddings of individual responses to all questions, for a total of 8,050 training examples, 272,418 words, and a vocabulary size of 7,411.

3.4.3.2 Experiment Result

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 a graph-based deep learning model to perform the task of depression prediction. Therefore, I compared the proposed GNN model with three baseline methods that only accept text features (see the Table 3.2).

Row 1, 2 and 3 learn a sentence-level embedding. Both row 1 and 2 implement sequence-based modeling, where it uses the last hidden state as the sentence-level representation of the text. Row 3 uses convolutional neural layers operating on word embeddings to get the sentence-level representation of text.

Row 4 learns a graph-level embedding. It shows the work of using a GNN model to learn a representation of an entire text graph. Compared with the state-of-the-art results on the same dataset, the approach based on the GNN achieves the best performance.

Table 3.3: Comparison of DL based approaches for measuring depression symptoms severity on development set using MAE. The task is evaluated: PHQ score regression. Modalities: A: audio, V: visual, L: linguistic(text), A+V+L: combination. The results marked with a * are taken from the cited papers. − represents that there are no reported results in the original work. Experimental results are obtained from the provided development set; therefore, there is no reported variance.

Regression: PHQ score			
DL-based models	Features	MAE ↓	RMSE ↓
* Baseline Challenge (Valstar et al., 2016)	A + V	5.52	6.62
* C-CNN (Haque et al., 2018)	A + V + L	5.18	-
* LSTM (Alhanai et al., 2018)	A + L	5.10	6.37
* LSTM (Haque et al., 2018)	A	5.78	-
* LSTM (Alhanai et al., 2018)	L	5.18	6.38
* CNN (Song et al., 2018)	V	5.15	6.29
* DepArt-Net (Du et al., 2019)	V	4.65	5.88
* DCGAN (Yang et al., 2020)	A	4.63	5.52
* BERT + CNN-LSTM (Yang et al., 2020)	L	-	4.22
* BERT + LSTM (Yang et al., 2020)	L	-	4.97
* ELMo + BiLSTM (Lin et al., 2020)	L	3.88	5.44
GNN (the proposed approach)	L	**4.2**	-
* CNN + GNN (Chen et al., 2022)	V	3.23	3.96

In conclusion, compared with baseline models adopting only text features, the proposed GNN model based on text features performs even better, with an MAE value of 4.2.

3.4.4 A Comparison Against the Literature

In Table 3.3, I compare the proposed method to prior work on measuring depressive symptom severity under the same condition of utilizing deep learning algorithms.

The major difference between the proposed method and prior work is that this method concentrates on learning context-aware semantic features. I 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.(Cummins et al., 2017) connected to the pre-trained word embeddings.

In conclusion, from Table 3.3 it can be seen that for the methods utilizing only one type of features, the methods based on text features perform better than those based on either audio or visual features in depression severity assessment task.

3.4.5 Metaparameter Setting Analysis of GNN

The hyperparameters are tuned with different sizes of hidden layers. The hyperparameters that were used in the GNN model are shown in Table 3.4. I adopted 10-fold stratified cross-validation and reported MAE values. Here I used the same window size of 9, a dropout of 0.5, and 200-dimensional GloVe word embeddings in the experiment.

Table 3.4: Result of 10-fold cross-validation using different hyperparameters of the GNN model. I report mean with standard deviation in the parentheses.

Regression: PHQ score				
Size of Node Embedding	Size of Readout	# Params	MAE (SD) ↓	RMSE (SD) ↓
16	8	3841	3.97 (0.5652)	4.86 (0.4258)
16	16	4321	3.85 (0.5074)	4.56 (0.4990)
32	8	8097	3.97 (0.2978)	4.85 (0.6520)
32	16	8833	4.02 (0.7055)	4.64 (0.5723)
32	**32**	**10689**	**3.83(0.5625)**	**4.41(0.5625)**
64	16	19393	3.93 (0.3104)	4.53 (0.5037)
64	32	22273	3.98 (0.5209)	4.62 (0.4771)
64	64	29569	3.86 (0.6831)	4.45 (0.5931)
128	16	46657	3.84 (0.3538)	4.70 (0.0583)
128	32	51585	3.95 (0.3199)	4.53 (0.3468)
128	64	62977	3.99 (0.3109)	4.61 (0.6185)
128	128	91905	4.02 (0.3108)	4.42 (0.5224)

3.4.6 Qualitative Analysis

Some incorrect results predicted by the model may be caused by the unequal distribution of dataset across PHQ scores (see figure 3.6 and 3.7). In figure 3.8, the model made worse predictions on those patients diagnosed with the most severe depressive symptoms. For example, the model predicts a depression score of 4[3] 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 8. Given that 43 out of 142 subjects (30%) have been labeled with a major depressive disorder of PHQ (a PHQ score $\geq$ 10), I suggest that additional samples with a PHQ score $\geq$ 10 can help the model learn more major depressive features from text. Consequently, the addition of more clinical transcripts can be useful as having more data could better reveal complex patterns for

[3] This score and all subsequent predicted PHQ scores in this thesis are rounded to the nearest integer.

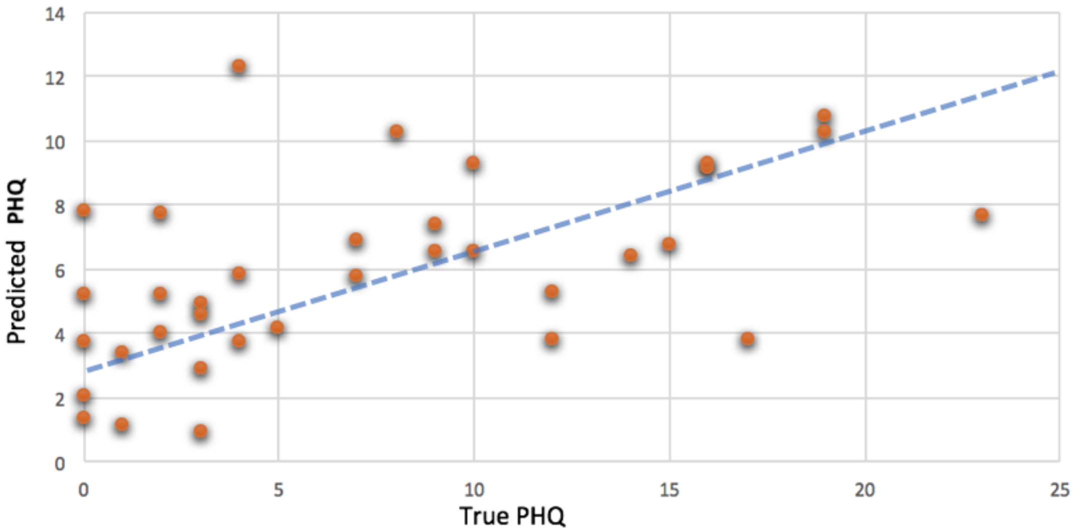


Figure 3.8: Results on development set for the graph-level PHQ prediction system, with predicted PHQ plotted as a function of true PHQ.

models to improve predictions with fewer features needed.

3.5 Conclusion

Instead of learning a transcript as a sequence input, which is most widely applied to many state-of-the-art methods, I treat a long text as a graph. To achieve this, I converted an unstructured raw transcript to a text graph. I learned a mapping encoding high dimensional probability distributions over a text-level graph. Most previous work 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. However, the proposed GNN model is capable of evaluating depression level for each subject in an end-to-end automated manner using just text. I showed the experimental results on the DAIC-WOZ benchmark demonstrating the effectiveness of the proposed approach and its superiority over other state-of-the-art methods.

By comparing performances of individual feature sets from visual, audio and text respectively, some research studies (Lin et al., 2020; Williamson et al., 2016) found that semantic analyses of dialogue transcript provided the highest performing features, and thus I suggest that future work on measuring depression intensity should also exploit semantic features to capture contextual information from patient transcripts.

Moreover, the accuracy improvement in the performance of the task of predicting the severity of depression indicates the advantage of using graph structures to learn a long text. In my hypothesis, graphs can help to capture contextual semantic features of depression from clinical text. For example, a deep graph model with its nature of relational inductive bias can exploit word co-occurrence information to capture general depressive features in an explicit way. This motivated us to learn a graph representation for a long text and to measure depressive symptoms in terms of PHQ scores.

Compared with other methods that learn sequences, the proposed approach quantifies depression levels by learning graph representation. The graph-based deep learning model is capable of evaluating depression levels for each subject in an end-to-end manner. To the best of our knowledge, this is the first study to use a GNN model to score depression.

In addition, this work focused on the evaluation of the efficacy of the use of the graph-level embedding by comparing with the other three sequence-based deep learning models that work on the text-to-numeric word embeddings. The experimental results on the DAIC-WOZ benchmark demonstrate the effectiveness of the proposed approach and its superiority over other state-of-the-art methods.

The natural strength of deep graph learning models enables to provide a straightforward interface for producing structured representations. These graph representations are interpretable in the structure of word entities and relations. The work in chapter 5 introduces multiple ways of visualizing the learned hidden layers of deep graph learning models. This could further convert the underlying weakness of the "black box" mechanism within deep learning architectures to its strength by utilizing its much more implicit and flexible learning capability dealing with more challenging work.

Chapter 4

Predicting the Severity of Depression Using a Schema-based GNN

4.1 Motivation

PHQ-oriented concepts are key to identifying certain depressive symptoms. We know that the 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” (Kroenke et al., 2010). This PHQ-8 checklist given for depressive disorders can create a PHQ-oriented context of eight elements, we therefore consider these eight elements as depression concepts. These depression concepts can be primarily identified in the PHQ-centered context of a transcript, for example, a PHQ-based concept of “sleep issue” may be described with a word pair (“not”, “sleep”), and a PHQ-based concept of “fatigue issue” may be described with another word pair (“feel”, “tired”). This indicates that word-word associations can naturally capture contextual information relevant to concepts of PHQ.

I observed that each text (see Figure 4.1) covers information relating to at least one of these eight topics and word association might provide complementary information relevant to PHQ-centered concepts. The more frequent the occurrence of PHQ-related concepts, the greater the degree of the severity of depressive symptoms. Consequently, I hypothesize that the context of words in a transcript can be used to generate PHQ scores. This motivates generating a context at each word of a transcript and learning a graph representation of the transcript to generate a PHQ score.

I first propose a novel form for node attributes within a GNN-based model that captures node-specific embeddings for every word in the vocabulary. The representation of each word is shared globally and can be updated according to associations among words in a transcript. Intuitively, the generated node-level embeddings maintain records linking PHQ-based concepts as facts that indicate the known symptoms of a subject.

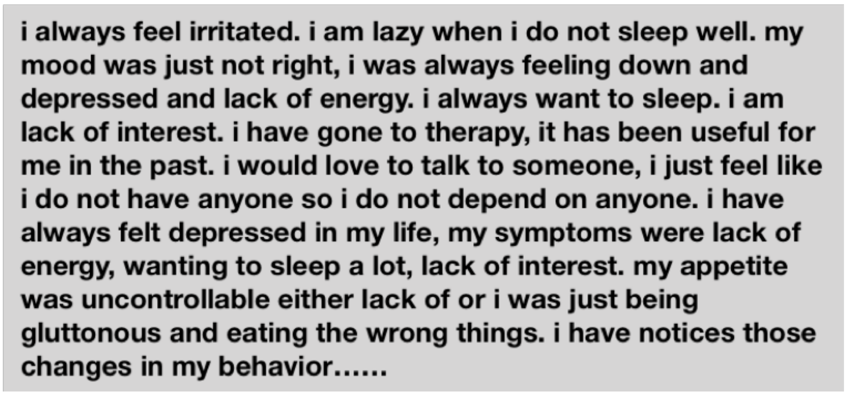
When subsequent facts are discovered which indicate additional depressive symptoms, records will be updated by aggregating both old and new facts. In other words, those records will be updated via a message passing mechanism over the transcript context until ideally all major depressive features (characterizing depressive symptoms) have been discovered.

I produce a graph-level embedding of a transcript by summarizing the representations of all the unique words in the transcript. Using graph structures to capture contextual features is an innovative pathway which I hypothesize can be used to measure different levels of depression.

In this chapter, the present research work demonstrates the efficacy of the proposed approach by showing that the accuracy of MDD measurement can be improved. Prior research has sought to make a diagnostic prediction of depression levels from patient data utilizing several modalities, including audio, video, and text. This research also demonstrates that the proposed method based on the text features achieves high performance with fewer parameters than methods based on multiple modalities. On the DAIC-WOZ benchmark, the proposed method outperforms the state-of-art methods by a substantial margin, including those using multiple modalities. It also outperforms the method presented in the previous chapter. Moreover, I evaluate the efficiency of 2-D node attributes and show that the proposed model outperforms a generic GNN model by leveraging 2-D node attributes.

4.2 Proposed Approach and Models

In this section I present a Schema-based Graph Neural Network (SGNN) in detail. Firstly I show how to build a graph and describe how to create 2D node attributes at each node of the graph using schema encoders. Secondly, I introduce how to use the message passing layer to



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 4.1: An extract from a raw transcript.

update node representations. Finally, I demonstrate how to extract node features to obtain a graph representation and learn a mapping between the graph embedding and a single value (the output is a PHQ score).

Building a Text Graph: For a given transcript, I build a text graph.

Let $G = (V, E)$ denote a text graph. V is a set of nodes representing all the unique words in a given transcript 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. V and E are defined as in Chapter 3.

Each node has an attribute which is a 2D array. To emphasise this representational structure of the attribute matrix, we refer to it as a 'schema' (Dozois & Beck, 2008; Hammen & Zupan, 1984; Rudolph et al., 1997; Soygüt & Savaşir, 2001) $U_i \in \mathbb{R}^{n \times d}$. 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.

4.2.1 Schema Encoders

I generate a schema for each word node which performs the 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 current training transcript.

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

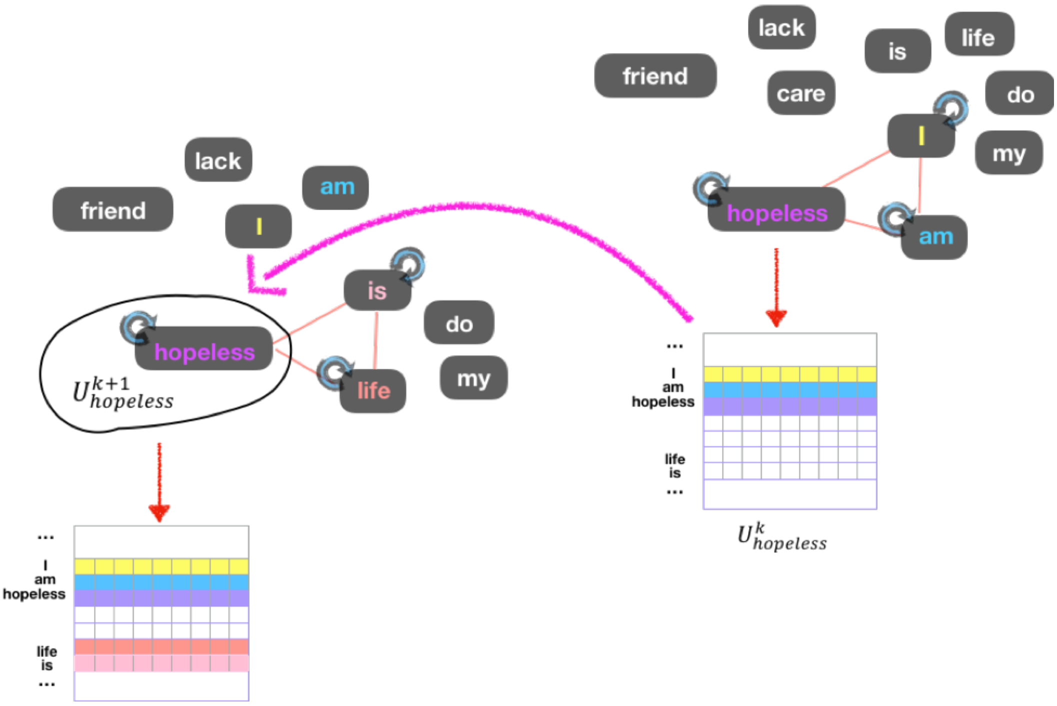


Figure 4.2: For one of the layers of an SGNN model k, the upper right part of the figure showing an output is a schema generated for the word node "hopeless". This schema can be treated as an "inner record" of the node "hopeless", which is formed as $U^k_{hopeless} \in n \times d_k$, where the generated d_k-dimensional representation of a node presents as a row. The blank rows correspond to word nodes that have not been encountered with the node labeled by "hopeless" in the current text context. The left part of the figure shows an example of a modified schema of the same word node "hopeless" after learning. The 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 the node labeled by "hopeless" is updated by the model. Moreover, its original record containing its other already existing neighbors has been explicitly preserved while learning. ***Note:*** For the convenience of display, in this figure, I set the window size of $+/-1$ for displaying associated edges among the nodes; in the actual experiments I use a larger window size.

4.2.2 Layer Initialization

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

I first initialize an $n \times d$ matrix as the schema $U^{(1)}_i$ at each node $v_i \in V$ using a linear transformation. The schema is all zeros apart from the row corresponding to the word associated with this node, which is a random d-dimensional vector.

4.2.3 Schema-based Message Passing Layer

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

I first modify the schema at each node v_i:

$$\hat{U_i}^k = U_i^k W_1^k + \frac{1}{n}\mathbf{1}\mathbf{1}^T U_i^k W_2^k + \frac{1}{n}\mathbf{1}_i\mathbf{1}^T U_i^k W_3^k \tag{4.1}$$

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, I compute the message function, which is defined as:

$$M^k\left(\hat{U_i}^k, \hat{U_j}^k\right) = \hat{U_j}^k + RELU\big([\hat{U_i}^k || \hat{U_j}^k] W_4^k\big) W_5^k \tag{4.2}$$

where $||$ denotes concatenation along the second axis. $\left(W_m^k\right)_{1 \leq m \leq 5}$ are learnable parameters. This is essentially a stack of two identical 1D convolution layers.

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

$$U_i^{k+1} = \sum_{j \in N(i)} M^k\left(\hat{U_i}^k, \hat{U_j}^k\right) \in \mathbb{R}^{n \times d_{k+1}} \tag{4.3}$$

I apply equation 4.3 as the sum aggregator over the k-th layer of the SGNN.

4.2.4 Pooling Layer

After all K message passing layers have been applied, I apply a max function to pool schema U_i^K in the word embedding space by taking the maximum across all rows of the schema:

$$h_{v_i} = max\big(U_i^K[i, :]\big) \tag{4.4}$$

As a result, the schema at each node by a row vector contains the maximum value in each column. Thus, the corresponding word is now represented by a row vector that captures the dominant values in the embedding spaces for all words in the vocabulary.

Next I apply a READOUT (Xu et al., 2018; Ying et al., 2018) function to capture graph-level features from row vectors. The READOUT (Xu et al., 2018; Ying et al., 2018) function aggregates node features by averaging them together:

$$h_G = \frac{1}{|V|} \sum_{v_i \in V} f(h_{v_i}^K) \tag{4.5}$$

where $f(.)$ is a final linear layer. The READOUT function performs a graph prediction task.

I use h_G to predict a PHQ score for each transcript. In the experiments, I apply a 2-layer MLP (1 hidden layer + 1 output layer) to generate a PHQ score. A ReLU layer and the following dropout layer are placed after the hidden layer of the MLP.

4.3 Evaluation

The results generated by the proposed method on the task of measuring the severity of depressive symptoms are shown in this section. I predict a PHQ score for each participant. The loss objective is RMSE for regression.

I evaluate the SGNN model on the DAIC-WOZ benchmark and I then compare the method to the state-of-art works including learning one single modality or multiple modalities on the same test set (i.e., the development set)

The model with randomly initialized word embeddings outperforms other works using pre-trained word embedding in the literature (Alhanai et al., 2018; Haque et al., 2018; Williamson et al., 2016). This shows that the approach is useful for text domains that may not have a large training corpus or pre-built dictionary.

I also investigate the effectiveness of pre-training for depression prediction. I employ GloVe word embeddings to generate 2-D node attributes. The SGNN model with pre-trained embeddings greatly improves the accuracy of assessing the severity of depression. I further investigate the effect of sliding window size on depression level prediction accuracy. I demonstrate the effectiveness of 2-D node attributes on minimizing errors in the prediction of depression severity and achieving a better generalization on a small group of subjects having PHQ scores higher than 14.

4.3.1 Experimental Setup

I set the learning rate as 10^{-3}, $L2$ weight decay as 10^{-4}, the dropout rate as 0.5, the window size as 4 to gather word-word occurrence statistics, $d_i = 32$ for all i, and the batch size as 7. The loss objective is RMSE. I trained the SGNN for a maximum of 300 epochs using the Stochastic Gradient Descent (SGD) optimizer (Kingma & Ba, 2014) and stopped training if the validation loss does not decrease for 10 consecutive epochs.

4.3.2 Experimental Performance & Analysis

I compare the method to prior work on measuring depressive symptom severity. The performance of the proposed method and eleven other methods, including the state-of-the-art method, is set out in Table 4.2. It is noted that results of some models are directly taken from their original papers. The method outperforms all other methods, despite using only the textual modality.

I found that the proposed model performs better than the standard GNN (Gilmer et al., 2017). The GNN without the schemas utilizes node representations which are initialized by pre-trained 300-dimensional GloVe word embeddings (Mikolov et al., 2013). However, the method constructs schemas which are initialized with random vectors. Moreover the standard GNN propagates node features in terms of vectors while the model uses 2-D node attributes. This change increases the expressive power of the MPM, resulting in better accuracy.

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

I also note that the model performs better than other prior work (Alhanai et al., 2018; Du et al., 2019; Haque et al., 2018; Lin et al., 2020; Ray et al., 2019; Song et al., 2018). 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 the model is trained to learn a graph-level embedding of each transcript. The method (Lin et al., 2020) based on the BiLSTM model reaches to a similar lower MAE. Their work employs pre-trained word2Vec embeddings. However, I train word embeddings initialized with random vectors.

4.3.2.1 Ablation Study

To further analyze the model, I perform three ablation studies (see Table 4.3). To give more reliable measures of performance, I perform a 10-fold cross-validation on the combined training and development dataset. I 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), I give equations 1-4 (in section 2.2) for 'point-wise' multiplication. According to the results in Table 4.3, we can see that using convolution layers can better model the relations between words compared with point-wise multiplication layers.

In (ii), I remove the second and third (equivariant) terms in (1). There was a significant reduction in performance (Table 4.3). From the ablation study, the representation power of the model can increase when parametrized with these layers, thereby improving the model's performance on the data set.

In (iii), the max operator, which operates as a max function (in section 4.2.3), is replaced by a mean operator to take an average over each row of a schema to obtain node features. From Table 4.3, we can see that the result was not good when applying an average operator. The max operator highlighting the strongest node features can enhance discriminating depressive features, which helps to achieve a better result.

4.3.3 Metaparameter Setting Analysis of SGNN

The hyperparameters are tuned to find an optimal setting of the SGNN model. The hyperparameters that were used in the model are shown in Table 4.1. I adopted 10-fold stratified cross-validation and reported MAE values. Here I used the same window size of 9, a dropout of 0.5. The model is initialized with 32-dimensional random word embeddings in the experiment.

4.3.4 The Effect of Window Size

In this experiment, I applied different sizes of a sliding window on each text document. A larger window size w captures long term dependency while a smaller p enforces the local dependency. I presented its impact in Table 4.4. Test accuracy on i)-iiii) expresses no clear patterns in model performance when different window sizes are employed. However, a much greater window has a negative impact on model performance. Since each patient's transcript is a long text, I finally

Table 4.1: Result of 10-fold cross-validation using different hyperparameters of the SGNN model. I report mean with standard deviation in the parentheses.

Regression: PHQ score				
Dim. of the Output of per Update Layer (node feature extractor)	Dim. of the Output of per Readout Layer (graph feature extractor)	# Params	MAE (SD) ↓	RMSE (SD) ↓
32	32	15745	3.84 (0.3482)	3.61(0.3106)
32	**16**	**13889**	**3.05(0.4565)**	**3.53(0.2311)**

Table 4.2: Comparison of machine learning approaches for measuring the severity of depressive symptoms on the DAIC-WOZ development set using MAE. The task evaluated is: PHQ score regression. Modalities: A: audio, V: visual, L: linguistic(text), A+V+L: combination. Note: the results marked with a * are taken from the cited papers.

Regression: PHQ score		
Methods	Modalities	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
*DCGAN (Yang et al., 2020)	A	4.63
*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
*BiLSTM (Lin et al., 2020)	L	3.88
*Multi-level Attention network (Ray et al., 2019)	L	4.37
* ELMo + BiLSTM (Lin et al., 2020)	L	5.44
* CNN + GNN (Chen et al., 2022)	V	3.96
the GNN (chapter 3)	L	4.24
Proposed Model	L	**3.54**

selected $w = 9$ as the relatively small sliding window for this domain task.

4.3.5 The Effectiveness of using pre-trained 2-D node attributes

I treat pre-trained word embeddings as node attributes x_i of node $v_i \in V$. I initialize $U_i^{(1)}$ at each node v_i by mapping the pre-trained word embeddings to rows corresponding to words associated with node v_i. I append these features to the same row of $U_i^{(1)}$: $U_i^{(1)}[i :,] = [1 : x_i]$ at first, then I apply a linear transformation to get an updated schema, which is a $n \times d_1$ matrix. I still set $d_1 = 32$ as same as the default size of schema.

I use the GloVe word embedding $50-, 100-, 200-$ dimensional training on the Wikipedia 2014+Gigaword 5 dataset. In Table 4.5, the results show equally high performances between

Table 4.3: Results of ablation studies. I apply 10-fold stratified cross-validation and give mean results with standard deviation in the parentheses.

Metric	MAE ↓		RMSE ↓	
Setting	Train	Test	Train	Test
Original(SGNN)	3.48(0.4521)	3.39(0.4260)	4.20(0.5430)	4.05(0.5491)
i) Fast(SGNN)	3.67(0.5222)	3.48(0.5273)	4.31(0.6036)	3.94(0.4151)
ii) Without equivariant linear layers (Maron et al., 2018)	3.60(0.6120)	4.40(0.6630)	4.16(0.5229)	4.01(0.5150)
iii) Mean Reduction	4.48(0.1595)	4.58(0.3739)	4.41(0.5541)	4.18(0.4147)

Table 4.4: Performance Effects of window size. The accuracy with multiple sizes of a sliding window is shown in the table. I adopt 10-fold stratified cross-validation and report mean with standard deviation in the parentheses.

	Train		Test	
Window Size	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
i) #W = 5	3.24(0.6041)	4.11(0.7579)	3.73(0.5034)	4.04(0.4301)
ii) #W = 7	3.72(0.5556)	4.75(0.5982)	3.63(0.4007)	4.35(0.4936)
iii)#**W = 9**	**3.32(0.3775)**	**4.18(0.4643)**	**3.19(0.4138)**	**3.75(0.6245)**
iiii)#W = 11	3.81(0.4631)	4.83(0.3211)	3.87(0.4572)	4.29(0.4043)

models utilizing three different word feature dimensions. I therefore choose 50-dimensional word embeddings for less computation.

I further evaluate the performance of the model with 50-dimensional word embeddings on the DAIC-WOZ development set; the result shown in Table 4.6 notes that the SGNN model with pre-trained word embeddings can effectively improve the accuracy with MAE and RMSE.

4.3.6 Performance Analysis of 2-dimensional Node Attributes

In Table 4.2, I compared the result of the performance of the GNN models with and without 2-D node embeddings. I argued that 2-D node embeddings generated by schema encoders obtained most informative information about the depressive state, particularly in the high-level depressive state. In this section, I analyze the model performance of a standard GNN model and the SGNN model using 2-D node attributes.

Figure 4.3 shows that the latter (in the middle and on the right of the figure) generalized the development set much better, particularly for the small group of samples in the class of having scores higher than 13. The graph representation learning framework using schemas achieves better generalization on the very limited dataset. I suggest that using 2-D node attributes

Table 4.5: Performance comparison using different word feature dimensions. I apply 10-fold stratified cross-validation and give mean (standard deviation) results.

Feature size	Train		Test	
	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
d = 50	3.12 (0.2255)	**3.36(0.2820)**	3.21 (0.1729)	3.36 (0.3368)
d = 100	2.68 (0.2144)	3.63 (0.6101)	2.61 (0.2485)	**3.08(0.4269)**
d = 200	**2.40(0.3007)**	3.90 (0.3640)	**2.37(0.6243)**	3.30 (0.2243)

Table 4.6: Results of experiments on the development set. Performances of SGNN model with and without pre-trained word embeddings.

Model	Test
	MAE ↓
Proposed Approach	3.54
Proposed Approach+50d Glove	**2.92**

(i.e., schemas) can improve the expressive power of the MPM. By setting a global context (matrices) of each node (as the input), I can obtain some parameters, in an explicit way, that capture the context of a word using each node's schema. Moreover, using pre-trained global word embeddings to learn 2-D node attributes can further improve the overall accuracy of the performance. On the contrary, using a 1D node embedding may result in losing contextual information when they are processed by several layers of a learning model.

Qualitative Analysis of Experimental Results: In figure 4.3, the SGNN model without pretrained embeddings (in the center) made more biased predictions on those subjects having PHQ scores in range [10, 15]: For instance, the model predicted a PHQ score of 3 for a patient with an actual PHQ score of 14, while the SGNN model which employed GloVe embeddings achieved a higher predictive score of 12 for the same patient. The original SGNN model predicted a PHQ score of 20 for a patient with an actual PHQ score of 15, while the SGNN model with GloVe embeddings achieved a predicted score of 14. Moreover, the patient who has the PHQ score of 23 has been underpredicted by both SGNN and SGNN+GloVe models. To improve the learning capacity of the model, additional training samples with PHQ scores higher than 20 would be needed.

Qualification Analysis of Word-Word Associations: I used a hidden layer output of the SGNN by equation 4.4 to qualitatively visualize word embeddings (= 32 dimensions) in the form

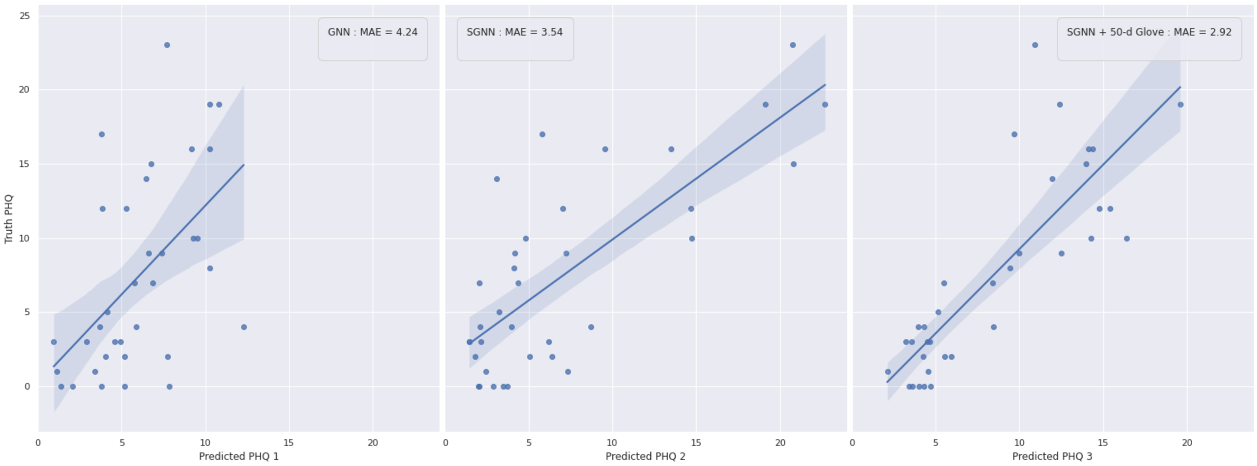


Figure 4.3: Results on the development set for the graph-based PHQ prediction system, with the true PHQ plotted as a function of predicted PHQ. The figure on the left describes the performance of applying a generic GNN model (Gilmer et al., 2017), while the figures in the middle and on the right respectively show the performance of implementing the SGNN model. It can be seen that GNN model cannot give a good generalization performance of a small group having scores greater than 13

.

of bi-grams. Table 4.7 shows the most important word pairs that have been learned to capture contextual semantic features, such as “feel, tough”, “therapy, asleep”, “sleeping, depression” or “depression, psychiatrist” related to PHQ topics. This demonstrates that high depression scores are predicted on the basis of appropriate semantics.

Table 4.7: Performance of top-10 bi-gram word associations on the DAIC-WOZ development set. The word-word connections, in the context of PHQ-related topics, are generated by the output of the final message passing layer of the SGNN model. gt: ground truth

Transcript 1	Transcript 2
gt score of 16 predicted score of 16.46	gt score of 19 predicted score of 17.30
(‘married’, ‘upset’)	(‘almost’, ‘thought’)
(‘getting’, ‘upset’)	(‘cheated’, ‘marry’)
(‘anyone’, ‘argued’)	(‘unconditional’, ‘trip’)
(‘family’, ‘issue’)	(‘depression’, ‘psychiatrist’)
(‘feel’, ‘tough’)	(‘certainly’, ‘argent’)
(‘energy’, ‘explore’)	(‘exhausted’, ‘never’)
(‘helping’, ‘sleep’)	(‘therapy’, ‘asleep’)
(‘lack’, ‘achieve’)	(‘development’, ‘issue’)
(‘missing’, ‘every’)	(‘married’, ‘upset’)
(‘sleeping’, ‘depression’)	(‘especially’, ‘breath’)

4.4 Conclusion

This chapter demonstrates a novel method to improve the performance of predicting depression states by training a deep graph learning model to learn contextual features from the text. The results have demonstrated that it is possible to apply deep learning methods to tackle more specific problems within the field, and not just a more general problem, even with limited data.

The proposed approach exploits the strong learning ability of schema encoders to capture important features from the clinical context of patients' verbal behaviors. The experiments demonstrate that good predictions can be made with little prior patient information based on their transcripts. The promising results of this chapter open the possibility for further applications in this or other domains.

Future work might address finding a way of explaining a GNN model such as visualizing the relationship between the underlying depression features and depression scores, which helps us better understand clinical context behind the data.

Chapter 5

Using the SGNN model to Interpret Fine-grained Prediction of Severity of Depression

5.1 Motivation

In chapter 4, we trained the SGNN model to learn a regression task that predicts a numerical rating scale representing the severity of depression. In this chapter, I investigate the learning process of the SGNN model. Since graph representation learning using graph structures naturally elucidates its predictive results, I provide a natural approach of interpretability to describe the model's predictions in terms of its gradients. I deliver human understandable post-hoc explanations of the model's decisions with semantic measures.

The following three ways of interpretability analysis for the depression level prediction task utilizing the output of an internal layer of the SGNN model are introduced:

1. I explore semantic measures to visualize patterns between depression levels using the output of the model's pooling layer.
2. I use the output of the model's final message passing layer to measure semantic similarity between words. I show by examples that the highlighted words generated by the model provide useful information.

3. I implement word cloud, a data visualization technique, to visualize the performance of the SGNN model. The output of the model's final message passing layer is applied to identify linguistic features of a transcript from a subject who may be suffering from depression over a time period.

My research demonstrates the explainability of the SGNN model with either embedded nodes (or words) as well as embed graphs (a graph represents a transcript), which provide linguistic insights into the degree of severity of subjects.

5.2 Semantic Network Relatedness

Depressed people recall bad or negative memories when they are exposed to information that is related to their knowledge (Dillon & Pizzagalli, 2018; Gorwood et al., 2008; MacQueen et al., 2002). The way their mind encodes knowledge is an associative semantic network (Bartczak & Bokus, 2017). An associative semantic network consists of a set of nodes and a set of edges. Each node in the network denotes a concept in semantic memory such as sleep, and edges representing associations between concepts can be used to indicate a semantic relation such as temporal co-occurrence, featural similarity, etc. I borrowed the concept of this associative semantic network to explore a network that identifies the relationship between different severity levels in depression. In my hypothesis, links between information in a semantic network can be determined by cognitive deficits in depression, such as cognitive schemas. As a result, connections within the semantic network are highly associative (Bartczak & Bokus, 2017).

My focus of interest is the relation between groups where individuals in each group have the same degree of severity of depression. My research question is how the levels of depression severity between groups can be indicated with semantic features extracted from their text. I exploit the structure of an associative semantic network to explore the difference in levels of severity of depression.

To achieve this, I treat an associative semantic network as a graph where the concept of a graph node represents an individual's transcript and the concept of the node features represents semantic features related to depression derived from the individual's transcript. The concept of a graph edge, as a feature-feature association, represents semantic similarities between individual nodes (or transcripts). I measure the influences of feature-feature relations in terms of semantic relatedness.

A semantic relatedness effect is an inverse correlation between semantic distance and relevance assessment (Budanitsky & Hirst, 2006; Miller & Charles, 1991; Weeds & Weir, 2005). I measure semantic relations between two nodes with semantic similarity metrics. A node pair has a strong association if their semantic similarity or relatedness is high, and otherwise they are "unrelated (or distant)". To estimate the scale of structure of an associative semantic network, semantic similarity metrics provide a quantitative way of measuring strengths of relationships between nodes in the network. The semantic relatedness features can be utilized to visualize the strengths of semantic networks, which capture a pattern in depression levels (from mild to severe).

Del and Fishel (2021) presents a model interpretation and analysis with cosine similarity to measure the similarity between the representations of sentences in different languages. I adopt this method to quantify the strength of an associative semantic network by measuring cosine similarity between a node pair:

$$\texttt{cosine\_similarity}(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|} \tag{5.1}$$

, where v_i and v_j are feature vectors of two distinct nodes (representing a pair of transcripts).

5.3 Semantic Measures Reflecting Patterns of Depression Levels

Since depression levels are indicated by scores 0–4 (no depression), 5–9 (mild), 10–14 (moderate), 15–19 (moderately severe) and 20–24(severe) as we discussed in section 1.3 of chapter 1, I applied four cut-points to partition the scales of depression into five groups. However, group five contains only a very small number of samples so I decided to combine group four and group five into one group. I partitioned the development dataset into four groups such that group one represents no depression (or the lowest level of depression) and group 4 represents the highest level of depression. Therefore I now have four new groups representing four distinct degrees of depression severity.

I applied t-SNE (van der Maaten & Hinton, 2008) to visualize high-dimensional embeddings of graphs (or transcripts). I quantified plots in the visualization (see figure 5.1) It can be shown in

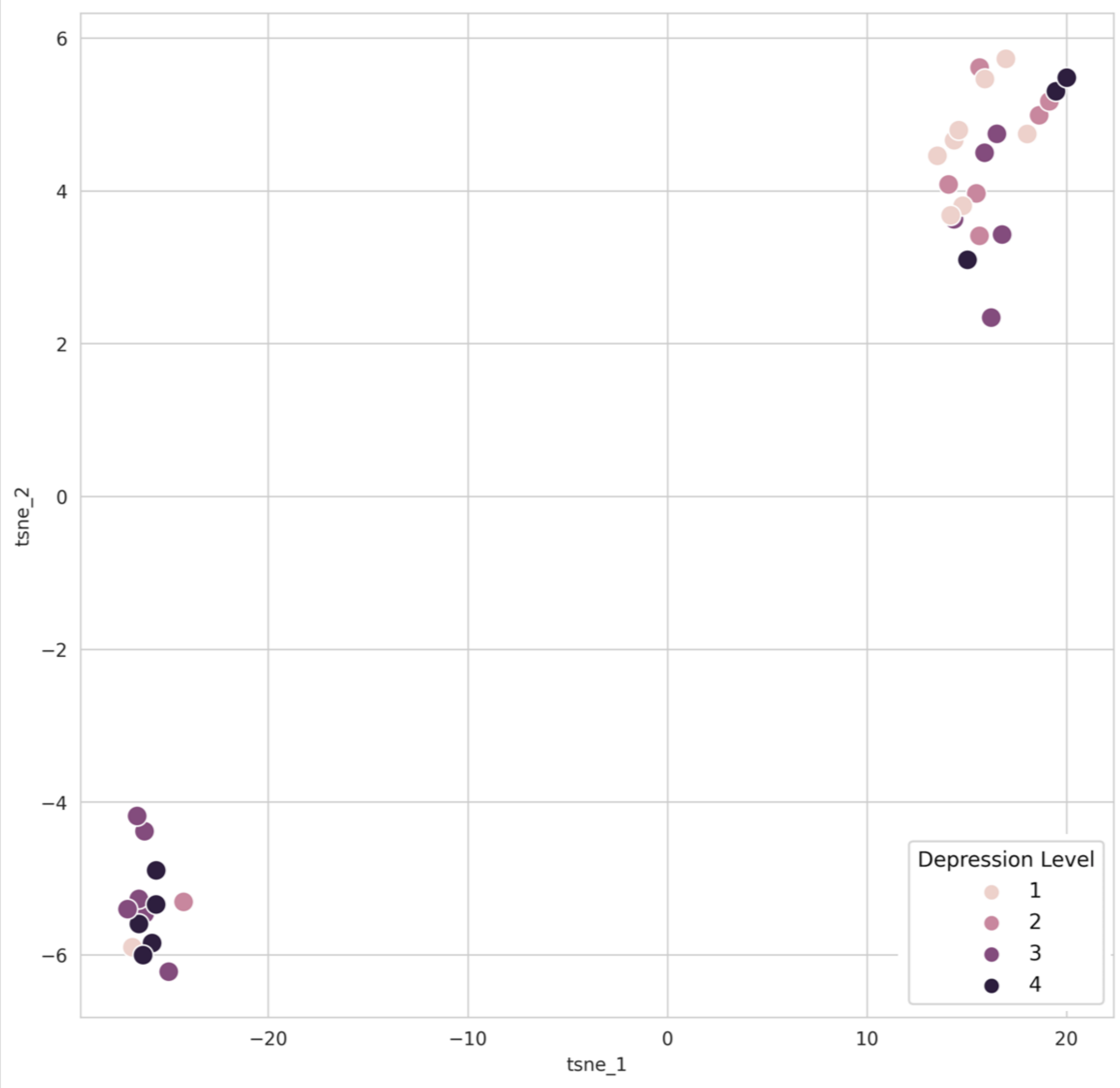


Figure 5.1: Visualizing semantic similarity of embeddings for transcripts in development set using t-SNE. The 16-dimensional embedding h_G is extracted from each transcript by equation 4.5.

figure 5.1 that graphs (in dark colors) with higher levels of depression (i.e. levels 3 and 4) are mostly clustered and distinct from graphs (in light colors) with lower levels of depression (i.e., level 1 and level 2). Furthermore, I also plotted node embeddings of transcripts with t-SNE. The visualization in figure 5.2 also illustrates correlations between relations of node embeddings and depression levels. I found that clusters of nodes (e.g., clusters coloured red) in the same depression level are not closely clustered, and some are even close to clusters of different levels.

To explore the reason why some transcript embeddings with high levels of depression are clustered together with low levels of depression shown in figure 5.3, I compute the similarity of two

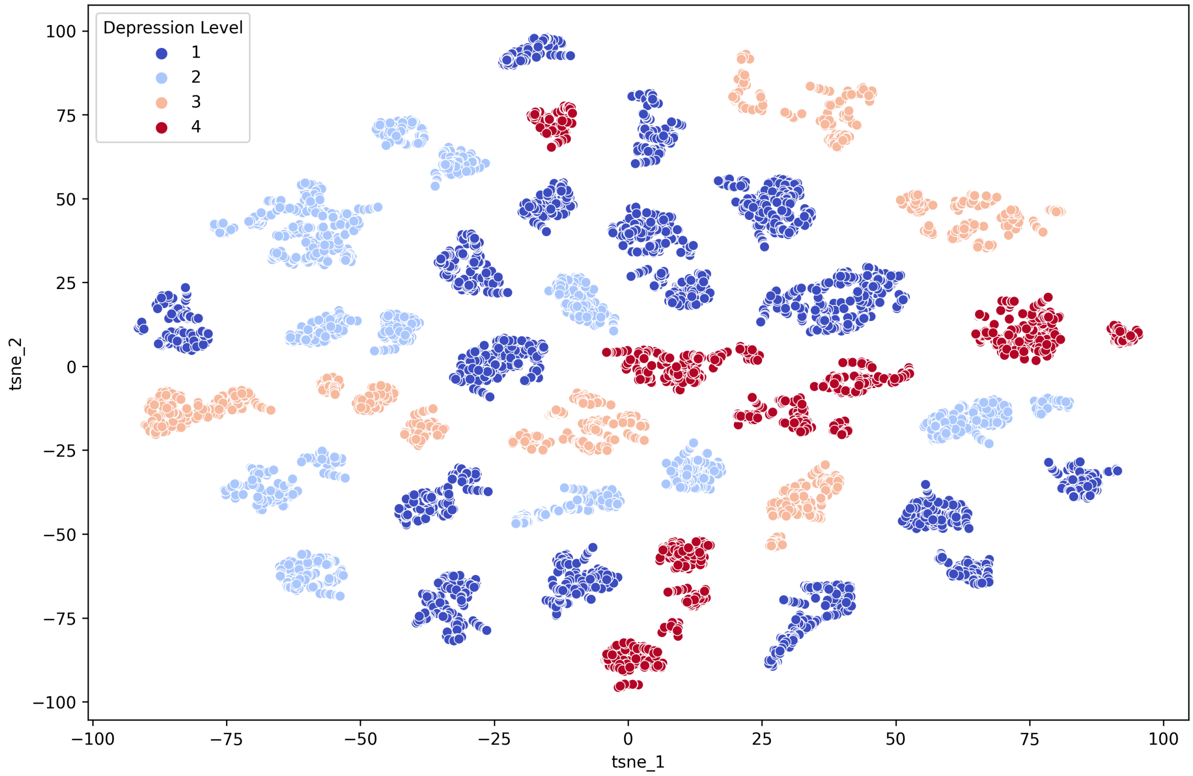


Figure 5.2: Visualizing semantic similarity of node embeddings of transcripts in development set using t-SNE. The 32-dimensional embedding h_{V_G} is extracted from each transcript by equation 4.4.

transcripts based on node embeddings with cosine similarity measure.

Node embeddings of a transcript i are the output of the 2^{nd} message passing layer by equation 4.4. $H_{v_{G_i}}$ is made by stacking row vectors h_{v_i}, which is 32-dimensional, for all $v_i \in V_{G_i}$.

The result of pairwise cosine similarity between node embeddings of transcript i and node embeddings of every other transcript, such that $cos\_sim(H_{v_{G_i}}, H_{v_{G_j}})$ for $v_{G_j} \in V_{\{G_1,\ldots,G_{N-1}\}}$ where N is the total number of transcripts, is concatenated to obtain a final embedding $H_{V_{G_i}} \in |V_{G_i}| \times d$.

t-SNE is applied to visualize the closeness of node representations (i.e., $H_{V_{G_i}}$) between four groups. Figure 5.3 shows that texts from various group levels have similar representations based on PHQ-related concepts, suggesting a shared context among the groups.

Furthermore, I quantified the semantic network relatedness between each pair of graphs in the high-dimensional representations. Graphs with the same level of depression are assigned to the same group. I treated each group as an associative semantic network. Graphs (i.e., transcripts) in the same group are network nodes and the features of each node are embeddings of each

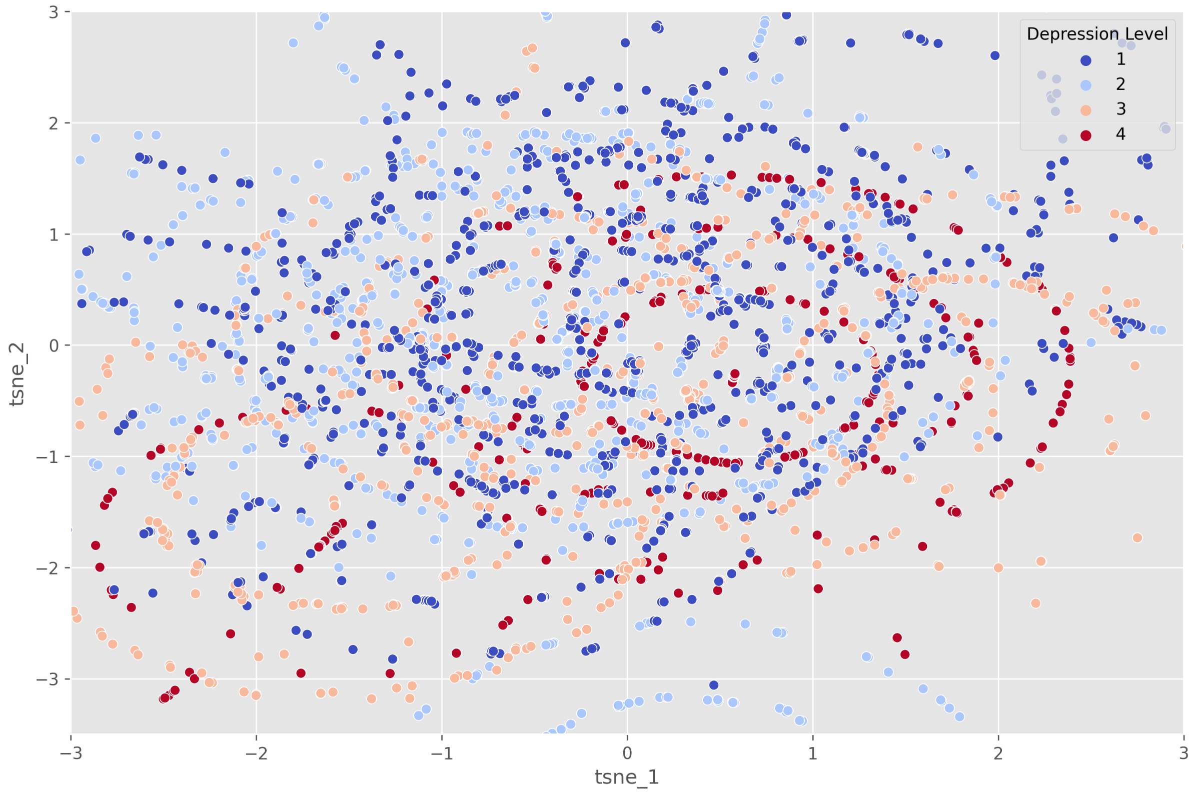


Figure 5.3: A bird's-eye view of semantic similarity of node embeddings of each transcript in development set via t-SNE.

graph. The learned features, for instance, h_{G_1} is the representation of the transcript 1, h_{G_2} is the representation of the transcript 2 and so on, of each transcript are extracted from the pooling layer of the SGNN model.

To measure the strength of each network, I calculated a relatedness score for a pair of nodes using the cosine similarity (as we introduced in section 5.2). For each pair of transcript i and j, I quantified the semantic network relatedness with a score c_{ij} : $c_{ij} = cos\_sim(h_{G_i}, h_{G_j})$.

I set the threshold value of 0.5 to filter out node pairs if their scores are low. A boxplot is applied to visualize the correlation between the quantity of relatedness and group categories.

Figure 5.4 shows that group four has the strongest semantic associative relations compared to the other three groups. It is noted that group one of subjects who have no or minor depressive symptoms is significantly distant from group four of subjects who suffer from severe depression. I argue that intergroup differences (or distances) shown in 5.4 reflect a pattern of depression levels.

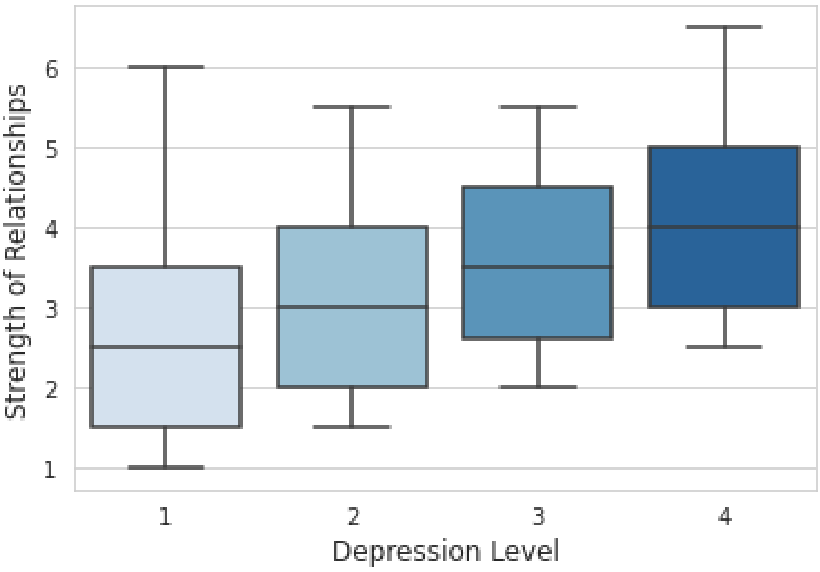


Figure 5.4: Quantifying semantic associative relations of each group with c_{ij}. The group distribution describes a certain trendline in depression levels on the development set: the median frequency indicates the strength of a network (of a group). The 16-dimensional embedding h_G is used.

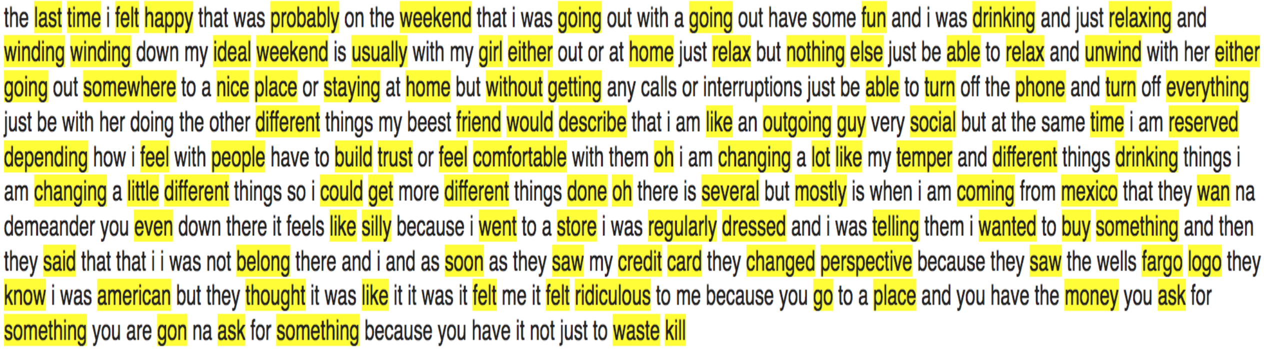
5.4 Illustration of Depression Level Prediction

Six examples are presented (see Figures 5.5,5.6,5.7,5.8,5.9,5.10) to reveal salient words selected by the model in a transcript. I applied features extracted from a hidden layer of the SGNN model for visualization.

I applied a cosine similarity function to calculate a similarity score between a pair of nodes using their representation features, such that $cos\_sim(h_{v_i}, h_{v_j})$. h_{v_i} and h_{v_j} which are a 32-dimensional embedding by equation 4.4 are used.

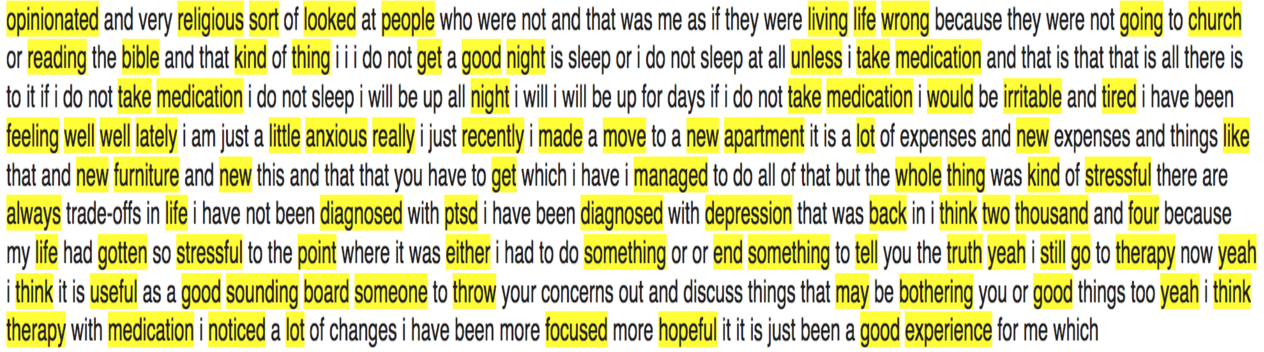
To measure the importance of a node such as v_i in a graph, I first computed all similarity scores between that node v_i and the rest of the nodes. After that, I averaged those similarities to obtain a final score for that node. A threshold of 0.996 is used to capture the most informative words that have scores higher than the threshold. According to the experimental results, applying a threshold of 0.996 captures an average of 22% of all words in each transcript, compared to applying threshold values of 0.997, 0.998, and 0.999, which capture an average of 22.05%, 21.83%, and 21.53% of words, respectively.

This method of utilizing word-word semantic associations allows us to visualize word indicators that contribute to the final decision. The high average similarity measure also does a good job of emphasizing words that are meaningful in the context of depression.



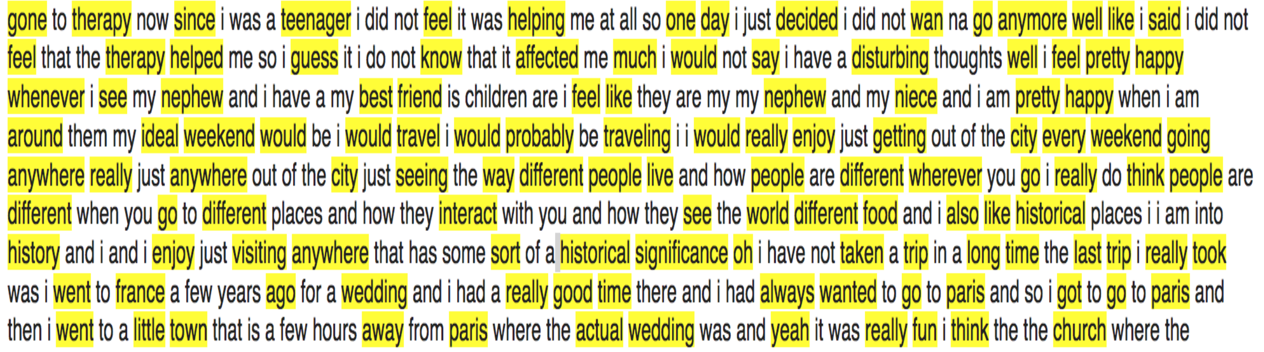
the last time i felt happy that was probably on the weekend that i was going out with a going out have some fun and i was drinking and just relaxing and winding winding down my ideal weekend is usually with my girl either out or at home just relax but nothing else just be able to relax and unwind with her either going out somewhere to a nice place or staying at home but without getting any calls or interruptions just be able to turn off the phone and turn off everything just be with her doing the other different things my beest friend would describe that i am like an outgoing guy very social but at the same time i am reserved depending how i feel with people have to build trust or feel comfortable with them oh i am changing a lot like my temper and different things drinking things i am changing a little different things so i could get more different things done oh there is several but mostly is when i am coming from mexico that they wan na demeander you even down there it feels like silly because i went to a store i was regularly dressed and i was telling them i wanted to buy something and then they said that that i i was not belong there and i and as soon as they saw my credit card they changed perspective because they saw the wells fargo logo they know i was american but they thought it was like it it was it felt me it felt ridiculous to me because you go to a place and you have the money you ask for something you are gon na ask for something because you have it not just to waste kill

Figure 5.5: an example of extraction of patient transcript with a ground truth PHQ score of 0



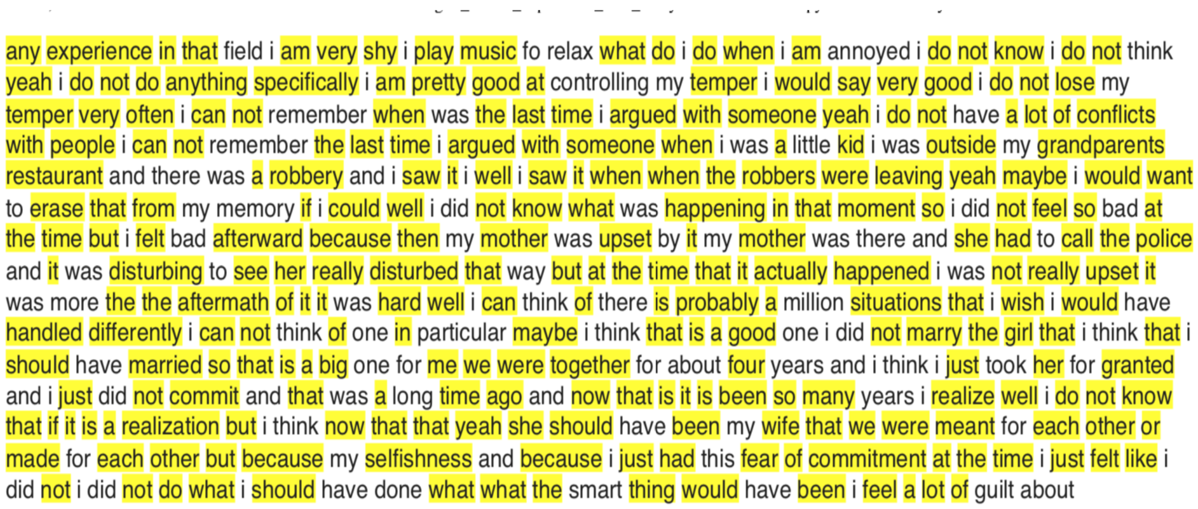
opinionated and very religious sort of looked at people who were not and that was me as if they were living life wrong because they were not going to church or reading the bible and that kind of thing i i i do not get a good night is sleep or i do not sleep at all unless i take medication and that is that that is all there is to it if i do not take medication i do not sleep i will be up all night i will i will be up for days if i do not take medication i would be irritable and tired i have been feeling well well lately i am just a little anxious really i just recently i made a move to a new apartment it is a lot of expenses and new expenses and things like that and new furniture and new this and that that you have to get which i have i managed to do all of that but the whole thing was kind of stressful there are always trade-offs in life i have not been diagnosed with ptsd i have been diagnosed with depression that was back in i think two thousand and four because my life had gotten so stressful to the point where it was either i had to do something or or end something to tell you the truth yeah i still go to therapy now yeah i think it is useful as a good sounding board someone to throw your concerns out and discuss things that may be bothering you or good things too yeah i think therapy with medication i noticed a lot of changes i have been more focused more hopeful it it is just been a good experience for me which

Figure 5.6: an example of extraction of patient transcript with a ground truth PHQ score of 10



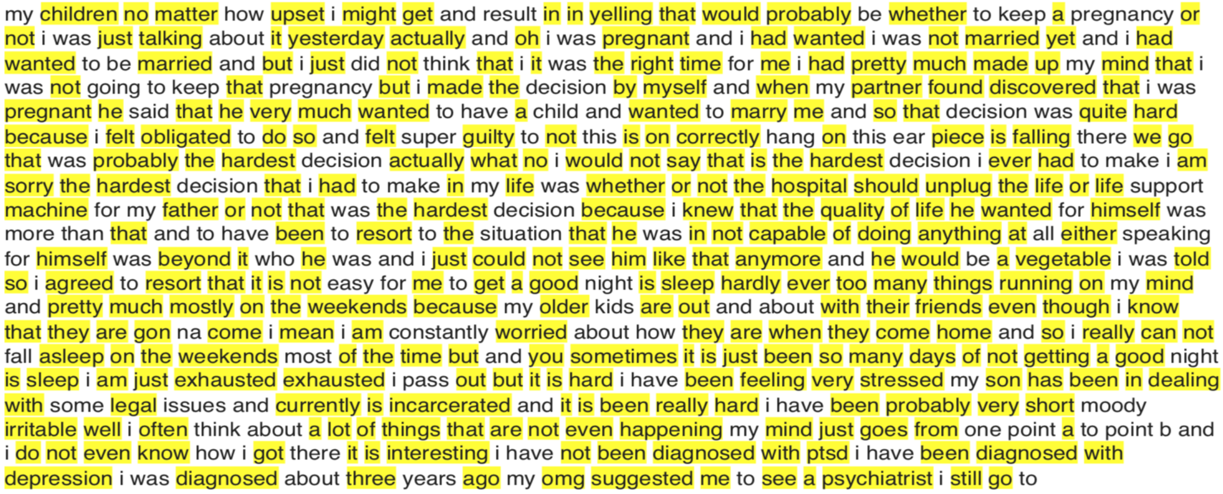
gone to therapy now since i was a teenager i did not feel it was helping me at all so one day i just decided i did not wan na go anymore well like i said i did not feel that the therapy helped me so i guess it i do not know that it affected me much i would not say i have a disturbing thoughts well i feel pretty happy whenever i see my nephew and i have a my best friend is children are i feel like they are my my nephew and my niece and i am pretty happy when i am around them my ideal weekend would be i would travel i would probably be traveling i i would really enjoy just getting out of the city every weekend going anywhere really just anywhere out of the city just seeing the way different people live and how people are different wherever you go i really do think people are different when you go to different places and how they interact with you and how they see the world different food and i also like historical places i i am into history and i and i enjoy just visiting anywhere that has some sort of a historical significance oh i have not taken a trip in a long time the last trip i really took was i went to france a few years ago for a wedding and i had a really good time there and i had always wanted to go to paris and so i got to go to paris and then i went to a little town that is a few hours away from paris where the actual wedding was and yeah it was really fun i think the the church where the

Figure 5.7: an example of extraction of patient transcript with a ground truth PHQ score of 16



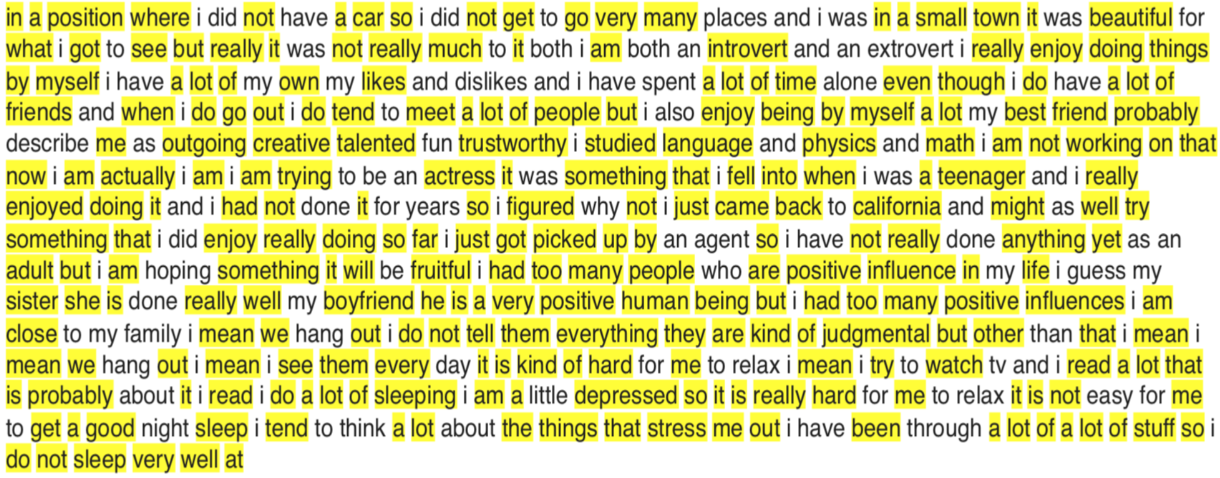
any experience in that field i am very shy i play music fo relax what do i do when i am annoyed i do not know i do not think yeah i do not do anything specifically i am pretty good at controlling my temper i would say very good i do not lose my temper very often i can not remember when was the last time i argued with someone yeah i do not have a lot of conflicts with people i can not remember the last time i argued with someone when i was a little kid i was outside my grandparents restaurant and there was a robbery and i saw it i well i saw it when when the robbers were leaving yeah maybe i would want to erase that from my memory if i could well i did not know what was happening in that moment so i did not feel so bad at the time but i felt bad afterward because then my mother was upset by it my mother was there and she had to call the police and it was disturbing to see her really disturbed that way but at the time that it actually happened i was not really upset it was more the the aftermath of it it was hard well i can think of there is probably a million situations that i wish i would have handled differently i can not think of one in particular maybe i think that is a good one i did not marry the girl that i think that i should have married so that is a big one for me we were together for about four years and i think i just took her for granted and i just did not commit and that was a long time ago and now that is it is been so many years i realize well i do not know that if it is a realization but i think now that that yeah she should have been my wife that we were meant for each other or made for each other but because my selfishness and because i just had this fear of commitment at the time i just felt like i did not i did not do what i should have done what what the smart thing would have been i feel a lot of guilt about

Figure 5.8: an example of extraction of patient transcript with a ground truth PHQ score of 16



my children no matter how upset i might get and result in in yelling that would probably be whether to keep a pregnancy or not i was just talking about it yesterday actually and oh i was pregnant and i had wanted i was not married yet and i had wanted to be married and but i just did not think that i it was the right time for me i had pretty much made up my mind that i was not going to keep that pregnancy but i made the decision by myself and when my partner found discovered that i was pregnant he said that he very much wanted to have a child and wanted to marry me and so that decision was quite hard because i felt obligated to do so and felt super guilty to not this is on correctly hang on this ear piece is falling there we go that was probably the hardest decision actually what no i would not say that is the hardest decision i ever had to make i am sorry the hardest decision that i had to make in my life was whether or not the hospital should unplug the life or life support machine for my father or not that was the hardest decision because i knew that the quality of life he wanted for himself was more than that and to have been to resort to the situation that he was in not capable of doing anything at all either speaking for himself was beyond it who he was and i just could not see him like that anymore and he would be a vegetable i was told so i agreed to resort that it is not easy for me to get a good night is sleep hardly ever too many things running on my mind and pretty much mostly on the weekends because my older kids are out and about with their friends even though i know that they are gon na come i mean i am constantly worried about how they are when they come home and so i really can not fall asleep on the weekends most of the time but and you sometimes it is just been so many days of not getting a good night is sleep i am just exhausted exhausted i pass out but it is hard i have been feeling very stressed my son has been in dealing with some legal issues and currently is incarcerated and it is been really hard i have been probably very short moody irritable well i often think about a lot of things that are not even happening my mind just goes from one point a to point b and i do not even know how i got there it is interesting i have not been diagnosed with ptsd i have been diagnosed with depression i was diagnosed about three years ago my omg suggested me to see a psychiatrist i still go to

Figure 5.9: an example of extraction of patient transcript with a ground truth PHQ score of 19



in a position where i did not have a car so i did not get to go very many places and i was in a small town it was beautiful for what i got to see but really it was not really much to it both i am both an introvert and an extrovert i really enjoy doing things by myself i have a lot of my own my likes and dislikes and i have spent a lot of time alone even though i do have a lot of friends and when i do go out i do tend to meet a lot of people but i also enjoy being by myself a lot my best friend probably describe me as outgoing creative talented fun trustworthy i studied language and physics and math i am not working on that now i am actually i am i am trying to be an actress it was something that i fell into when i was a teenager and i really enjoyed doing it and i had not done it for years so i figured why not i just came back to california and might as well try something that i did enjoy really doing so far i just got picked up by an agent so i have not really done anything yet as an adult but i am hoping something it will be fruitful i had too many people who are positive influence in my life i guess my sister she is done really well my boyfriend he is a very positive human being but i had too many positive influences i am close to my family i mean we hang out i do not tell them everything they are kind of judgmental but other than that i mean i mean we hang out i mean i see them every day it is kind of hard for me to relax i mean i try to watch tv and i read a lot that is probably about it i read i do a lot of sleeping i am a little depressed so it is really hard for me to relax it is not easy for me to get a good night sleep i tend to think a lot about the things that stress me out i have been through a lot of a lot of stuff so i do not sleep very well at

Figure 5.10: an example of extraction of patient transcript with a ground truth PHQ score of 23

Those figures show that informative keywords that are semantically relevant to diagnostic categories of PHQ-8 are highlighted for prediction. For instance, in figure 5.10 the highlighted words “i”, “do”, “not”, “sleep” and “well” are highly informative to identify a “sleep problem”, one of PHQ-8 topics. The visualization of highlighted words indicates that the model captured information associated with depressive features, and took them into account when making its final decision.

5.5 Visualization of Linguistic Features

Horizontal bar charts have been applied to qualitatively visualize word embeddings in terms of the 20 top keywords. The output of the hidden layer of the SGNN model as described by equation 4.4 is used.

I used the method described in section 5.4 to compute cosine scores for each word and thus to compute the 20 most informative words. For comparison, I also computed the 20 most common words based on their frequency. Each of the figures 5.11, 5.12, 5.13, 5.14, 5.15, 5.16, 5.17, 5.18, 5.19, 5.20, 5.21, 5.22, 5.23 and 5.24 presents the 20 most informative keywords for each text in development set. A horizontal bar plot shows the 20 words in a transcript, with the length of each bar representing the total number of times the word appears. The longer the bar, the higher the frequency of appearance. Based on the figures, the low frequency of keywords does not affect the decisions made by the model. In fact, the model is able to successfully capture PHQ-related context-level keywords, even if they appear less frequently in a transcript.

Moreover, bar graphs, for example figures 5.21, 5.22, 5.23 and 5.24, can explain the model’s incorrect predictions – the model generated high scores but not close to the actual scores. Figures show that the model successfully captures global keywords from different texts, including “ptsd”, “therapy”, “tired”, “bad”, “annoyed” and “stress”, that are highly indicative of certain PHQ-8 topics, such as ”depressed mood,” ”feelings,” and ”mental health impairments.” This demonstrates that high depression scores are predicted on the basis of appropriate semantics. The bar plot graphs demonstrate that the SGNN model can distinguish between semantic features in the text that are related to depressive symptoms and those that are not.

5.6 Conclusion

I showed multiple visualizations illustrating how the model made its prediction with learned graph features and word features. I quantified the strength of associative semantic networks with overall semantic relatedness. This approach provides a quantitative framework for showing the correlation between semantic closeness between networks and their nodes. Each network represents a category of depression levels. I found a pattern reflecting depression levels in categories.

Furthermore, I also applied semantic measures to highlight the words that are most informative from the model's perspective. In this way, I provided an alternative human understandable explanation for the predictive results by taking the advantage of context-aware word features. For instance, depression level reveals depressing words revealing depressive features in the text. The presence of the most common keywords indicates that the model has primarily focused on them, which helps us understand its learning process to some extent

In summary, I presented a way to interpret gradients of the SGNN model using semantic measures. I provided several intuitive visualizations measuring depression levels in text in a manner that may also be applied to different clinical domain tasks.

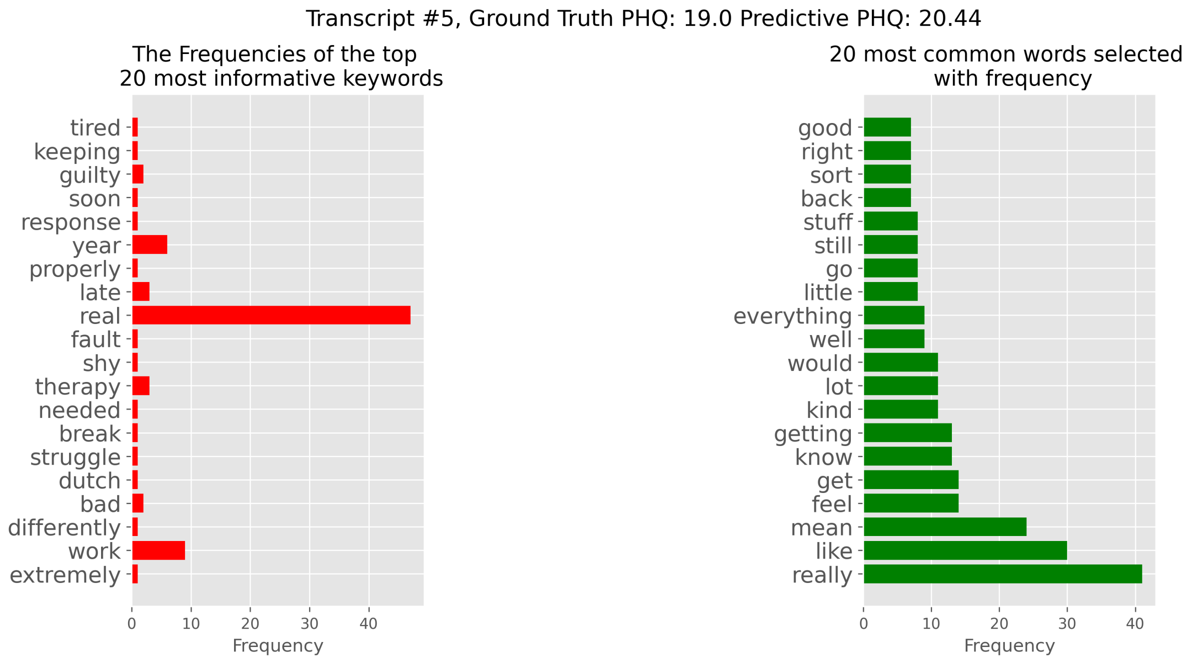


Figure 5.11: The bar graph on the left side shows that the 20 most informative keywords (e.g., 'tired', 'guilty', 'shy', 'therapy', 'struggle', 'bad', 'work', and 'extremely'), selected by the SGNN model, are relevant to PHQ topics, indicating the score of 20.44 is estimated on the basis of appropriate semantics. The bar graph on the right side displays 20 most common words selected based on word frequency.

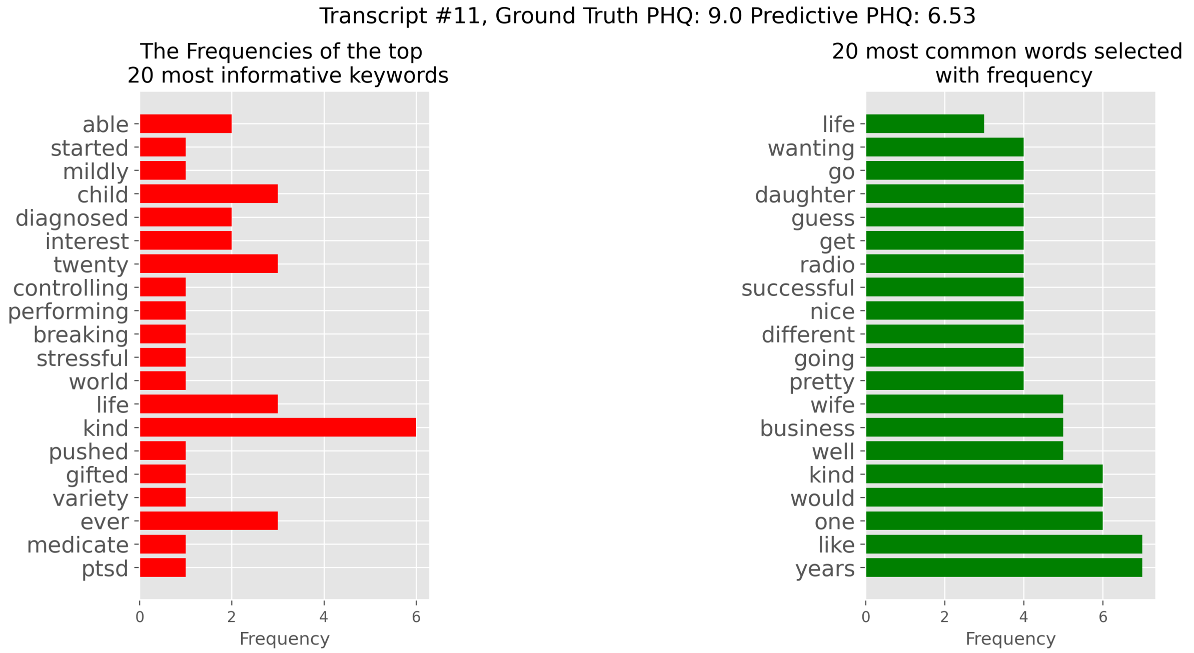


Figure 5.12: The bar graph on the left side shows that the 20 most informative keywords (e.g., 'diagnosed', 'interest', 'child', 'stressful', 'life', 'pushed', 'ptsd', and 'medicate'), selected by the SGNN model, are strongly related to PHQ topics. The bar graph on the right side displays 20 most common words selected based on word frequency.

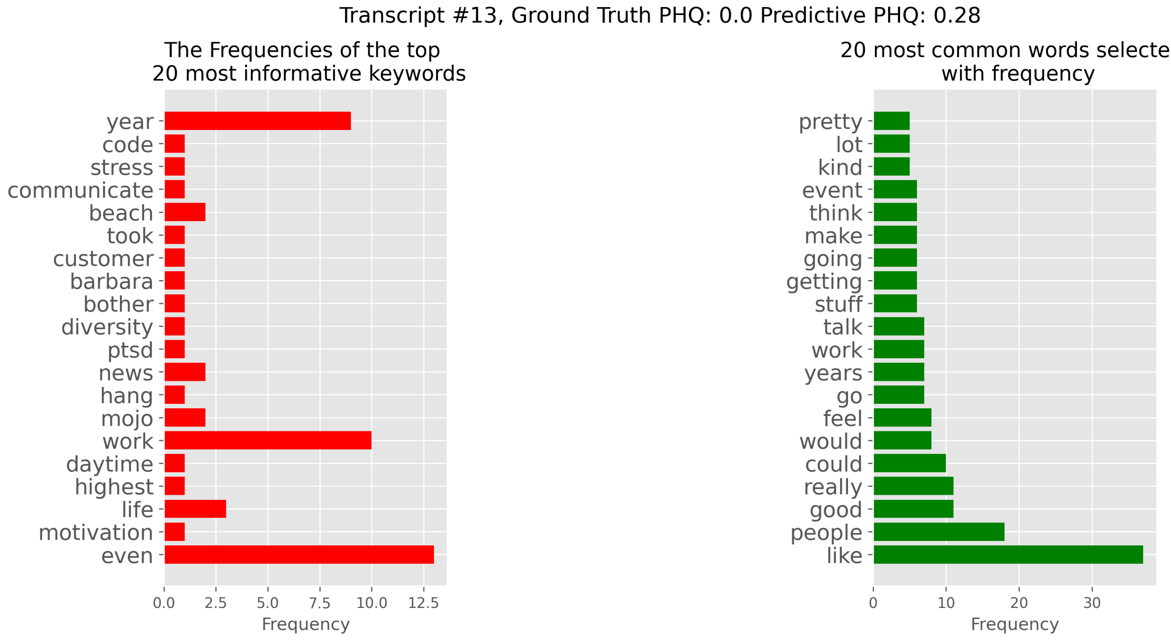


Figure 5.13: Horizontal bar chart of 20 most common words. The bar graph on the left side shows that the 20 most informative keywords (e.g., 'stress', 'communicate', 'ptsd', 'work', 'daytime', 'life', and 'motivation'), selected by the SGNN model, are irrelevant to PHQ topics, indicating the score of 2.59 is estimated on the basis of appropriate semantics. The bar graph on the right side displays 20 most common words selected based on word frequency.

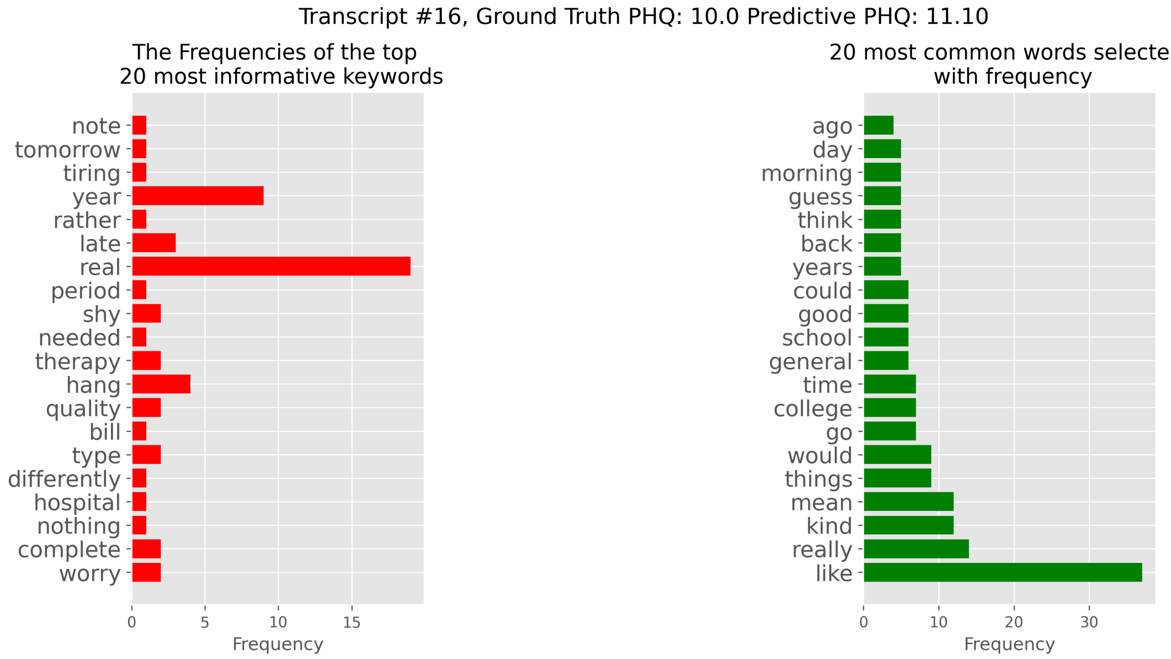


Figure 5.14: Horizontal bar chart of 20 most common words. The bar graph on the left side shows that the 20 most informative keywords (e.g., 'tiring', 'shy', 'therapy', 'hospital', 'nothing' and 'worry'), selected by the SGNN model, are strongly related to PHQ topics. The bar graph on the right side displays 20 most common words selected based on word frequency.

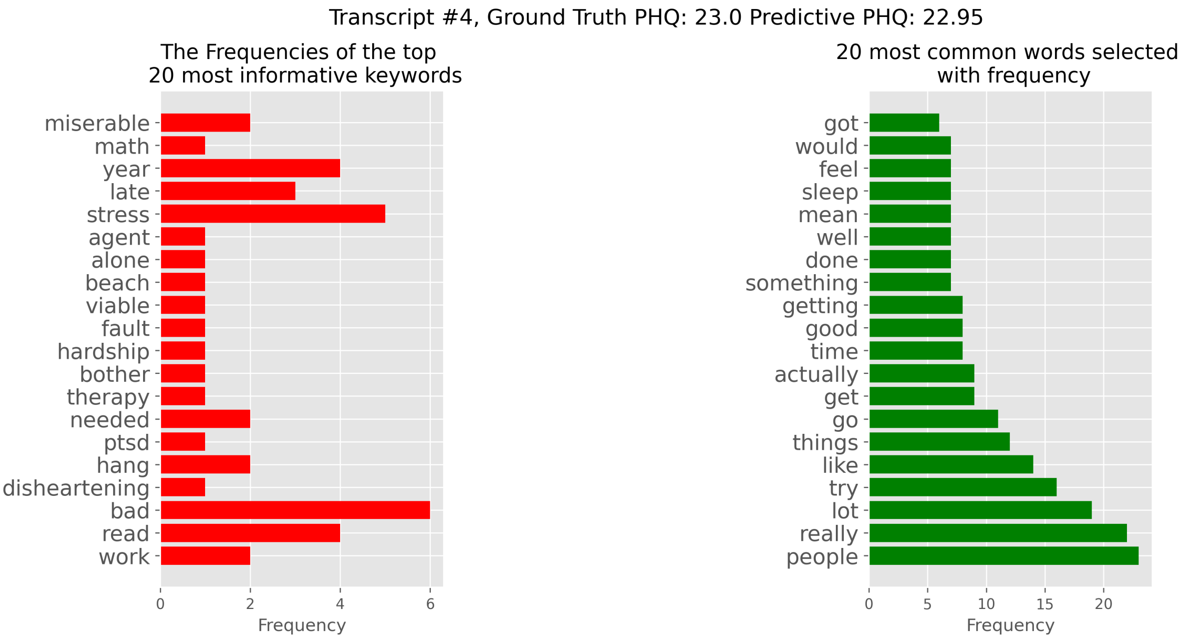


Figure 5.15: Horizontal bar chart of 20 most common words. The bar graph on the left side shows that the 20 most informative keywords (e.g., 'miserable', 'stress', 'alone', 'work', 'hardship', 'ptsd', 'bother', 'bad', 'therapy' and 'disheartening'), selected by the SGNN model, are related to PHQ topics. The bar graph on the right side displays 20 most common words selected based on word frequency.

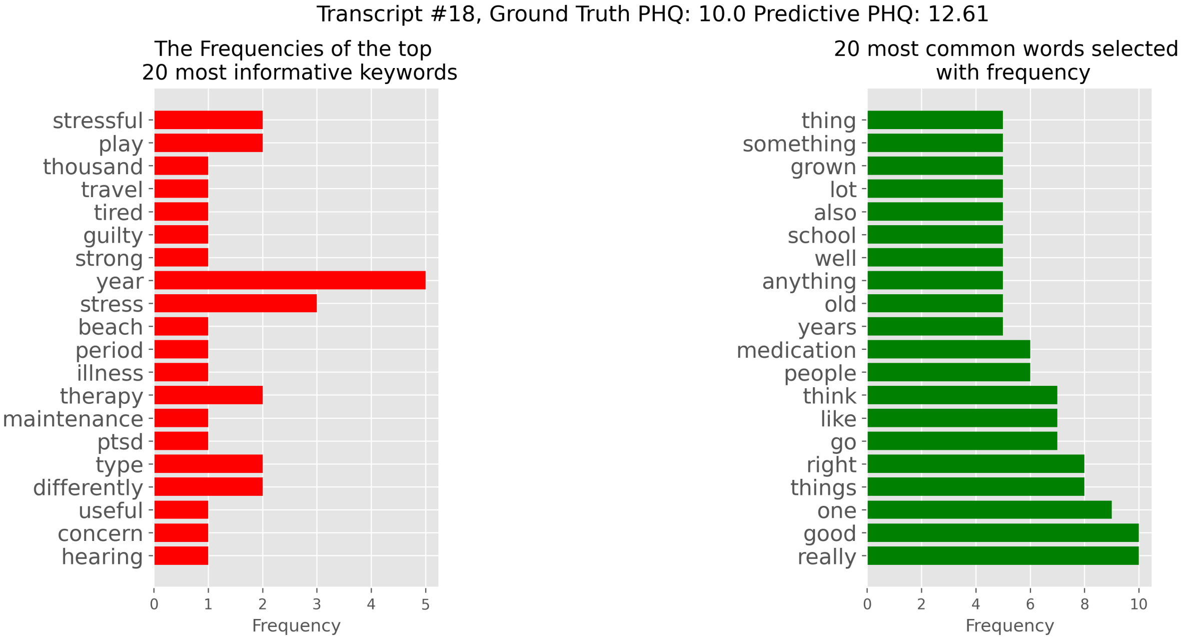


Figure 5.16: Horizontal bar chart of 20 most common words. The bar graph on the left side shows that the 20 most informative keywords (e.g., 'stressful', 'travel', 'tired', 'stress', 'illness', 'therapy', 'ptsd' and 'concern'), selected by the SGNN model, are strongly related to PHQ topics. The bar graph on the right side displays 20 most common words selected based on word frequency.

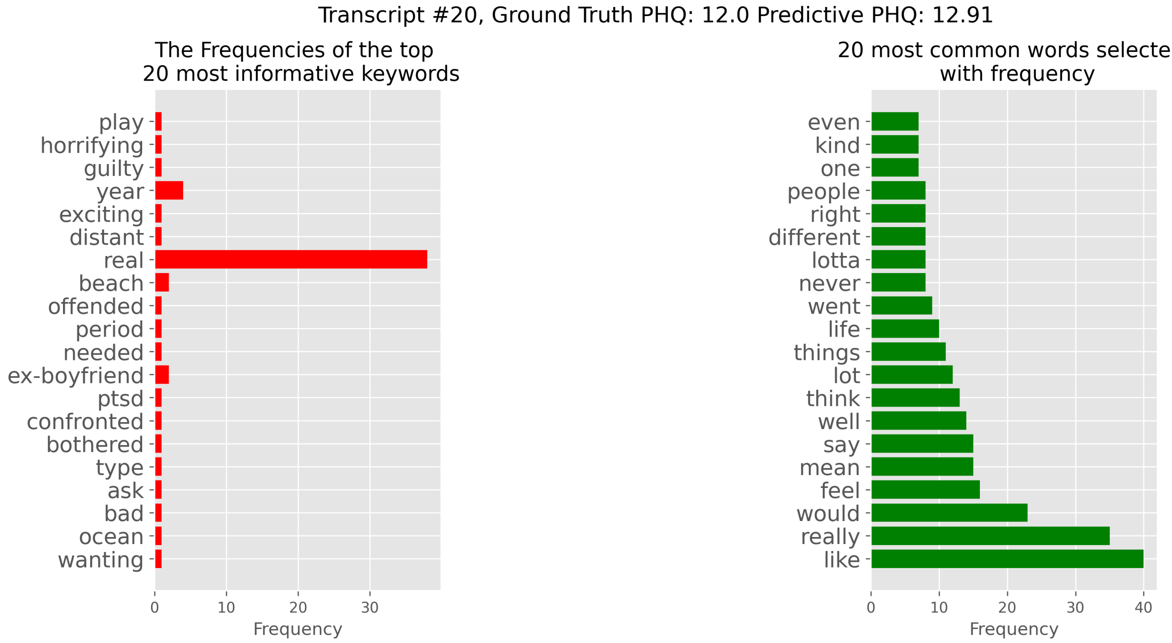


Figure 5.17: Horizontal bar chart of 20 most common words. The bar graph on the left side shows that the 20 most informative keywords (e.g., 'ptsd', 'bothered', 'bad', 'horrifying', 'guilty', 'exciting', 'offended', 'ex-boyfriend', and 'confronted'), selected by the SGNN model, are related to PHQ topics. The bar graph on the right side displays 20 most common words selected based on word frequency.

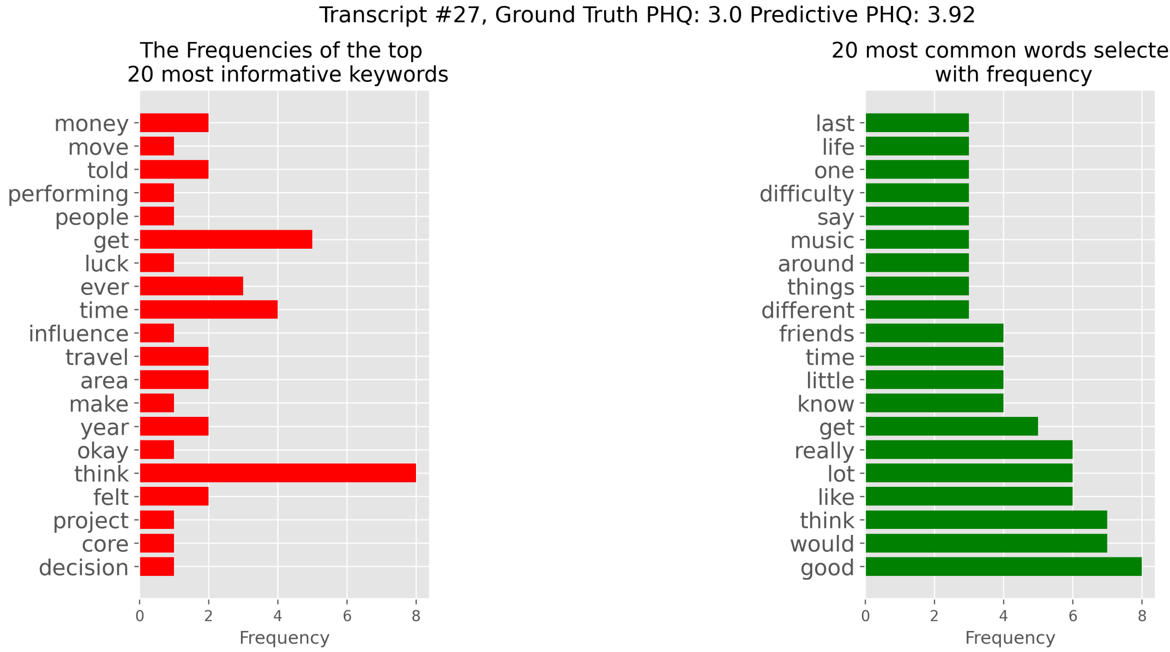


Figure 5.18: Horizontal bar chart of 20 most common words. The bar graph on the left side shows that the 20 most informative keywords (e.g., 'move', 'people', 'luck', 'travel', and 'project'), selected by the SGNN model, are less relevant to PHQ topics. The bar graph on the right side displays 20 most common words selected based on word frequency.

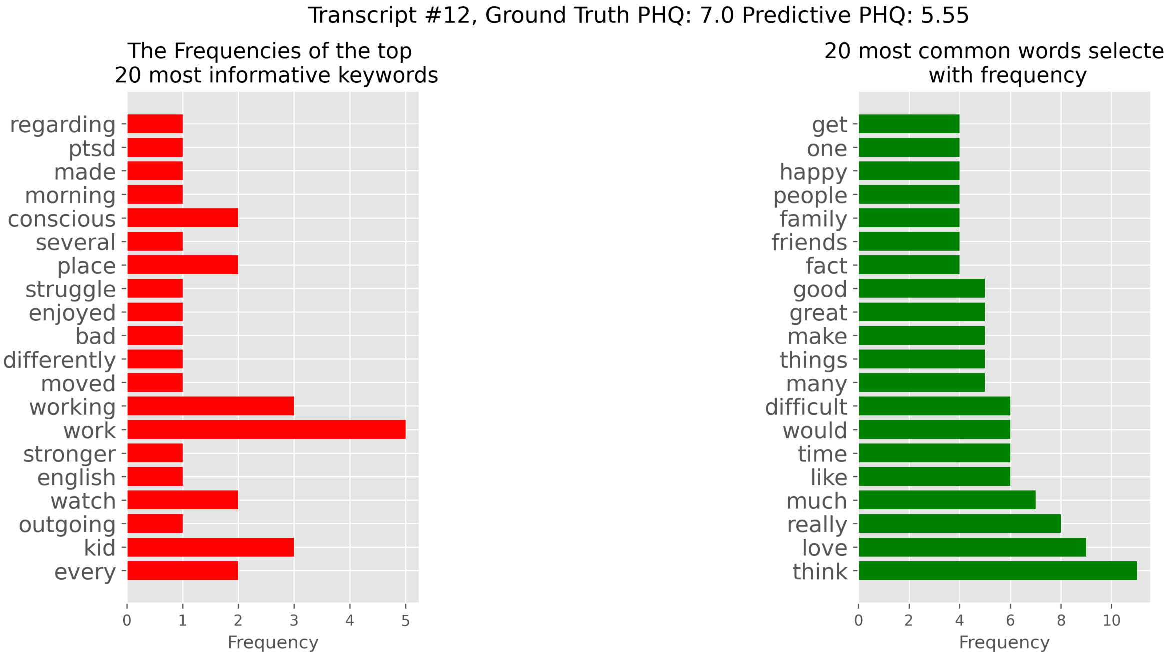


Figure 5.19: Horizontal bar chart of 20 most common words. The bar graph on the left side shows that the 20 most informative keywords (e.g., 'ptsd', 'struggle', 'bad', 'enjoyed', 'work', 'kid', 'stronger', and 'outgoing'), selected by the SGNN model, are related to PHQ topics. The bar graph on the right side displays 20 most common words selected based on word frequency.

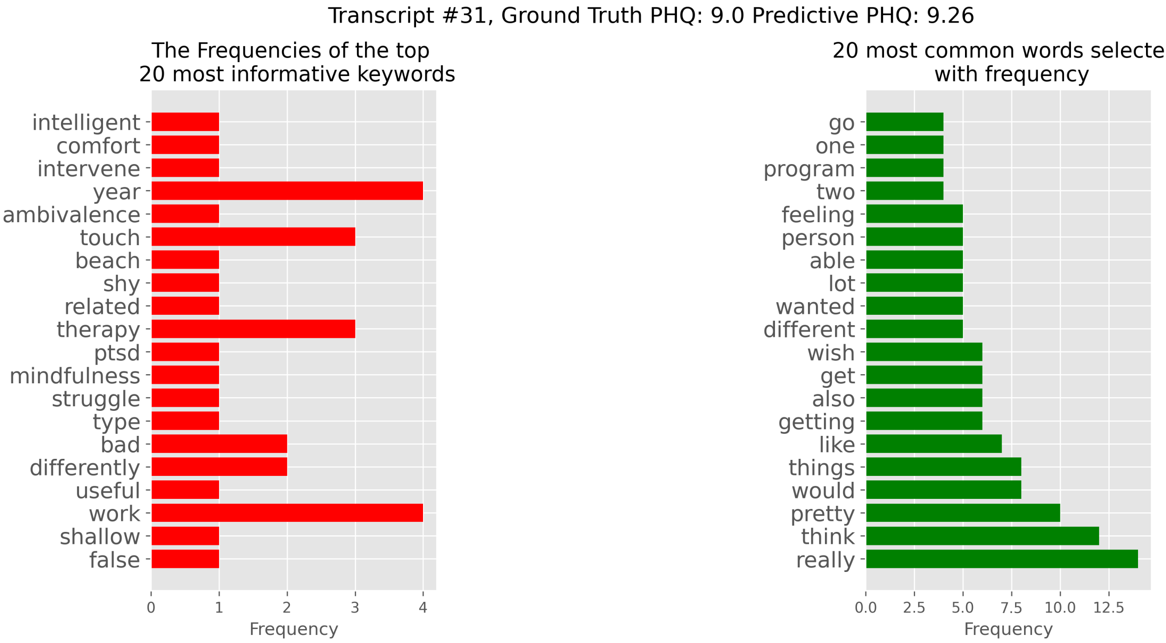


Figure 5.20: Horizontal bar chart of 20 most common words. The bar graph on the left side shows that the 20 most informative keywords (e.g., 'intervene', 'ambivalence', 'shy', 'therapy', 'ptsd', 'struggle', 'bad' and 'false'), selected by the SGNN model, are related to PHQ topics. The bar graph on the right side displays 20 most common words selected based on word frequency.

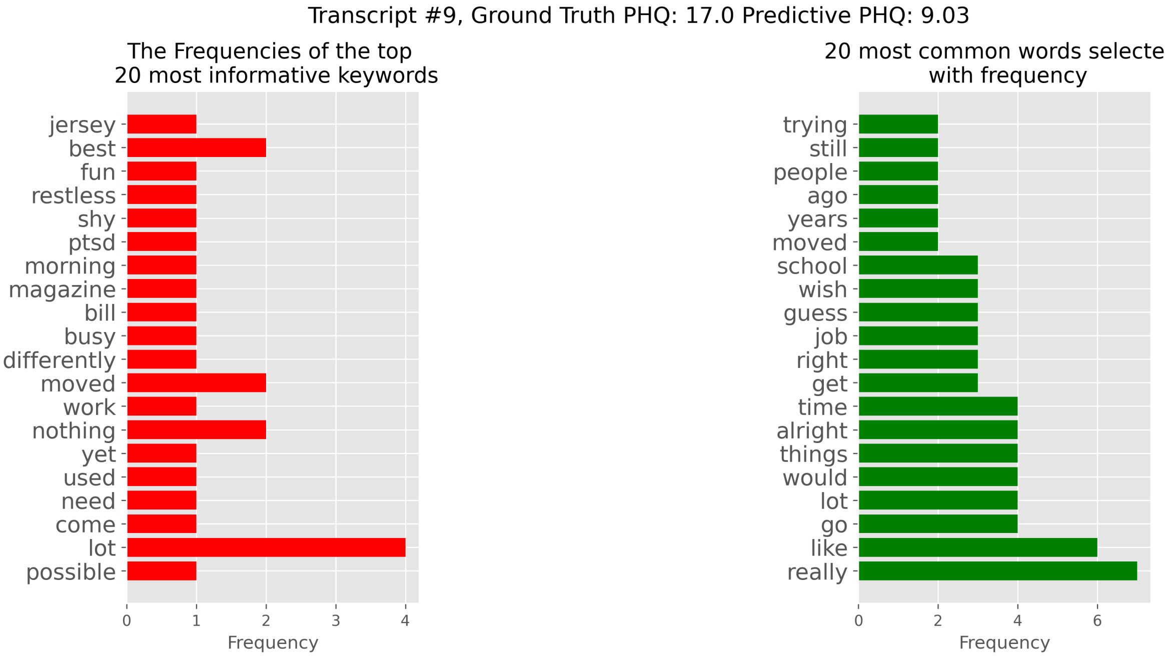


Figure 5.21: Horizontal bar chart of 20 most common words. The bar graph on the left side shows that the 20 most informative keywords (e.g., 'restless', 'fun', 'shy', 'ptsd', 'busy' and 'nothing'), selected by the SGNN model, are related to PHQ topics. The bar graph on the right side displays 20 most common words selected based on word frequency.

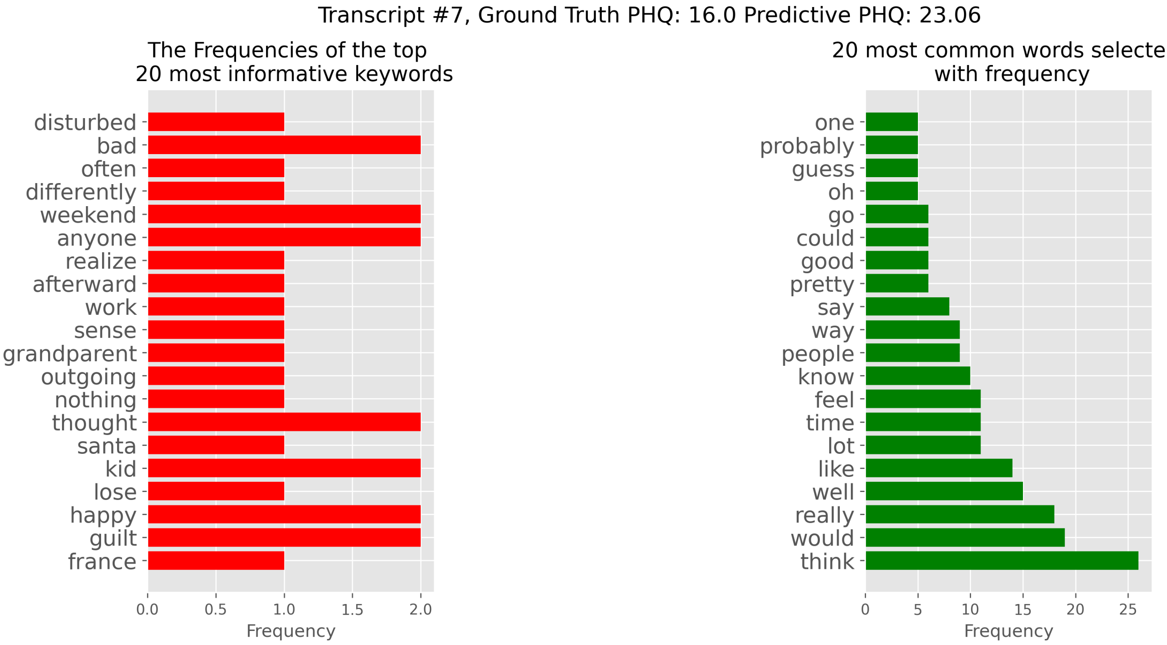


Figure 5.22: Horizontal bar chart of 20 most common words. The bar graph on the left side shows that the 20 most informative keywords (e.g., 'disturbed', 'bad', 'work', 'outgoing', 'nothing', 'kid', 'happy', 'lose', and 'guilt'), selected by the SGNN model, are related to PHQ topics. The bar graph on the right side displays 20 most common words selected based on word frequency.

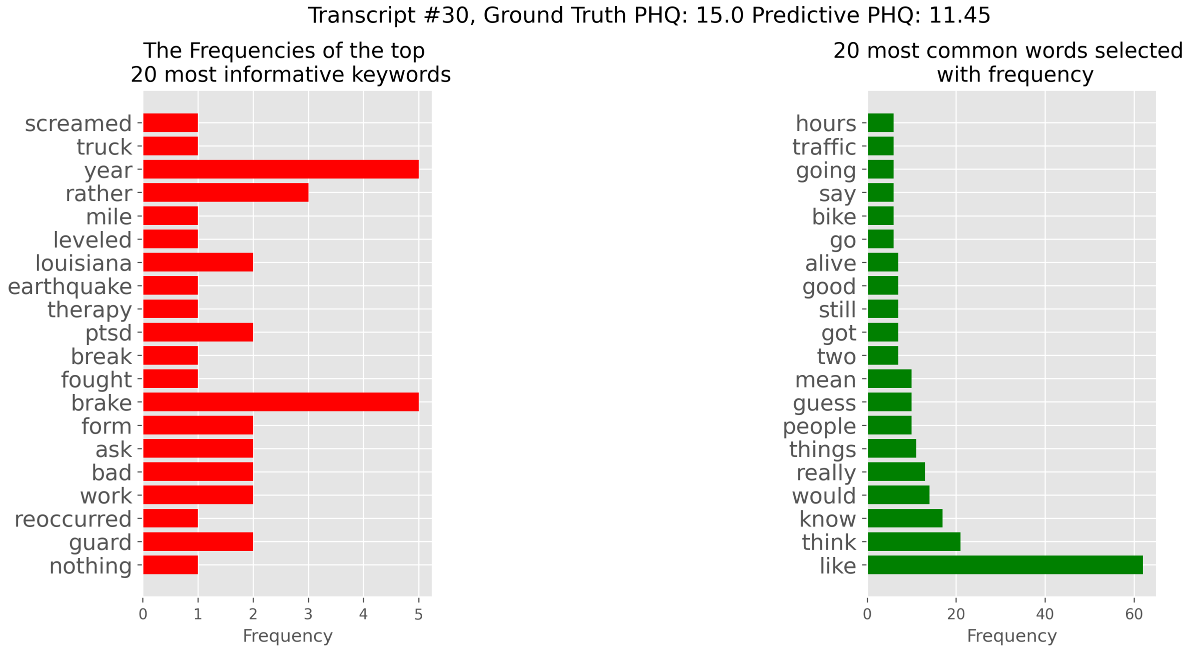


Figure 5.23: Horizontal bar chart of 20 most common words. The bar graph on the left side shows that the 20 most informative keywords (e.g., 'screamed', 'ptsd', 'therapy', 'break', 'work', and 'nothing'), selected by the SGNN model, are related to PHQ topics. The bar graph on the right side displays 20 most common words selected based on word frequency.

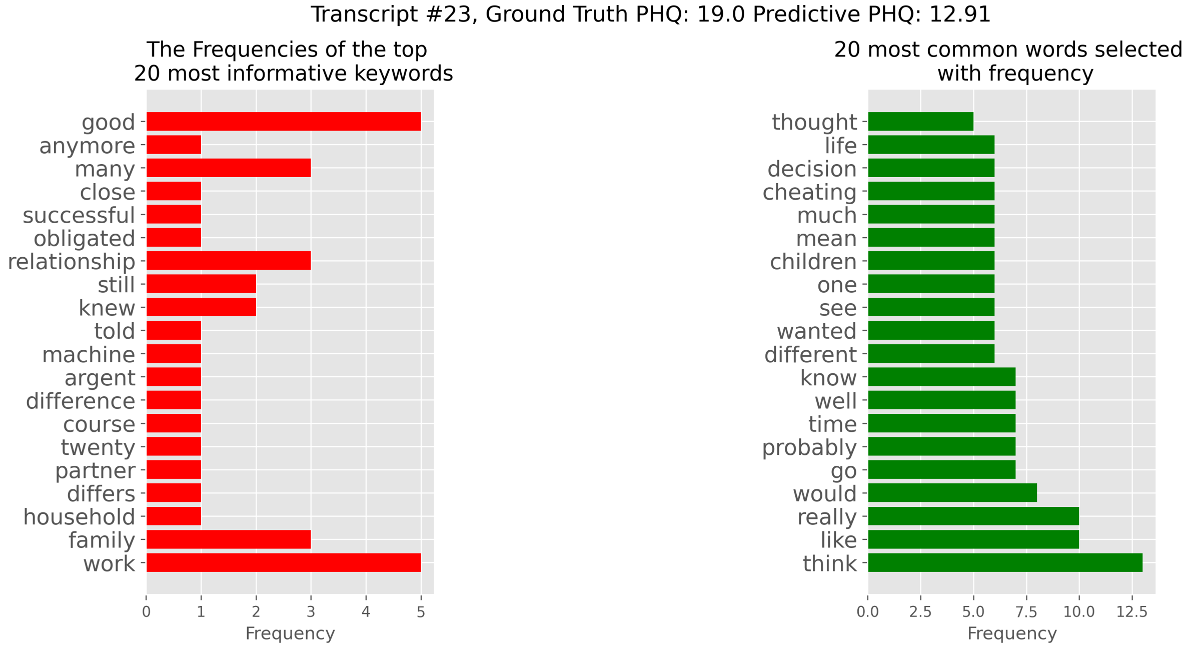


Figure 5.24: Horizontal bar chart of 20 most common words. The bar graph on the left side shows that the 20 most informative keywords (e.g., 'relationship', 'partner', 'household', 'family', and 'work'), selected by the SGNN model, are related to PHQ topics. The bar graph on the right side displays 20 most common words selected based on word frequency.

Chapter 6

Evaluation on Two Further Datasets

6.1 Motivation

This chapter shows that the proposed model can well generalize to two other domains with good experimental results.

1. **Twitter Sentiment Classification** I chose the Twitter Sentiment dataset because it is large, with millions of annotated tweets, which makes it suitable for training deep learning models. The sentiment analysis technique is widely used to analyze mental knowledge of web, social media, and related references. Online social media such as Twitter is a popular platform for people to express their emotions on various topics. Twitter has become an important data source to analyze public emotional reactions and mood oscillations to traumatic events (Pellert et al., 2022). Individual words in tweets cumulatively add predictive information to the sentiment classification performance. Processing raw text is the crucial first step in text classification and sentiment analysis. For example, some conventional stop words like (am, be, at, my, your, on, to), considered to carry little information by default, are bearing low importance scores based on their word embedding. Thus the use of a widely applied dictionary of stop words can limit the overall performance of sentiment classification in Twitter domain.

   Moreover, due to the noisy nature of tweets and ambiguous usage of words, traditional lexicon-based algorithms using pre-built word dictionaries, such as a default sentiment dictionary, lack optimization for tweets (domain), sentiment analysis and a specific context (e.g., sentiment towards a subject matter). For instance, two words of different polarity (good and bad) appear in the same tweet which can complicate Twitter sentiment analysis in a particular context.

To address this problem, I provided a novel way of a DL-based text-to-graph algorithm to obtain in-domain and context-aware embeddings using word co-occurrence information in tweets. I build a text graph for each tweet and I regard all unique words appearing in a tweet as the nodes of the graph. I apply a sliding window in a tweet to detect word co-occurrences and express these as edges.

I suggest that the strength and polarity of words can be captured using co-occurrence patterns of words. I demonstrate that the schema effectively builds all word associations for a given word into the corresponding node. This innovation improves the accuracy of sentiment classification on Twitter data.

2. **Alzheimer's Disease Detection**

   Alzheimer's Disease (AD) is a debilitating disease with no known cure that affects all facets of cognition, including the use of language. Unrecognized dementia has adverse effects that range from anxiety, unexplained symptoms, family discord, and catastrophic events; early diagnosis can help manage symptoms and minimize impact on a person's quality of life (Stokes et al., 2015). However, early diagnosis of neurodegenerative disorders such as AD and related dementia is currently a challenge. Diagnosis of AD dementia, at present, relies on patient and caregiver reports, extensive neuropsychological examinations such as the Mini-Mental State Examination or the Montreal Cognitive Assessment, and invasive imaging and diagnostic procedures (Orimaye et al., 2014). Therefore, there may be a number of obstacles to early diagnosis, including cost, location, mobility and time. Our interest lies in the binary diagnosis of AD dementia using verbal utterances of patients only. Analyzing spoken language which reflects cognitive status fits this purpose.

   Deep learning on graphs nowadays is well developed on learning brain images of a patient for the purpose of AD diagnosis (Cuingnet et al., 2011; Hosseini-Asl et al., 2016; Suk et al., 2014). Applying a graph representation for the same purpose with the use of word embeddings obtained from verbal utterances of patients, although, is unstudied.

   I evaluate the performance of the SGNN model on AD detection. On a sparse clinical language dataset, the SGNN model could predict AD-type dementia with an accuracy of nearly 96%. Moreover the proposed approach performs better than the state-of-the-art methods by 1–4% in accuracy in patient data.

The proposed model captures semantic contextual clues such as semantic abnormality from patient speech, and these semantic deficits are highly associated with AD (Revonsuo et al., 1998). I suggest that patient utterances help to identify semantic deficits in AD. I identify dementia from language samples using dementia and control groups. I also exploit the model's learned features to explain its prediction, allowing individuals and medical experts to understand its decision-making.

6.2 Twitter Sentiment Classification

6.2.1 Dataset

Borderlands Sentiment Twitter[1]dataset includes 74682 tweets labeled as positive (28%), negative (30%) and neutral (42%) respectively. Tweets shorter than two characters are removed. I classified 32,800 tweets in the experiment. Tweets range in length from 3 to 126 words with an average length of 25.3 words per tweet. The distribution of the dataset (32,800 tweets) is shown in Figure 6.1.

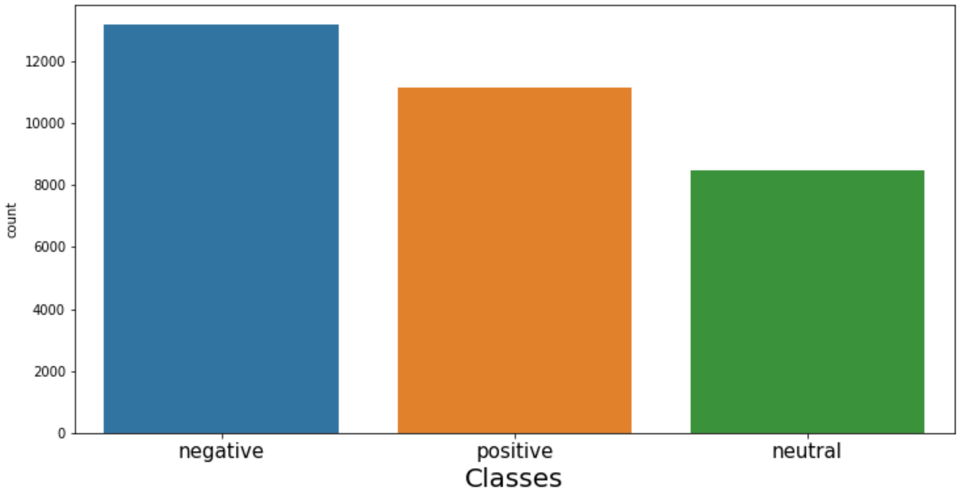


Figure 6.1: A histogram showing the distribution of Twitter sentiment classes on the dataset of 32800 tweets. The Y-axis represents the number of tweets for each class.

.

[1] https://www.kaggle.com/cameronwatts/bag-of-words-sentiment-analysis-with-keras/data?scriptVersionId=78350767last accessed on 11/11/21

6.2.2 Data Pre-processing

I preprocess raw text of tweets before training the model:

URL link & HTML reference character elimination. There are many URLs such as http or https, or HTML entities such as &lt; &gt; &amp, which are embedded in the original Twitter dataset. The removal of links helps improve the performance of sentiment analysis.

Non-Letter character & emotion elimination. Non-Letters, non-English characters, numbers, and the punctuation that is commonly used in emojis or hash marks are removed from the dataset. Punctuation which is not usually associated with an emoji is also removed. In future work, we plan to study the effect of emojis by including emojis to see any improvement in the accuracy of the classification.

Stopword elimination. The words that happen most often do not bear much information. Common words such as “a” and “on” are automatically removed because no other discriminative information of sentiment is added.

Lemmatization. Lemmatization is applied to pre-process raw Twitter dataset.

6.2.3 Experimental Setup

The model is trained to minimize the cross-entropy loss function of classifying sentiments in the training set. I set the learning rate as 10^{-3}, $L2$ weight decay as 10^{-4}, the dropout rate as 0.5, and the window size as 7 to gather word-word occurrence statistics. The batch size used in training is 128. I set the size of the word embedding as 300 and initialized with GloVe for baseline models. I used stochastic SGD (Kingma & Ba, 2014) as the optimizer for training and stopped training if the validation loss does not decrease for 10 consecutive epochs.

6.2.4 Experimental Performance & Analysis

I train the model with a total of 32800 tweets. 10-fold stratified cross-validation is adopted for all experiments. I compare the SGNN model with a GNN baseline model (Johanson & Bejerholm, 2017) by evaluating the performance of sentiment classification.

6.2.4.1 Model Comparison

I evaluate the extent to which the 2-D structure is key to improving performance as opposed to simply having more parameters. I scale the size of feature vectors in the GNN to keep it having the same number of training parameters as of the SGNN, so that the number of parameters to be estimated in training is the same as in the SGNN. Table 6.1 shows the performance gains of the proposed model on the Borderlands Sentiment Twitter data when 2-D node attributes are applied. This gain demonstrates that a GNN model with 2-D node attributes can capture more context-level features of sentiments from text, and thus can greatly improve the accuracy of this domain task.

I further evaluate the efficiency of the proposed method by comparing the performance of text-to-numeric word embedding generated by two different types of deep learning models respectively:

1. I used a general LSTM model, which performs as a sequence-based deep learning method, with 300-dimensional pre-trained GloVe embeddings.
2. I used a standard 1D-CNN model, which performs as a sequence-based deep learning method, with 300-dimensional pre-trained GloVe embeddings.

Compared to the performance of general deep learning methods, the SGNN model with and without pre-trained word embeddings performed well in text sentiment classification tasks (see Table 6.1). The experimental results demonstrate that an SGNN model can also achieve a good result on short texts like tweets.

6.2.4.2 Test Performance Analysis

Table 6.1 shows the effectiveness of the proposed method on a short text dataset like Twitter. Furthermore, the results demonstrate the effectiveness of the proposed schema encoders in modeling consecutive and short-distance semantics.

The 1-D CNN model performs better than the proposed SGNN with and without pre-trained word embeddings for short text. Similarly, Bi-LSTM models using pre-trained word embeddings perform better than the proposed SGNN model. This is likely due to the fact that word orders are important in sentiment classification or short text. Another reason might be that tweet graphs have few nodes and few edges. The lack of edges limits the message passing among the nodes. Thus there are only a few word connections that can be encoded for learning because the tweet text is very brief.

For more in-depth performance analysis, I note that SGNN with randomly initialized word embedding outperforms the SGNN model with GloVe embedding and improves the accuracy of sentiment classification. This increase in accuracy may suggest that the proposed model helps to develop and analyze word embeddings to capture directly from domain-specific contexts and trending keywords using large corpora.

Moreover, Kumawat et al. (2021) proposes a hybrid deep-learning method that combines a Long Short-Term Memory model and a Transformer model to capture long-distance contextual semantics. Their results demonstrate the effectiveness of this method in the same task. GCN-based (Tang et al., 2020) models aggregate information from only the direct neighbor nodes. They have limited capacity to capture long-range contextual dependency information.

Table 6.1: Performance comparison of the proposed deep neural network and state-of-the-art multi-class sentiment classification methods on Twitter dataset. I used a 10-fold cross-validation scheme and gave mean(standard deviation) results. – represents that standard deviations are not reported in the original work.

Sentiment Classification	
DL Methods	Accuracy
LSTM+GloVe	0.918 (0.0212)
RNN+GloVe	0.837 (0.0221)
Bi-LSTM+GloVe	0.977 (0.0043)
1D-CNN+GloVe	0.970 (0.0051)
GNN+GloVe	0.878 (0.0963)
SGNN+GloVe	0.948 (0.0019)
SGNN	**0.958(0.0013)**
Bert (Kumawat et al., 2021)	0.812
Transformer+LSTM (Kumawat et al., 2021)	0.914
GAT (Tang et al., 2020)	0.699 (–)
GCN (Tang et al., 2020)	0.684 (–)
GCN+BiLSTM(Tang et al., 2020)	0.738 (–)

6.3 Model Visualization

To intuitively show the learned embeddings of tweets, I visualize graph representations in 2D space using the t-SNE algorithm (van der Maaten, 2014). From the visualization (see figure 6.2), it shows that the SGNN did a good job of clustering tweets in a low-dimensional space. The model successfully classified all tweets in the test set into three classes.

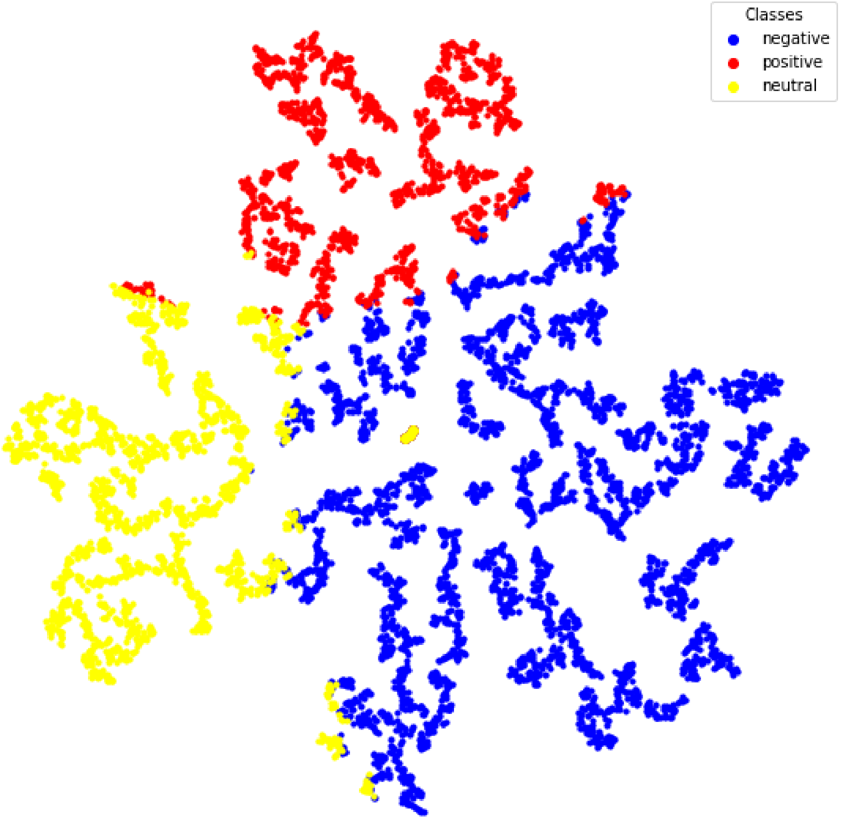


Figure 6.2: 2D visualization of node representations on Twitter using t-SNE. The figure shows the t-SNE visualization of the graph embedding layer (i.e. the pooling layer) of the proposed model.

6.4 Alzheimer's Disease Detection

6.4.1 Dataset & Data Pre-processing

Dementia Bank Pitt Corpus[2] I used an existing Dementia Bank clinical dataset in this study. A detailed description of this dataset is available in Becker et al. (1994). The dataset was created during a longitudinal study conducted by the University of Pittsburgh School of Medicine on Alzheimer's and related Dementia. The dataset contains transcripts of verbal interviews with AD and related Dementia patients. The average length for each speech transcript is nearly 83 words.

[2] https://https://dementia.talkbank.org/last accessed on 01/02/22

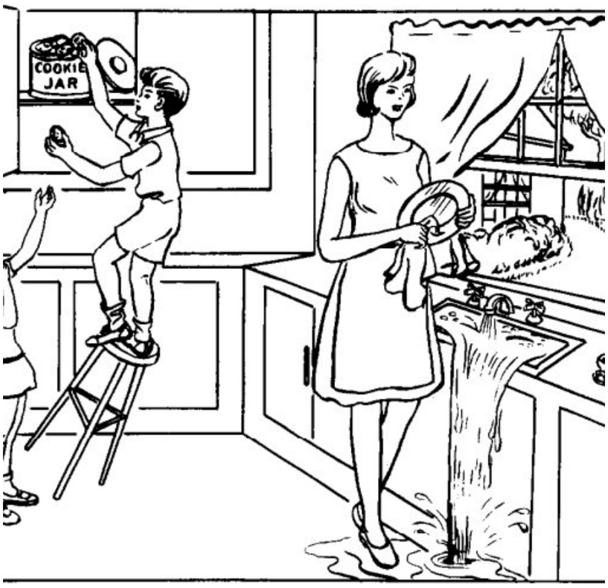
Cookie Theft stimuli.	A Raw Speech Transcript Sample	A Binary Prediction Score
	“oh any anything I see well there’s a little girl here she’s pointing at something and that looks like her brother is and here on her right side there’s a young lady there she’s washing the what she she’s cleaning the she’s cleaning what is she she”	1 (=AD) or 0 (=CT) ?

Table 6.2: A sample from Pitt Dataset

There are 169 subjects classified as an AD dementia group on the basis of clinical or pathological examination, and 99 subjects classified as Control (CT) group. Many participants had multiple visits over the duration of this longitudinal study. I use 309 transcript samples from those in the AD group, and 243 from those in the CT group.

In brief, interviews were conducted in the English language and were based on the Cookie-Theft picture description from the Boston Diagnostic Aphasia Examination (Kaplan, 1983), a widely-used diagnostic test for language abnormality detection. During the interview, patients were presented with a “Cookie Theft” picture stimulus (see Figure 6.2) and were told to discuss everything they could see happening in the picture. The patients’ verbal utterances were recorded and then subsequently transcribed verbatim (MacWhinney, 2000). Thus, in this study, I train the model using verbatim transcriptions of the audio recordings. I pre-processed raw transcripts using lemmatization and stopword removal (Camacho-Collados & Pilehvar, 2017; HaCohen-Kerner et al., 2020).

6.4.2 Experimental Setup

The model is trained to minimize the cross-entropy loss function of predicting the class label of participants' speech records in the training set. The window size of 9 is used to detect word co-occurrences. I set the learning rate as 10^{-3}, $L2$ weight decay as 10^{-4}, and dropout rate as 0.5. The batch size of the model is 32. For models $\big($1-6 in Table 6.3$\big)$ using pre-trained word embeddings, I used 300-dimensional GloVe word embeddings. Each model is trained using the stochastic SGD optimizer.

6.4.3 Experimental Performance & Analysis

In the experiments, I compare the proposed model with five state-of-the-art deep models on the task of binary detection of AD-type dementia. It can be seen from Table 6.3 that the method with and without pre-trained embeddings outperforms the state-of-the-art methods with accuracy of around 95% and 96% respectively, and F1 scores of 0.958 and 0.962 respectively.

I evaluate the efficiency of 2D node structure employed in SGNN model. I scale the feature vectors in the GNN model so that the number of parameters in the GNN model to be estimated in training is the same as in the SGNN (A total of around $30,000$ parameters). I compare the GNN using 1D node attributes (in row 5) to SGNN model using 2D node attributes (in row 7). From Table 6.3, it can be seen that the proposed model yields higher performance with a detection accuracy of around 95% and F1 score of 0.958. Thus the schema-based method can effectively improve the AD disease detection accuracy with more than 40% relative gain compared to the GNN model.

I also explore the effectiveness of utilizing GloVe word embedding in AD detection. As shown in Table 6.3, the proposed SGNN model with GloVe embedding outperforms the SGNN model with randomly initialized embedding and achieves a detection accuracy of around 96% and F1 score of 0.962.

Moreover, I compare the proposed method to four deep learning models that usually treat text as sequences and encode them for classification: 1. LSTM; 2. Bi-LSTM; 3. RNN; 4. CNN models (models are introduced in section 2.3 in chapter 2.).

These four models require lengths of inputs to be the same. I set the total length of utterances (of a speech transcript input) to 80 because 99% of transcripts have words less than 80. I truncated lengths of the speech transcripts that contain more than 80 words and added padding for the others that contain less than 80. On the contrary, the SGNN model accepts flexible size inputs, which allows us to learn representations of text inputs with variable lengths.

Unlike the sequence-based deep learning models that learn a sentence (or sequence)-level mapping, the proposed method can learn a node-level embedding to obtain word features that can be updated internally based on the associations between words in a transcript. I demonstrate that encoding the information from word-word associations can help to learn word embeddings from the dementia context. According to the experimental results, the SGNN model can enhance the overall performance of AD detection of dementia. The results of the comparison experiments using different deep models are shown in Table 6.3.

Table 6.3: Performance of evaluated models. I adopted 10-fold stratified cross-validation and report mean(standard deviation). – represents that standard deviations are not reported in the original work.

Detection:AD or CT				
DL models	Accuracy	F1	Precision	Recall
LSTM+GloVe	0.929 (0.0929)	0.934 (0.0911)	0.938 (0.1011)	0.931 (0.0863)
Bi-LSTM+GloVe	0.944 (0.0731)	0.955 (0.0727)	0.952 (0.0800)	0.958 (0.0652)
RNN+GloVe	0.916 (0.1348)	0.923 (0.1285)	0.918 (0.1258)	0.928 (0.1328)
CNN+GloVe	0.948 (0.0798)	0.958 (0.0843)	0.951 (0.0962)	**0.960(0.0720)**
GNN+GloVe	0.675 (0.0165)	0.723 (0.0180)	0.697 (0.0129)	0.760 (0.0315)
Attention-CNN+Attention-BiGRU (Chen et al., 2019)	97.42 (3.0900)	–	–	–
BERT (Searle et al., 2020)	0.84 (–)	0.82 (–)	0.86 (–)	0.85 (–)
textBert (Zhu et al., 2021)	0.82 (2.8300)	0.80 (3.5500)	0.88 (2.0900)	0.74 (5.5300)
GCN (Millington & Luz, 2021)	0.75 (–)	–	–	–
SGNN+GloVe	**0.957(0.0837)**	**0.962(0.0789)**	0.964 (0.0771)	0.952 (0.1067)
SGNN	0.948 (0.1158)	0.958 (0.0880)	**0.970(0.0600)**	0.942 (0.1181)

6.4.4 Qualitative Analysis

Table 6.4 shows some samples of false positive and false negative respectively produced by the proposed model. Label 0 indicates that the description was uttered by a healthy individual (in the CT group) and label 1 indicates that the description was uttered by a patient (in the AD group).

According to samples of patient voice transcripts listed in the Table 6.4, there are several obvious language cues such as semantically related but incorrect words (i.e., calling a stool a chair) and language fluency problems. These false predictions therefore indicate that the model failed to capture these types of language cues that are very useful in distinguishing AD from controls (Jarrold et al., 2014; Nebes, 1989; Nicholas et al., 1985).

False Negative samples	• "faucet turned water overflowing sink grass growing tree growing indication wind coming window mother washing drying standing water johnny cookie out of handing cookie falling sister putting finger nose i action johnny of course reaching cookie jar" • "boy stool ladder stool tipped ladder step stool i upset cookie sink flowing floor mother boy i i wanna step stool step stool stool water overflowing floor water back i i i" • "boy cookie jar fall floor chair stool mom drying dish paying attention water running water floor girl i begging brother give cookie i summer time window open grass shrubbery dish mom dried"
False Positive sample	• "child falling chair taking cookie jar girl standing floor cookie door cabinet door open mother washing dish sink overflowing water running i drying washing kitchen window curtain window open view back dish counter"

Table 6.4: Samples of false positive and false negative respectively produced by the proposed model.

6.5 Model Visualization

The advantage of the proposed model is that we can interpret the binary classification process of the model. I exploit the graph structures to explain model's decision-making process of classifying AD group and CT group. I argue that the anomaly of the words chosen in the language of dementia patients can be identified by the relationship between words in the content of the image description. To achieve this, I use the schema encoders to encode word-word co-occurrence information from the context derived from descriptions of a cookie theft picture.

Some research (Nebes, 1989; Nicholas et al., 1985; Rohrer et al., 2008) indicates that patients with dementia struggle to communicate their thoughts. For instance, a person with dementia may have a hard time finding the right words to express themselves such as describing an image. I treat word-finding difficulties as symptoms of dementia. I quantify correlations between words using semantic contextual word embeddings extracted from the latent layer of the SGNN model.

I plot in figures 6.3, 6.4 and 6.5 for 3 words ("girl", "boy", "mother") in a transcript from the test set.

Figure 6.3 shows the input of the 1st message passing layer to obtain a schema where all rows are identical except the row corresponding to the word associated with this node. It applies a linear mapping to the schema rows and adds in a schema of identical rows derived from the second and third invariant terms of the update equation 4.1.

Figure 6.4 shows the schemas in the output of the 1st message passing layer. Figure 6.5 shows the schemas in the output of the 2nd message passing layer. Rows which are identical in a schema correspond to nodes that have not been encountered in the neighborhood of the current node. The schema of each node is updated via propagation by encoding information from its neighbors.

After pooling the 2nd message passing layer by equation 4.4, H is made by stacking row vectors h_{v_i} (h_{v_i} is 64-dimensional) for all $v_i \in V$. I take variance value across all row vectors of H in each dimension of all nodes to measure the salience or influence of the node. H_{ij} denotes the value of j-th dimension of word i. The method is taken from the work of Li et al. (2015). The salience of each word is defined by the expression: $||H_{ij} - \frac{1}{|V|}\sum_{i' \in |V|} H_{i'j}||$. Figure 6.6 shows the saliency heatmap for the same transcript.

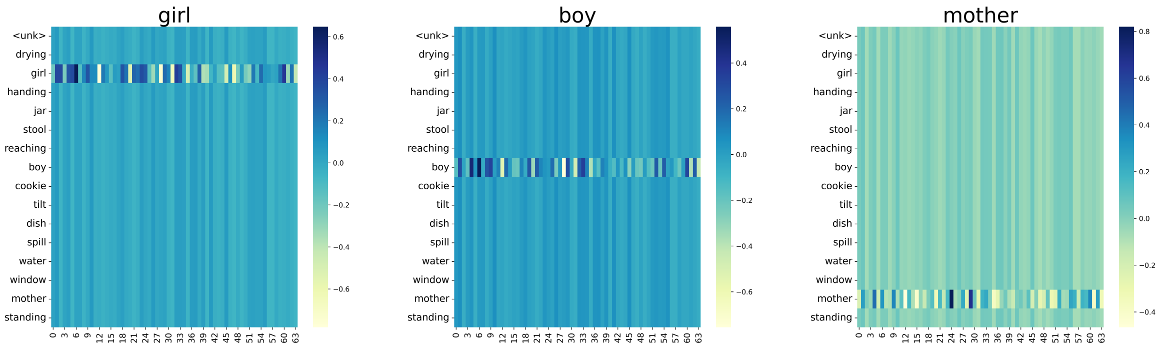


Figure 6.3: Input of the 1st message passing layer for a particular transcript. The x-axis represents the size of a row vector.

Moreover, I present eight saliency heatmaps for eight speech transcripts, including four predictive AD patients and four healthy control participants. Each row corresponds to a saliency score for the correspondent word embedding with each grid representing each dimension. The visualization of heatmaps describes abnormalities (the absence of essential indicator words) existing in the language of demented individuals and healthy individuals respectively.

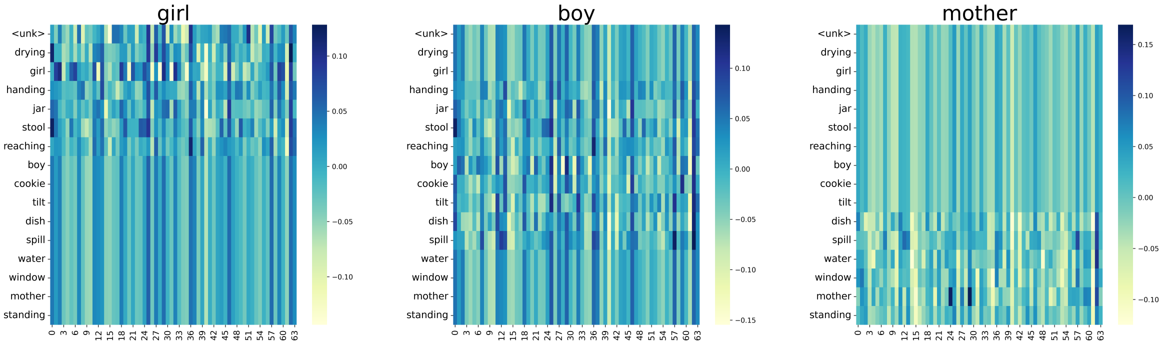


Figure 6.4: Schemas in the output of the 1st message passing layer for the same transcript applied in figure 6.3. The x-axis represents the size of a row vector.

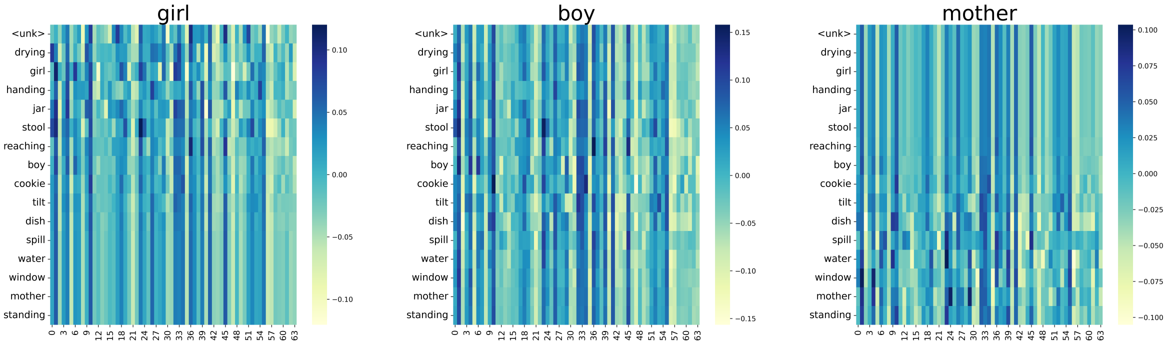


Figure 6.5: Schemas in the output of the 2nd message passing layer for the same transcript applied in figure 6.3. The x-axis represents the size of a row vector.

I found a set of salient words in plots that are recognized as word anomalies by the model. For instance, I discovered several keywords, such as “mom”, “mother”, “child”, “girl”, “boy”, and “sister”, in four heatmaps (see figures 6.11, 6.12, 6.13, 6.14) that were recognized as inconsequential by the model. The lack of keywords reveals a lack of understanding of the content of the cookie theft picture and may therefore be identified as abnormalities by the model. Furthermore, some terms (words) such as “standing”, “washing”, and “running”, treated as abnormalities, did not appear salient in the heatmaps, referring to another sign of word-finding difficulty in speech of AD patients. As a result, the absence of content words in the context of cookie theft picture reflects that AD patients tend to produce less or no content words and low information content.

In contrast, heatmaps from predictive healthy controls (see figures 6.7, 6.8,6.9,6.10) showed a few words, such as “pointing”, “coming”, “grass”, and “talking”, which are identified as non-salient by the model in some speech transcripts of healthy individuals. The abnormalities

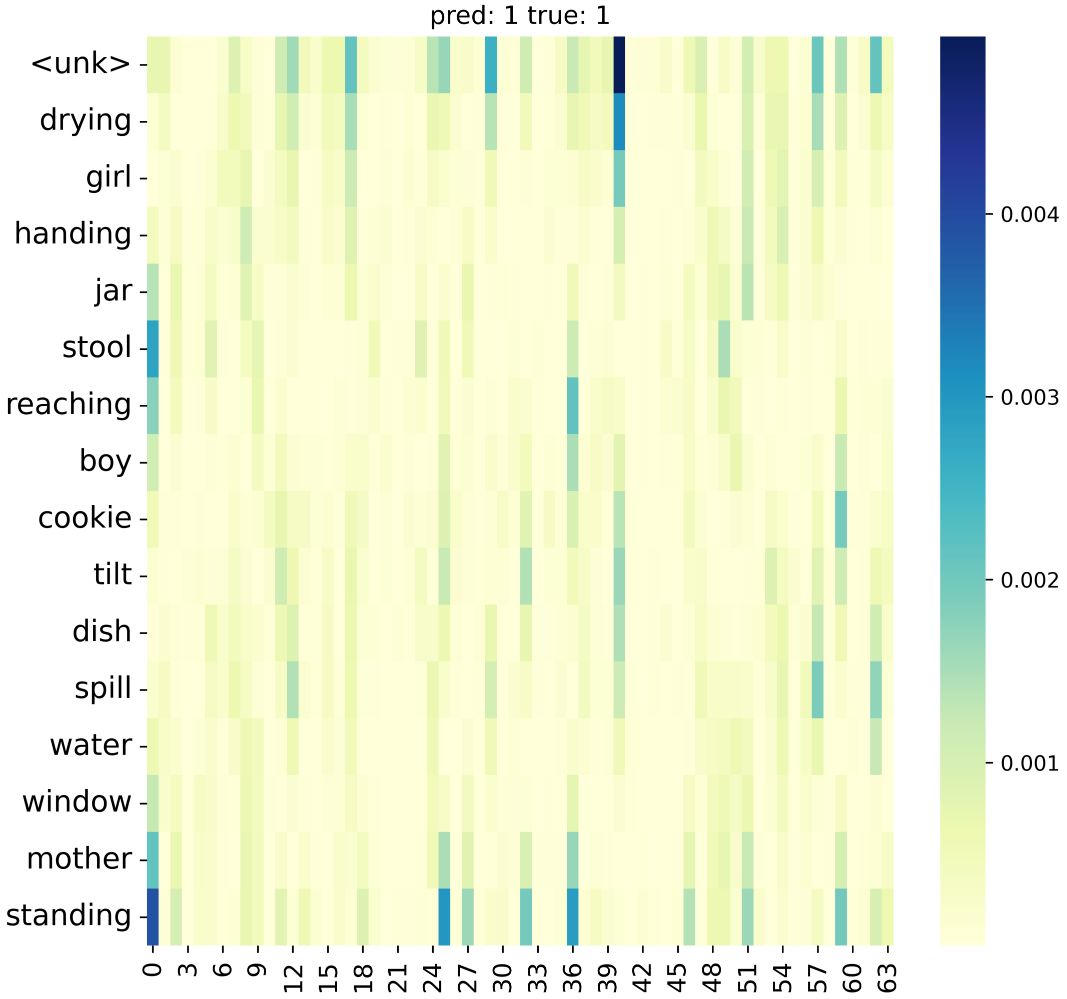


Figure 6.6: Saliency heatmap of the same transcript applied in figure 6.3. The x-axis represents the size of a row vector.

captured from healthy controls convey less pertinent information about the picture.

Consequently, I found that heatmaps created by the model's latent embedding can explain the differences between two groups of individuals. I observed that AD patients produced fewer content elements than healthy people (from the CT group). Also, the AD patients produced more irrelevant, unrelated, or inappropriate content elements than the healthy group. These findings corroborate reports in the literature (Croisile et al., 1996; Groves-Wright et al., 2004; Nebes, 1989; Nicholas et al., 1985; Rohrer et al., 2008). Features extracted from the model help to discover latent meanings capable of capturing psychologically interesting dimensions of language.

6.6 Conclusion

This thesis shows that the SGNN model outperforms the GNN model that we have discussed in chapter 3 by leveraging the novel 2-D node attributes. I evaluate the performance of the novel model on both a Twitter sentiment dataset and a dementia dataset. The relative performance demonstrates the benefit of using 2-D schema node attributes as opposed to vectors.

I trained the proposed SGNN model as a classifier and applied the trained classifier in Twitter sentiment classification, a different domain task from depression severity regression. I demonstrate the effectiveness of deep graph model through experiments by comparing it with sequence-based deep models. In this research, I achieved an accuracy of around 98% over three sentiments: positive; negative; neutral. I note that the proposed model with pre-trained word embeddings performs much better on Twitter data than the one with randomly initialized word embeddings. The possible reason might be that each tweet is very short and has few word-word connections. Therefore, the edges in a tweet (as a graph) are much fewer than in a transcript of conversations, which limits the message passing among the nodes.

Moreover, I tested the proposed model on dementia data. I trained the model to detect AD using patient speech transcripts. When pre-trained GloVe word embeddings are provided, the proposed SGNN model performs the best and outperforms all baseline models, achieving an accuracy of around 96% in classifying AD patients and healthy controls. I extracted and used the output of a latent layer of the model to explain its results. Below, I present heatmaps that illustrate abnormalities existing in the context of patient utterances and this finding is also reported in psychological research. The model visualization explains the difference in semantic context between the language use of AD and CT groups.

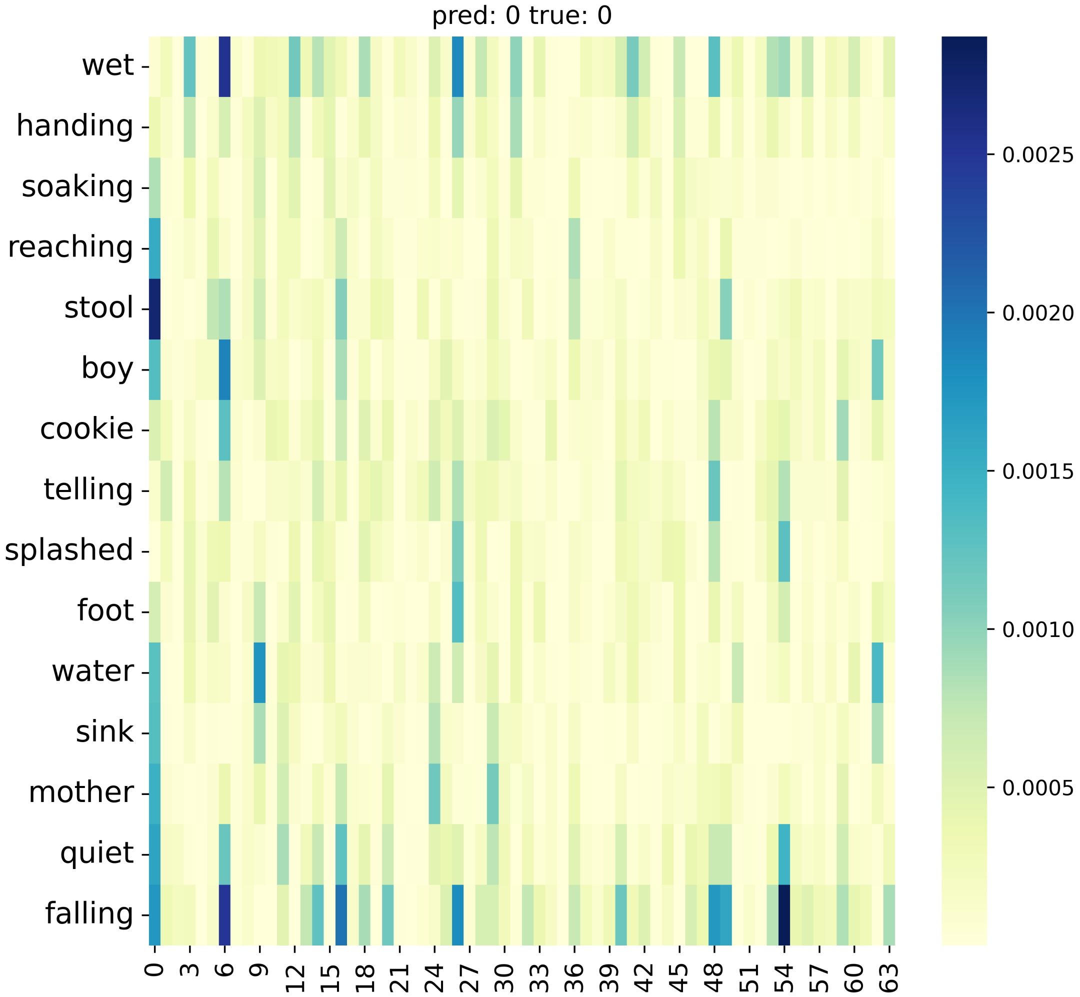


Figure 6.7: Example of a saliency heatmap for a predicted control. Emphasized words: wet, stool, boy, cookie, water, sink, mother, falling. true: ground truth label. pred: predicted label. The x-axis represents the size of a row vector.

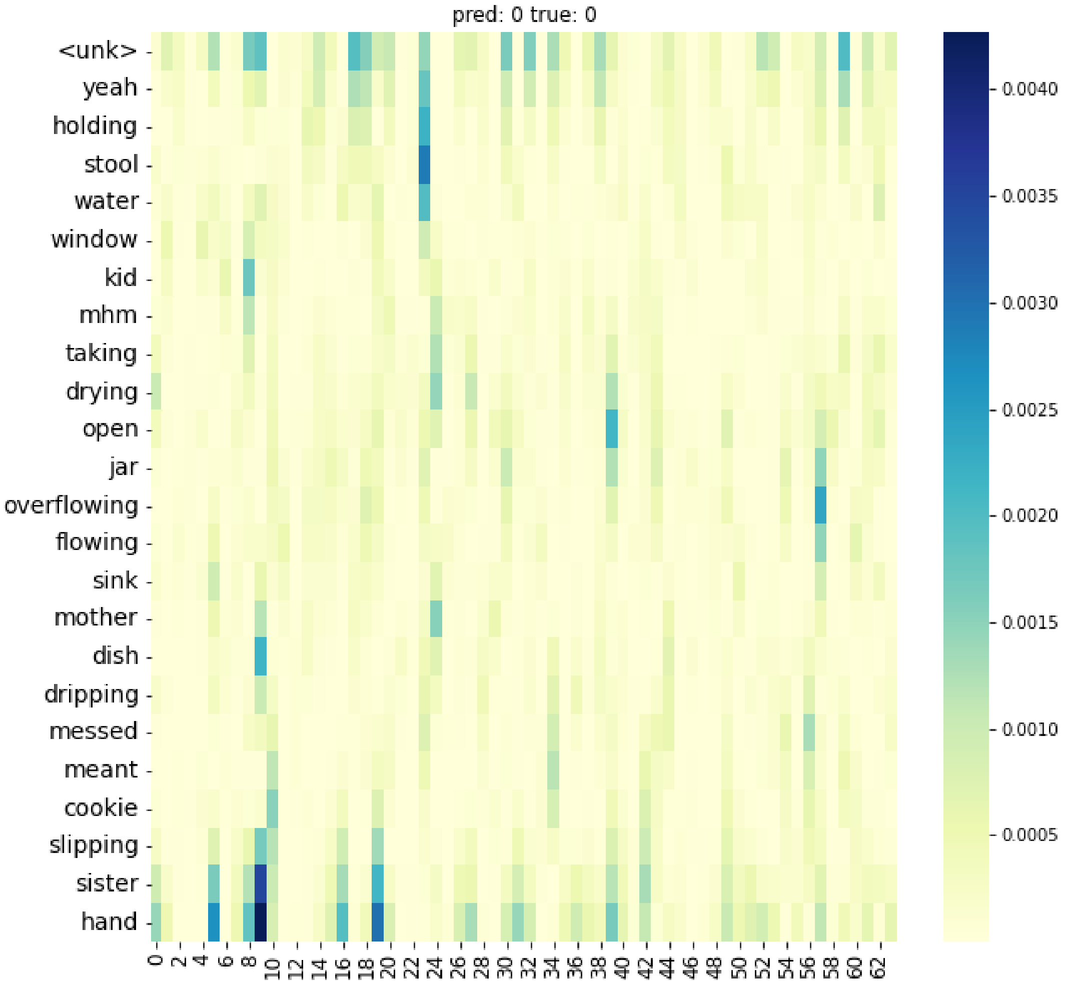


Figure 6.8: Example of a predicted saliency heatmap for a control. Emphasized words: stool, water, kid, dish, mother, slipping, sister, hand. true: ground truth label. pred: predicted label. The x-axis represents the size of a row vector.

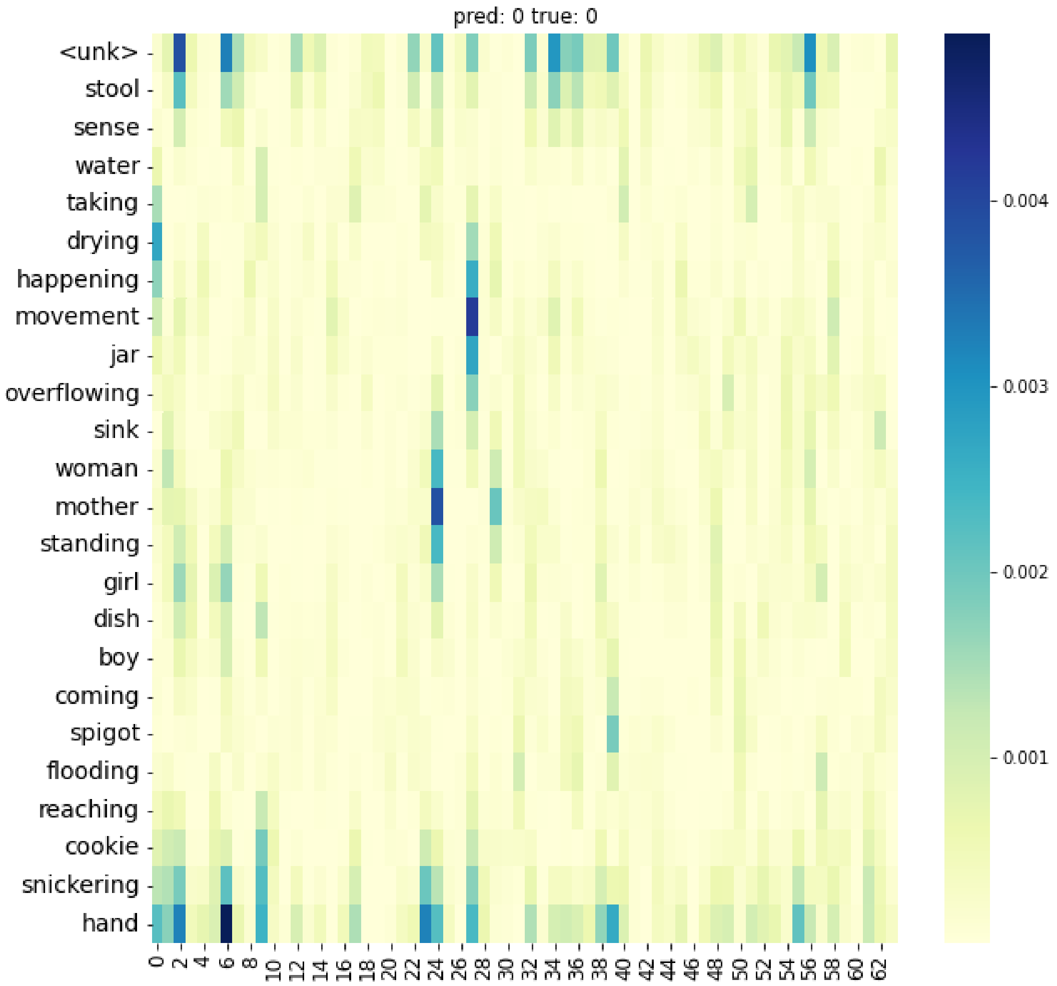


Figure 6.9: Example of a predicted saliency heatmap for a control. Emphasized words: stool, taking, drying, happening, movement, jar, overflowing, woman, mother, standing, girl, dish, cookie, snickering, hand. true: ground truth label. pred: predicted label. The x-axis represents the size of a row vector.

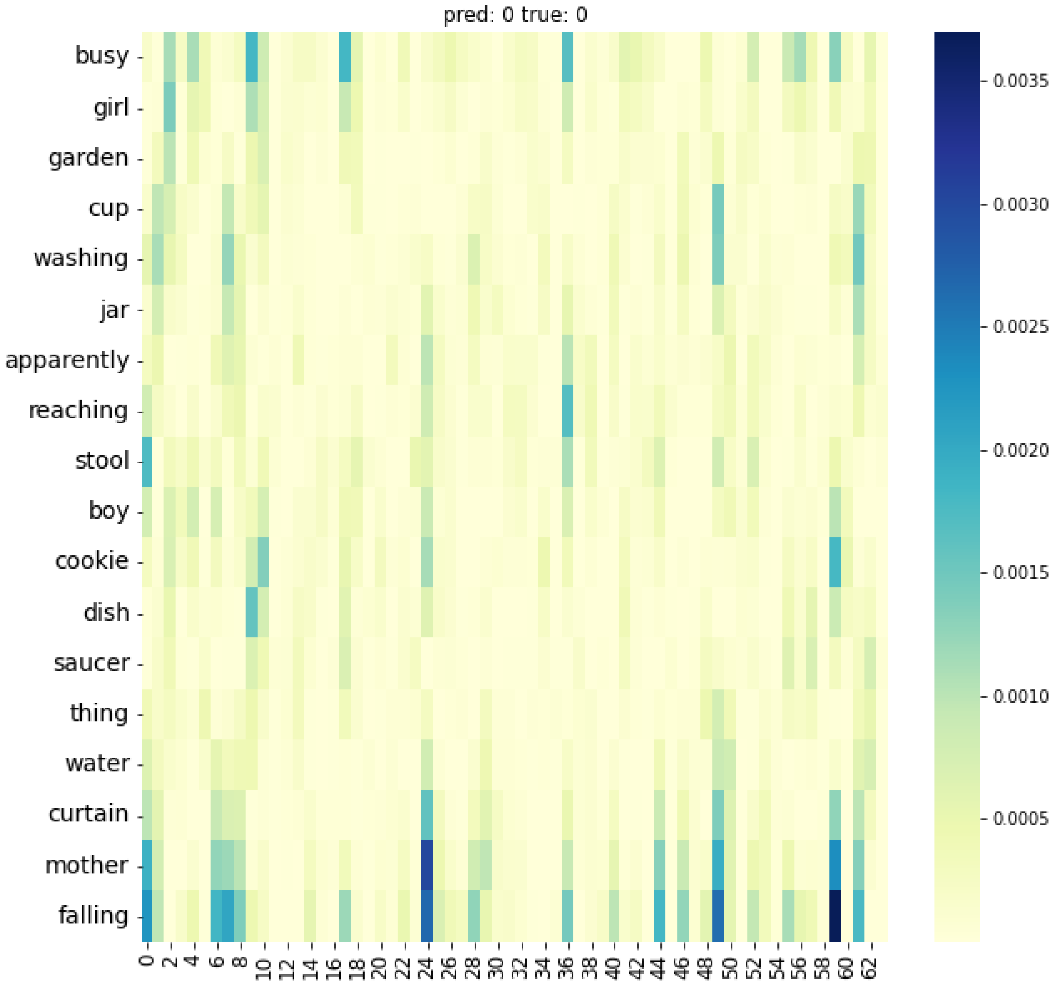


Figure 6.10: Example of a predicted saliency heatmap for a control. Emphasized words: busy, girl, cup, washing, reaching, stool, boy, cookie, dish, curtain, mother, falling. true: ground truth label. pred: predicted label. The x-axis represents the size of a row vector.

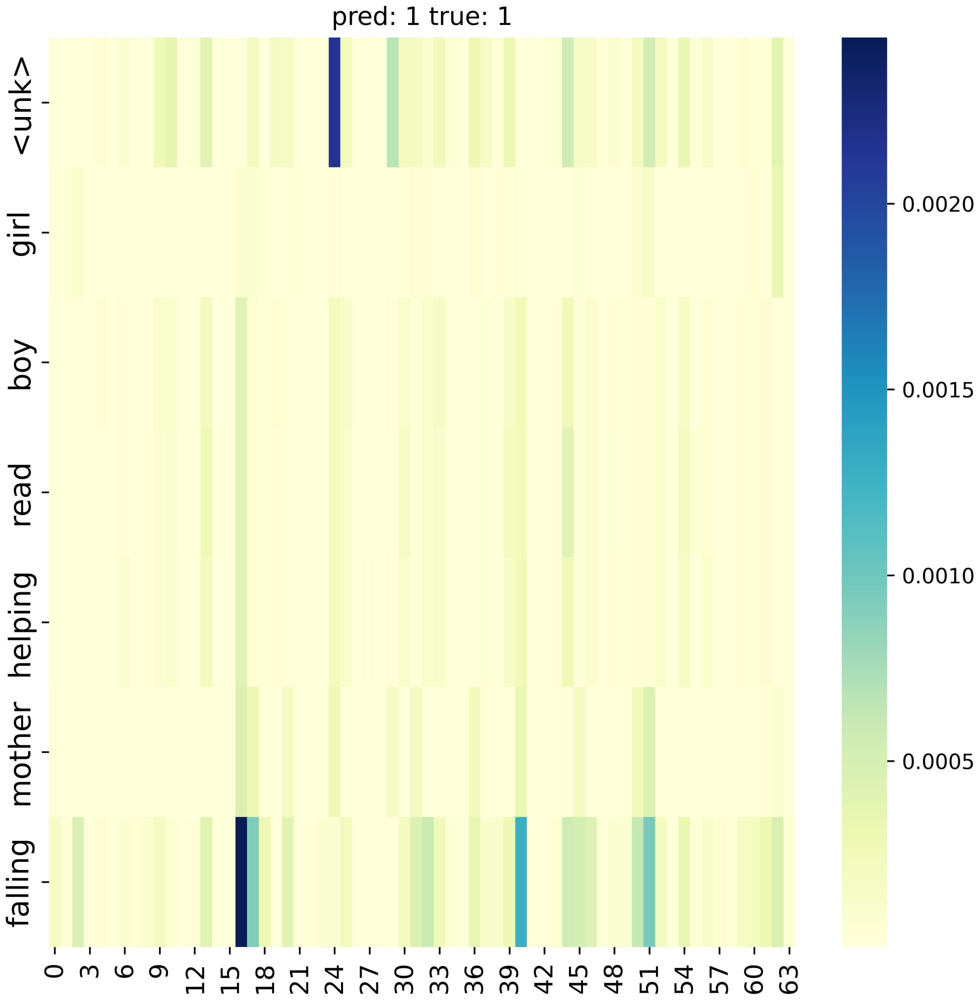


Figure 6.11: Example of a predicted saliency heatmap for an AD. Emphasized words: falling. true: ground truth label. pred: predicted label. The x-axis represents the size of a row vector.

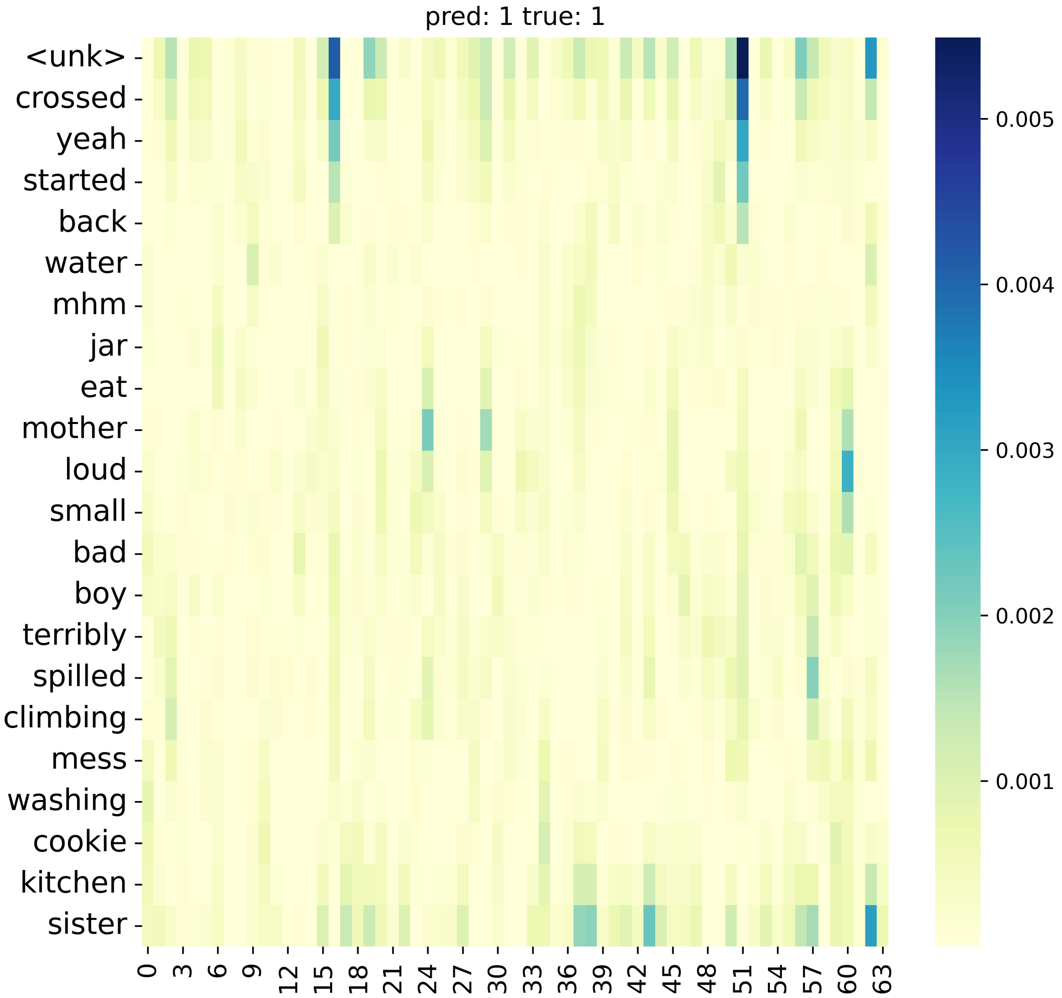


Figure 6.12: Example of a predicted saliency heatmap for an AD. Emphasized words: crossed, yeah, started, mother, sister. true: ground truth label. pred: predicted label. The x-axis represents the size of a row vector.

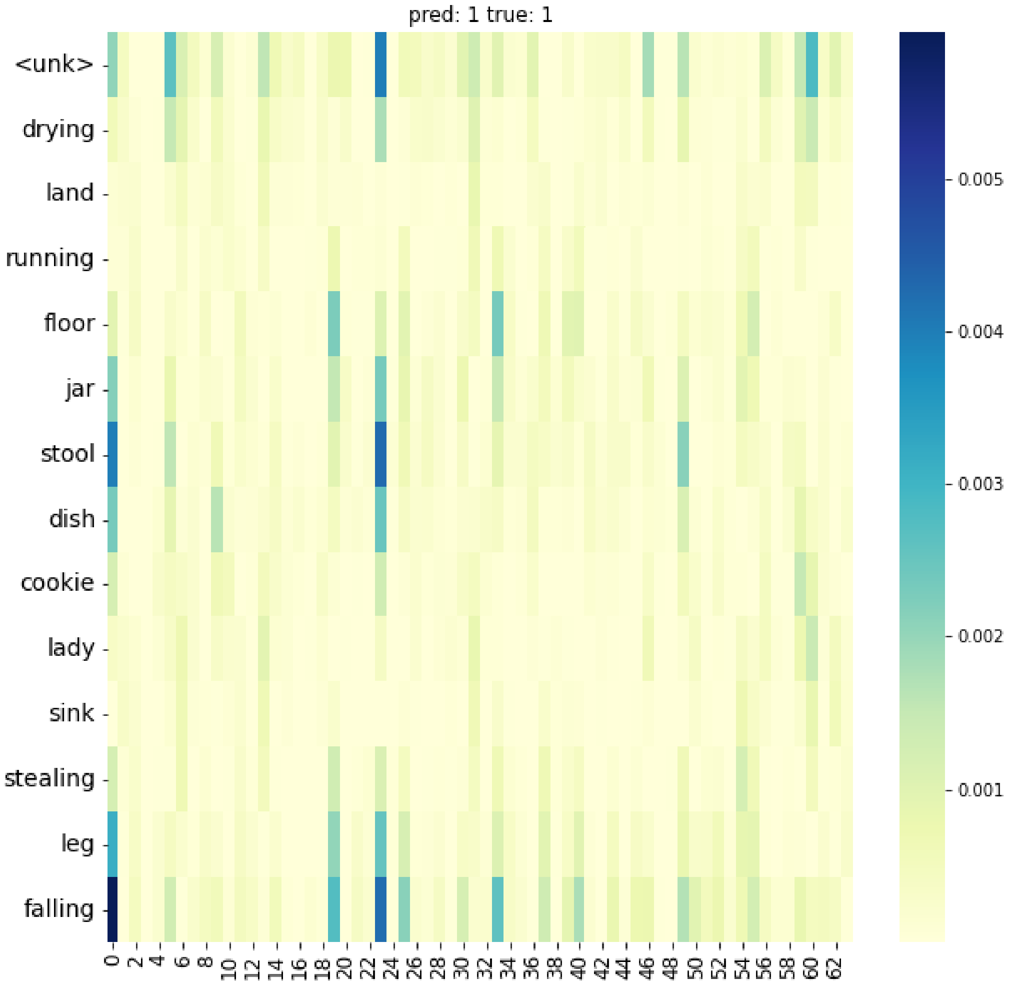


Figure 6.13: Example of a predicted saliency heatmap for an AD. Emphasized words: drying, floor, jar, stool, dish, leg, falling. true: ground truth label. pred: predicted label. The x-axis represents the size of a row vector.

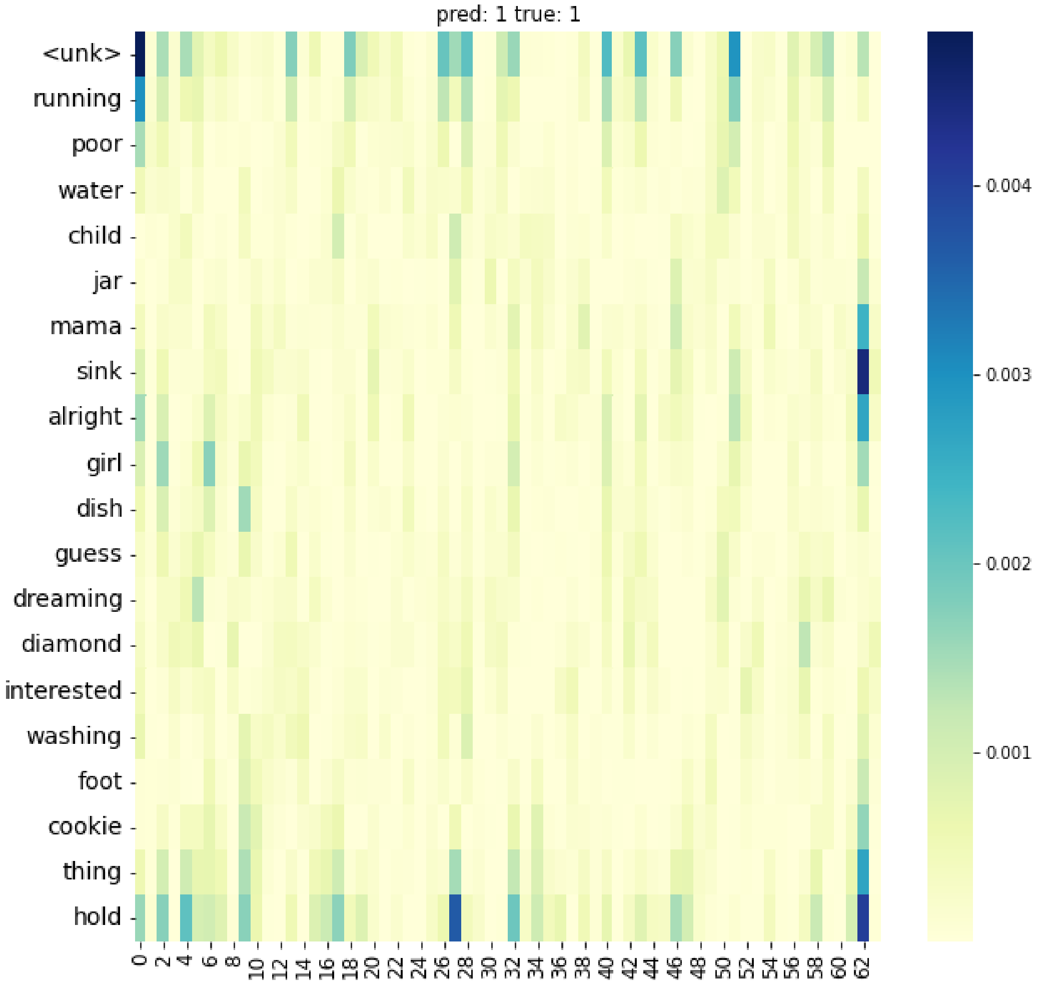


Figure 6.14: Example of a predicted saliency heatmap for an AD. Emphasized words: running, mama, sink, alright, girl, thing, hold. true: ground truth label. pred: predicted label. The x-axis represents the size of a row vector.

Chapter 7

Conclusions

In the research, I found that the relationship between language use and depression is significant and could lead to opportunities to automatically detect individuals at risk of depression from textual data (i.e., clinical interviews and questionnaire surveys conducted by hospitals or agencies). The proposed text-only method outperforms the state-of-the-art methods, including multi-modal methods, on the DAIC-WOZ depression benchmark.

Deep learning methods have been widely applied in text sentiment analysis. The most popular deep neural networks used for capturing linguistic indicators through patients' interview transcripts are CNNs and LSTMs. In a CNN, the text features representation is constructed by acquiring the local information in the filter region, which makes it difficult to learn the dependencies at the level of individual words between distant positions. In comparison, an LSTM (or RNN) connects contextual memory and stores more long-term global information than a CNN. However, these models focus on learning a mapping of consecutive word sequences, they do not explicitly use word co-occurrence information, and the complex model structures are not easy to interpret their results.

My research addressed the above barriers by introducing a deep graph learning model to predict depression on different scales. I created a graph for a patient's self-reported transcript and learned a mapping between a graph and a depression score. Contrasting with a standard vector representation, I introduced a novel way of using graphs to learn word features at the node level.

My research demonstrated that graph representation learning enhanced the automation of decision-making without human intervention. Furthermore, the performance of this work on the evaluation of two new datasets demonstrated the generality of the proposed method. Besides, the proposed graph-based deep model is explainable, which helps domain experts better understand the clinical context behind the data.

7.1 Major Contributions

The main contributions of this research work are:

- I introduced a novel deep graph learning model on a patient transcript (a long text) to capture contextual semantics for depression language. I utilized a graph representation algorithm to generate a fine-grained depression score ranging from 0 to 24. My method enhances the prediction accuracy of the generic depression detection as well as facilitating fine grain analysis of depression from mild to severe.

- I designed innovative 2D "schema" encodings that provide global representations of every vocabulary word for node attributes. Schemas are updated using a GNN-based message passing algorithm. My text-only method outperforms the state-of-the-art, including multimodal methods, on the DAIC-WOZ depression benchmark.

- I investigated the benefits of leveraging a graph representation to learn a mapping of a long text. I exploited graphs to learn the relative importance of word-word connections to capture semantic features from domain context. The results using both depression clinical transcripts and Alzheimer speech transcripts indicate that the proposed graph representation learning network can significantly improve the domain task of estimating a patient's clinical symptoms from their language.

- Sequence-based deep models require all text inputs to be represented in equal length vectors for training. If the length of a transcript input exceeds the default length, it should be split into multiple inputs, and the corresponding PHQ scores should become new ground truth labels assigned for each separate input of that transcript. This can lead to an erroneous diagnosis of a patient if one trains a model to learn from such samples because the original PHQ score is based on assessing a complete transcript.

  Moreover, the proposed deep graph learning model allows transcripts to be represented regardless of the length. This novel methodology addressed the limitation of learning variable-length text. I evaluated different text data and achieved the best performance compared to the state-of-the-art sequence-based deep models. My research at the nexus of NLP and DL contributes to the enhancement of automated medical decision-making for diagnosis.

- In chapter 6, I investigated the effectiveness of a 2D node embedding mechanism by comparing the performance of SGNN and GNN models with the same number of parameters. I evaluated the performance of the method across two domains: 1. Twitter sentiment classification; 2. AD binary detection.
- I explained the proposed deep model using its gradients. The generated examples can be explained by visualizing the model's hidden layers using either node-level features or graph-level features. In the depression severity prediction task and AD binary detection task, the proposed model generated explanations that can provide some clinical domain-related findings and these findings are consistent with the findings reported in the clinical domain literature.
- My research has empirically demonstrated the generality of the proposed method that outperforms the current state-of-the-art models on different text data. My research suggests directions for future work that we expect to further improve accuracy performance.

7.2 Limitations

More research is needed on the effects of the word processing step and how this could impact the regression step on the final performance, such as medical decision processing. I observed that routinely removing stop words from corpora emphasizes more informative words. However, this exclusion may inadvertently leave out elements that can be informative in a clinical context. Most importantly, aggressive removal of words can dramatically reduce a piece of word message that determines a patient's current status seriously. This substantial loss of data can adversely affect the performance of deep learning models. We suggest the importance of paying attention to this usually-overlooked step in the pipeline, particularly when applying this step to some particular domain areas, such as psychological medicine and sentiment analysis.

Diagnostic modelings for profiling depressive symptoms for detecting depression always use different assessment tools, different data sources (i.e., text, video or speech), and datasets.

Since there are no standards to define depression severity, it is difficult to compare different approaches based on their experimental results. Instead of utilizing a gold standard metric from self-report diagnostic scales, such as the PHQ, some other self-report measures should be taken into account for evaluating ML/DL models' performance. There is some evidence that indicates mismatches between a depression score and depressive symptoms in a patient's self-reporting. For example, a patient mentions I cannot even fathom happiness while reporting a PHQ-8 score just above the cutoff for mild depression. The lack of standardized definitions of depression requires to use multiple physicians' ratings as well as a variety of other clinical and non-clinical measures of depression. In turn, comparing errors across these metrics might also shed light on the future improvement of the automated medical decision-making system, which could better support corresponding domain experts.

Another limitation is the absence of expert judgment on the importance of keywords in the data determined by the SGNN model. While machine learning algorithms can identify patterns and relationships in the data, they do not have the same level of understanding and context as humans. As a result, experts in the field are needed to interpret the results and determine the relevance and the significance of specific keywords identified by the model.

7.3 Future Work

MDD is a growing problem that may cause people to commit suicide. Future research could better utilize longitudinal and temporal information such as depression scores across interview sessions that are weeks or months apart, therefore improving prediction performance.

As language functions play an important role in the detection of cognitive biases across different states of depression, clinical transcripts can assist in measuring disease with confidence. My hypothesis is that cognitive biases due to depression underlying an individual's opinions can be manifested from the language of depression. Cognitive biases could form an effective and robust framework that is invariant to personal biases underlying each individual's behavior. Thus they could be used to efficiently filter out individual differences, and focus on symptom similarities as the key features to assessing depression levels for each individual. Future work may consider incorporating insights from psychology into designing a model algorithm. For instance, training a deep learning model to represent concept-based depressive features, such as cognitive biases, that capture features from the context of depression. Using psychological heuristics to test a psychological hypothesis regarding cognitive bias can help develop a common language that draws together the fields of psychological medicine, automated medical decision-making and NLP.

Another direction for further work is to explore the combination of transformers and graphs for learning node representations. Transformer-based language models, such as BERT, have shown great success in a variety of natural language processing tasks, but they are not designed to capture the structural relationships between nodes in a graph. Graph embeddings, on the other hand, are specifically designed to represent the relationships between nodes in a graph, but they do not have the same ability to capture the meaning and context of the nodes. By combining transformers and graph embeddings, it may be possible to improve the performance of inductive representation learning on large graphs. For example, the work of Dai et al. (2022) applies a way of incorporating dependency trees into a transformer-based machine translation model. Their work improves the quality of machine translation. This shows the potential for combining transformers and graphs for inductive representation learning on large graphs and suggests that this is an important direction for further research.

Moreover, understanding why the model made predictions could also be valuable. In the domain of mental health, work on explainable AI has started to emerge in importance to address the problem of trustworthy of AI systems. Future work should employ adequate explainability in the model, for example, proposing an explainable AI model for measuring depression states using attention mechanisms. It may provide a specific way of delivering the explainability of the learning algorithm with inter and intra-feature attentions that capture the relative importance between different feature classes as well as the relative importance of features within a class. These can be used to provide explanations of the model's conclusions.

7.4 Social Impact

Knowing all possible facts and data is not knowing the meaning of a certain situation. Big data and machine learning approaches are not only limited to efficiently producing a promising result but also help humans explore the meaning behind those data and facts. This motivated us to explore a possible way of enhancing both the interpretability and explainability of deep learning systems. This research work shared insights on leveraging GNN baseline models. I found that deep graph networks operating over word entities and relations provide a straightforward interface to produce structure behaviors. As a result, the decision-making process of GNN models becomes straightforward – perhaps more straightforward in understanding the reasoning behind its decisions – than general DL models.

The area of mental health needs both interpretable learning process of models and explicable results if we aim to build an automated diagnostic system deployed in practice. The main concern in this area is that if AI systems make illogical decisions and are not able to explain their decision-making. So it brings us to the question of how to interface a deep learning model with domain knowledge to solve that domain problem. Thus we could improve the explainability of the deep model such as why certain predictions were made.

To better align machine learning technology with domain challenges, it is important to understand how machine learning algorithms actually work. For instance, in healthcare applications, questions of accountability and transparency are particularly significant. If we aim to deploy artificial intelligence and deep learning systems to such areas, we have to properly deliver enhanced interpretability and ultimately explainability in our algorithms to address potential limitations of artificial intelligence.

Further, automatic detection of MDD using a single modality or multiple modalities is not new. My research focuses on modeling depression as a continuous rather than binary outcome, and models might detect specific symptoms in addition to detecting depression as an overall construct. Diversity in how mental distress or cognitive disease is expressed in many literature reviews (Beck et al., 1961; Kroenke et al., 2001). Apart from the absence of a gold standard, model performance and errors should be evaluated in depth. For example, there might be consistent types of symptoms or depression experiences not being detected. It is possible that some linguistic features are better predictors of depressive symptoms (or types of depression) than audio/visual features (De Choudhury et al., 2013b; Williamson et al., 2016).

We need to review our understanding of mental health and what we are exactly detecting. And we need to think about how to develop predictive models that incorporate the uncertainty in our understanding of depression and other clinical and non-clinical measures of depression. Research efforts can then turn to realize the vision that underpins these models: their deployment for early, scalable, and low-burden intervention and diagnosis of depression.

References

Adam-Troian, J., & Arciszewski, T. (2020). Absolutist words from search volume data predict state-level suicide rates in the united states. *Clinical Psychological Science*, *8*(4), 788–793.

Adel, H., & Shi, S. (2021). Proceedings of the 2021 conference on empirical methods in natural language processing: System demonstrations. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.*

Ahmed, U., Mukhiya, S. K., Srivastava, G., Lamo, Y., & Lin, J. C.-W. (2021). Attention-based deep entropy active learning using lexical algorithm for mental health treatment. *Frontiers in Psychology*, *12*, 642347.

Al Hanai, T., Ghassemi, M. M., & Glass, J. R. (2018). Detecting depression with audio/text sequence modeling of interviews. *Interspeech*, 1716–1720.

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Gedeon, T., Breakspear, M., & Parker, G. (2013). A comparative study of different classifiers for detecting depression from spontaneous speech. *2013 IEEE international conference on acoustics, speech and signal processing*, 8022–8026.

Alhanai, T., Ghassemi, M., & Glass, J. (2018). Detecting depression with Audio/Text sequence modeling of interviews. *Interspeech*, 1716–1720.

Al-Mosaiwi, M. (2018). People with depression use language differently–here's how to spot it. *Retrieved form http://theconversation. com/people-with-depression-use-language-differently-heres-how-to-spot-it-90877.*

Al-Mosaiwi, M., & Johnstone, T. (2018). In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, *6*(4), 529–542.

American Psychiatric Association, DSM-5 Task Force. (2013). *Diagnostic and statistical manual of mental disorders, 5th edition* (5th ed.). Washington, DC: American psychiatric association.

Andrews, G., Brugha, T., Thase, M. E., Duffy, F. F., Rucci, P., & Slade, T. (2007). Dimensionality and the category of major depressive episode. *International journal of methods in psychiatric research*, *16*(S1), S41–S51.

Arseniev-Koehler, A., Mozgai, S., & Scherer, S. (2018). What type of happiness are you looking for? - A closer look at detecting mental health from language. *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 1–12.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Bakioğlu, F., Korkmaz, O., & Ercan, H. (2021). Fear of COVID-19 and positivity: Mediating role of intolerance of uncertainty, depression, anxiety, and stress. *International journal of mental health and addiction*, *19*(6), 2369–2382.

Balani, S., & De Choudhury, M. (2015). Detecting and characterizing mental health related self-disclosure in social media. *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 1373–1378.

Bartczak, M., & Bokus, B. (2017). Semantic distances in depression: Relations between ME and PAST, FUTURE, JOY, SADNESS, HAPPINESS. *Journal of Psycholinguistic Research*, *46*(2), 345–366.

Bathina, K. C., Ten Thij, M., Lorenzo-Luaces, L., Rutter, L. A., & Bollen, J. (2021). Individuals with depression express more distorted thinking on social media. *Nature human behaviour*, *5*(4), 458–466.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al. (2018). Relational inductive biases, deep learning, and graph networks. arxiv 2018. *arXiv preprint arXiv:1806.01261*.

Beck, A. T. (2002). Cognitive models of depression. *Clinical advances in cognitive psychotherapy: Theory and application*, *14*(1), 29–61.

Beck, A. T. (1979). *Cognitive therapy and the emotional disorders*. Penguin.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of general psychiatry*, *4*(6), 561–571.

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of neurology*, *51*(6), 585–594.

Beckham, E. E., Leber, W. R., Watkins, J. T., Boyer, J. L., & Cook, J. B. (1986). Development of an instrument to measure Beck's cognitive triad: The Cognitive Triad Inventory. *Journal of consulting and clinical psychology*, *54*(4), 566.

Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. *Advances in neural information processing systems*, *13*.

Birjali, M., Beni-Hssane, A., & Erritali, M. (2017). Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. *Procedia Computer Science*, *113*, 65–72.

Blanco, I., & Joormann, J. (2017). Examining facets of depression and social anxiety: The relation among lack of positive affect, negative cognitions, and emotion dysregulation. *The Spanish Journal of Psychology*, *20*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Brewin, C. R., Smith, A. J., Power, M., & Furnham, A. (1992). State and trait differences in depressive self-perceptions. *Behaviour research and therapy*, *30*(5), 555–557.

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational linguistics*, *32*(1), 13–47.

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

Burnard, P. (1991). A method of analysing interview transcripts in qualitative research. *Nurse education today*, *11*(6), 461–466.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*(4), 335–359.

Camacho-Collados, J., & Pilehvar, M. T. (2017). On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. *In Proc. EMNLP BlackboxNLP. ACL.*

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, *538*(7623), 20.

Chen, J., Zhu, J., & Ye, J. (2019). An attention-based hybrid network for automatic detection of Alzheimer's disease from narrative speech. *Interspeech*, 4085–4089.

Chen, M., Xiao, X., Zhang, B., Liu, X., & Lu, R. (2022). Neural architecture searching for facial attributes-based depression recognition. *arXiv preprint arXiv:2201.09799*.

Chen, Y. (2015). *Convolutional neural network for sentence classification* (Master's thesis). University of Waterloo.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Clough, J. R., Oksuz, I., Puyol-Antón, E., Ruijsink, B., King, A. P., & Schnabel, J. A. (2019). Global and local interpretability for cardiac MRI classification. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 656–664.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, *12*(ARTICLE), 2493–2537.

Croisile, B., Ska, B., Brabant, M.-J., Duchene, A., Lepage, Y., Aimard, G., & Trillet, M. (1996). Comparative study of oral and written picture description in patients with alzheimer's disease. *Brain and language*, *53*(1), 1–19.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., Initiative, A. D. N., et al. (2011). Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *neuroimage*, *56*(2), 766–781.

Cummins, N., Joshi, J., Dhall, A., Sethu, V., Goecke, R., & Epps, J. (2013). Diagnosis of depression by behavioural signals: A multimodal approach. *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 11–20.

Cummins, N., Sethu, V., Epps, J., Schnieder, S., & Krajewski, J. (2015). Analysis of acoustic space variability in speech affected by depression. *Speech Communication*, *75*, 27–49.

Cummins, N., Sethu, V., Epps, J., Williamson, J. R., Quatieri, T. F., & Krajewski, J. (2017). Generalized two-stage rank regression framework for depression score prediction from speech. *IEEE Transactions on Affective Computing*, *11*(2), 272–283.

Dabkowski, P., & Gal, Y. (2017). Real time image saliency for black box classifiers. *Advances in neural information processing systems*, *30*.

Dai, Y., de Kamps, M., & Sharoff, S. (2022). BERTology for machine translation: What BERT knows about linguistic difficulties for translation. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 6674–6690.

De Choudhury, M., Counts, S., & Horvitz, E. (2013a). Social media as a measurement tool of depression in populations. *Proceedings of the 5th annual ACM web science conference*, 47–56.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013b). Predicting depression via social media. *Seventh international AAAI conference on weblogs and social media.*

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, *29*.

Del, M., & Fishel, M. (2021). Establishing interlingua in multilingual language models. *arXiv preprint arXiv:2109.01207.*

Delahunty, F., Wood, I. D., & Arcan, M. (2018). First insights on a passive major depressive disorder prediction system with incorporated conversational chatbot. *AICS*, 327–338.

Deshpande, M., & Rao, V. (2017). Depression detection using emotion artificial intelligence. *2017 international conference on intelligent sustainable systems (ICISS)*, 858–862.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Dham, S., Sharma, A., & Dhall, A. (2017). Depression scale recognition from audio, visual and text analysis. *arXiv preprint arXiv:1709.05865.*

Dillon, D. G., & Pizzagalli, D. A. (2018). Mechanisms of memory disruption in depression. *Trends in neurosciences*, *41*(3), 137–149.

D'mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, *47*(3), 1–36.

Dozois, D. J., & Beck, A. T. (2008). Cognitive schemas, beliefs and assumptions. *Risk factors in depression*, 119–143.

Du, M., Liu, N., Song, Q., & Hu, X. (2018). Towards explanation of DNN-based prediction with guided feature inversion. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1358–1367.

Du, Z., Li, W., Huang, D., & Wang, Y. (2019). Encoding visual behaviors with attentive temporal convolution for depression prediction. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1–7.

Durkheim, E. (2005). *Suicide: A study in sociology*. Routledge.

Ekman, P., & Rosenberg, E. L. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)*. Oxford University Press, USA.

Falagas, M., Vardakas, K., & Vergidis, P. (2007). Under-diagnosis of common chronic diseases: Prevalence and impact on human health. *International journal of clinical practice*, *61*(9), 1569–1579.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. *International conference on machine learning*, 1263–1272.

Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S. M., Hammal, Z., & Rosenwald, D. P. (2014). Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, *32*(10), 641–647.

Gladstone, G., Gladstone, G., & Parker, G. (2001). Depressogenic cognitive schemas: Enduring beliefs or mood state artefacts? *Australian & New Zealand Journal of Psychiatry*, *35*(2), 210–216.

Gong, Y., & Poellabauer, C. (2017). Topic modeling based multi-modal depression detection. *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 69–76.

Gorwood, P., Corruble, E., Falissard, B., & Goodwin D Phil, G. M., F Med Sci. (2008). Toxic effects of depression on brain function: Impairment of delayed recall and the cumulative length of depressive disorder in a large sample of depressed outpatients. *American Journal of Psychiatry*, *165*(6), 731–739.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., et al. (2014). The distress analysis interview corpus of human and computer interviews. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 3123–3128.

Groves-Wright, K., Neils-Strunjas, J., Burnett, R., & O'Neill, M. J. (2004). A comparison of verbal and written language in alzheimer's disease. *Journal of Communication Disorders*, *37*(2), 109–130.

HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PloS one*, *15*(5), e0232525.

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, *30*.

Hammen, C., & Zupan, B. A. (1984). Self-schemas, depression, and the processing of personal information in children. *Journal of Experimental Child Psychology*, *37*(3), 598–608.

Haque, A., Guo, M., Miner, A. S., & Fei-Fei, L. (2018). Measuring depression symptom severity from spoken language and 3D facial expressions. *arXiv preprint arXiv:1811.08592*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Holtzman, N. S. et al. (2017). A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality*, *68*, 63–68.

Hosseini-Asl, E., Keynton, R., & El-Baz, A. (2016). Alzheimer's disease diagnostics by adaptation of 3D convolutional network. *2016 IEEE international conference on image processing (ICIP)*, 126–130.

Islam, M., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., Ulhaq, A., et al. (2018). Depression detection from social network data using machine learning techniques. *Health information science and systems*, *6*(1), 1–12.

Islam, M. S., Tasnim, R., Sujan, M. S. H., Ferdous, M. Z., Sikder, M. T., Masud, J. H. B., Kundu, S., Tahsin, P., Mosaddek, A. S. M., & Griffiths, M. D. (2021). Depressive symptoms associated with COVID-19 preventive practice measures, daily activities in home quarantine and suicidal behaviors: Findings from a large-scale online survey in Bangladesh. *BMC psychiatry*, *21*(1), 1–12.

Jackson, R. G., Patel, R., Jayatilleke, N., Kolliakou, A., Ball, M., Gorrell, G., Roberts, A., Dobson, R. J., & Stewart, R. (2017). Natural language processing to extract symptoms of severe mental illness from clinical text: The Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ open*, *7*(1), e012012.

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., & Ogar, J. (2014). Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 27–37.

Johanson, S., & Bejerholm, U. (2017). The role of empowerment and quality of life in depression severity among unemployed people with affective disorders receiving mental healthcare. *Disability and rehabilitation*, *39*(18), 1807–1813.

Joshi, G., Wozniak, J., Petty, C., Martelon, M. K., Fried, R., Bolfek, A., Kotte, A., Stevens, J., Furtak, S. L., Bourgeois, M., et al. (2013). Psychiatric comorbidity and functioning in a clinically referred population of adults with autism spectrum disorders: A comparative study. *Journal of autism and developmental disorders*, *43*(6), 1314–1325.

Kaplan, E. (1983). *Boston diagnostic aphasia examination booklet*. Lea & Febiger Philadelphia, PA.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N., & Welling, M. (2016a). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kipf, T. N., & Welling, M. (2016b). Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

Korobkin, S. B., Herron, W. G., & Ramirez, S. M. (1998). Severity of symptoms of depression and anxiety as predictors of duration of psychotherapy. *Psychological reports*, *82*(2), 427–433.

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of general internal medicine*, *16*(9), 606–613.

Kroenke, K., Spitzer, R. L., Williams, J. B., & Löwe, B. (2010). The patient health questionnaire somatic, anxiety, and depressive symptom scales: A systematic review. *General hospital psychiatry*, *32*(4), 345–359.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of affective disorders*, *114*(1-3), 163–173.

Kumawat, S., Yadav, I., Pahal, N., & Goel, D. (2021). Sentiment analysis using language models: A study. *In Proceedings of the 11th International Conference on Cloud Computing, Data Science and Engineering (Confluence)*, 984–988.

Lam, G., Dongyan, H., & Lin, W. (2019). Context-aware deep learning for multi-modal depression detection. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3946–3950.

Lewinsohn, P. M., Solomon, A., Seeley, J. R., & Zeiss, A. (2000). Clinical implications of "subthreshold" depressive symptoms. *Journal of abnormal psychology*, *109*(2), 345.

Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2015). Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066.*

Li, Y., Masitah, A., & Hills, T. T. (2020). The emotional recall task: Juxtaposing recall and recognition-based affect scales. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(9), 1782.

Liang, Y., Meng, F., Zhang, Y., Chen, Y., Xu, J., & Zhou, J. (2022). Emotional conversation generation with heterogeneous graph neural network. *Artificial Intelligence*, *308*, 103714.

Liao, W., Zeng, B., Liu, J., Wei, P., Cheng, X., & Zhang, W. (2021). Multi-level graph neural network for text sentiment analysis. *Computers & Electrical Engineering*, *92*, 107096.

Lin, L., Chen, X., Shen, Y., & Zhang, L. (2020). Towards automatic depression detection: A BiLSTM/1D CNN-based model. *Applied Sciences*, *10*(23), 8701.

Lin, W., Ji, D., & Lu, Y. (2017). Disorder recognition in clinical texts using multi-label structured SVM. *BMC bioinformatics*, *18*(1), 1–11.

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour research and therapy*, *33*(3), 335–343.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025.*

MacQueen, G. M., Galway, T., Hay, J., Young, L., & Joffe, R. (2002). Recollection memory deficits in patients with major depressive disorder predicted by past depressions but not current mood state or treatment status. *Psychological medicine*, *32*(2), 251–258.

MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.

Mallol-Ragolta, A., Zhao, Z., Stappen, L., Cummins, N., & Schuller, B. (2019). A hierarchical attention network-based approach for depression detection from transcribed clinical interviews.

Maron, H., Ben-Hamu, H., Shamir, N., & Lipman, Y. (2018). Invariant and equivariant graph networks. *arXiv preprint arXiv:1812.09902*.

McDermott, L. M., & Ebmeier, K. P. (2009). A meta-analysis of depression severity and cognitive function. *Journal of affective disorders*, *119*(1-3), 1–8.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111–3119.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and cognitive processes*, *6*(1), 1–28.

Millington, T., & Luz, S. (2021). Analysis and classification of word co-occurrence networks from Alzheimer's patients and controls. *Frontiers in Computer Science*, *3*, 649508.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Moore, M. T., & Fresco, D. M. (2007). Depressive realism and attributional style: Implications for individuals at risk for depression. *Behavior Therapy*, *38*(2), 144–154.

Morales, M., Scherer, S., & Levitan, R. (2017). A cross-modal review of indicators for depression detection systems. *Proceedings of the fourth workshop on computational linguistics and clinical psychology—From linguistic signal to clinical reality*, 1–12.

Morales, M. R., & Levitan, R. (2016). Speech vs. text: A comparative analysis of features for depression detection systems. *2016 IEEE spoken language technology workshop (SLT)*, 136–143.

Mukherjee, S. S., Yu, J., Won, Y., McClay, M. J., Wang, L., Rush, A. J., & Sarkar, J. (2020). Natural language processing-based quantification of the mental state of psychiatric patients. *Computational Psychiatry*, *4*.

Mundt, J. C., Vogel, A. P., Feltner, D. E., & Lenderking, W. R. (2012). Vocal acoustic biomarkers of depression severity and treatment response. *Biological psychiatry*, *72*(7), 580–587.

Naseem, U., Khan, S. K., Razzak, I., & Hameed, I. A. (2019). Hybrid words representation for airlines sentiment analysis. *Australasian Joint Conference on Artificial Intelligence*, 381–392.

Naseem, U., Razzak, I., Musial, K., & Imran, M. (2020). Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems*, *113*, 58–69.

Nebes, R. D. (1989). Semantic memory in Alzheimer's disease. *Psychological bulletin*, *106*(3), 377.

Nguyen, T., Phung, D., Dao, B., Venkatesh, S., & Berk, M. (2014). Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, *5*(3), 217–226.

Nicholas, M., Obler, L. K., Albert, M. L., & Helm-Estabrooks, N. (1985). Empty speech in Alzheimer's disease and fluent aphasia. *Journal of Speech, Language, and Hearing Research*, *28*(3), 405–410.

Niepert, M., Ahmed, M., & Kutzkov, K. (2016). Learning convolutional neural networks for graphs. *International conference on machine learning*, 2014–2023.

Oh, K., Chung, Y.-C., Kim, K. W., Kim, W.-S., & Oh, I.-S. (2019). Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning. *Scientific Reports*, *9*(1), 1–16.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, *2*(11), e7.

Orimaye, S. O., Wong, J. S.-M., & Golden, K. J. (2014). Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, 78–87.

Parker, G., Gladstone, G., Roussos, J., Wilhelm, K., Mitchell, P., Hadzi-Pavlovic, D., Austin, M.-P., & Hickie, I. (1998). Qualitative and quantitative analyses of a 'lock and key' hypothesis of depression. *Psychological medicine*, *28*(6), 1263–1273.

Pellert, M., Metzler, H., Matzenberger, M., & Garcia, D. (2022). Validating daily social media macroscopes of emotions. *Scientific reports*, *12*(1), 1–8.

Peng, H., Li, J., He, Y., Liu, Y., Bao, M., Wang, L., Song, Y., & Yang, Q. (2018). Large-scale hierarchical text classification with recursively regularized deep graph-cnn. *Proceedings of the 2018 world wide web conference*, 1063–1072.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, *54*(1), 547–577.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237.

Poulin, C., Shiner, B., Thompson, P., Vepstas, L., Young-Xu, Y., Goertzel, B., Watts, B., Flashman, L., & McAllister, T. (2014). Predicting the risk of suicide by analyzing the text of clinical notes. *PloS one*, *9*(1), e85733.

Pyszczynski, T., & Greenberg, J. (1987). Self-regulatory perseveration and the depressive self-focusing style: A self-awareness theory of reactive depression. *Psychological bulletin*, *102*(1), 122.

Quatieri, T. F., Williamson, J. R., Smalt, C. J., Patel, T., Perricone, J., Mehta, D. D., Helfer, B. S., Ciccarelli, G., Ricke, D., Malyska, N., et al. (2015). Vocal biomarkers to discriminate cognitive load in a working memory task. *Sixteenth Annual Conference of the International Speech Communication Association*.

Ray, A., Kumar, S., Reddy, R., Mukherjee, P., & Garg, R. (2019). Multi-level attention network using text, audio and video for depression prediction. *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 81–88.

Revonsuo, A., Portin, R., Juottonen, K., & Rinne, J. O. (1998). Semantic processing of spoken words in Alzheimer's disease: An electrophysiological study. *Journal of Cognitive Neuroscience*, *10*(3), 408–420.

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., Schmitt, M., Alisamir, S., Amiriparian, S., Messner, E.-M., et al. (2019). Avec 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition. *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, 3–12.

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., & Pantic, M. (2017). Avec 2017: Real-life depression, and affect recognition workshop and challenge. *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 3–9.

Riskind, J. H., Castellon, C. S., & Beck, A. T. (1989). Spontaneous causal explanations in unipolar depression and generalized anxiety: Content analysis of dysfunctional-thought diaries. *Cognitive Therapy and Research*, *13*(2), 97–108.

Rohrer, J. D., Knight, W. D., Warren, J. E., Fox, N. C., Rossor, M. N., & Warren, J. D. (2008). Word-finding difficulty: A clinical analysis of the progressive aphasias. *Brain*, *131*(1), 8–38.

Rottenberg, J. (2017). Emotions in depression: What do we really know. *Annual Review of Clinical Psychology*, *13*(1), 241–263.

Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, *18*(8), 1121–1133.

Rudolph, K. D., Hammen, C., & Burge, D. (1997). A cognitive-interpersonal approach to depressive symptoms in preadolescent children. *Journal of abnormal child psychology*, *25*(1), 33–45.

Santomauro, D. F., Herrera, A. M. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., Abbafati, C., Adolph, C., Amlag, J. O., Aravkin, A. Y., et al. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet*, *398*(10312), 1700–1712.

Sarker, I. H. (2021). Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, *2*(6), 1–20.

Scherer, S., Stratou, G., Mahmoud, M., Boberg, J., Gratch, J., Rizzo, A., & Morency, L.-P. (2013). Automatic behavior descriptors for psychological disorder analysis. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 1–8.

Schoene, A. M., & Dethlefs, N. (2016). Automatic identification of suicide notes from linguistic and sentiment features. *Proceedings of the 10th SIGHUM workshop on language technology for cultural heritage, social sciences, and humanities*, 128–133.

Searle, T., Ibrahim, Z., & Dobson, R. (2020). Comparing natural language processing techniques for Alzheimer's dementia prediction in spontaneous speech. *arXiv preprint arXiv:2006.07358*.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.-S., & Zhu, W. (2017). Depression detection via harvesting social media: A multimodal dictionary learning solution. *IJCAI*, 3838–3844.

Shestyuk, A. Y., & Deldin, P. J. (2010). Automatic and strategic representation of the self in major depression: Trait and state abnormalities. *American Journal of Psychiatry*, *167*(5), 536–544.

Shickel, B., Siegel, S., Heesacker, M., Benton, S., & Rashidi, P. (2020). Automatic detection and classification of cognitive distortions in mental health text. *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, 275–280.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Socher, R., Lin, C. C., Manning, C., & Ng, A. Y. (2011). Parsing natural scenes and natural language with recursive neural networks. *Proceedings of the 28th international conference on machine learning (ICML-11)*, 129–136.

Solieman, H., & Pustozerov, E. A. (2021). The detection of depression using multimodal models based on text and voice quality features. *2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, 1843–1848.

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

Soygüt, G., & Savaşir, I. (2001). The relationship between interpersonal schemas and depressive symptomatology. *Journal of Counseling Psychology*, *48*(3), 359.

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine*, *63*(4), 517–522.

Stokes, L., Combes, H., & Stokes, G. (2015). The dementia diagnosis: A literature review of information, understanding, and attributions. *Psychogeriatrics*, *15*(3), 218–225.

Stolar, M. N., Lech, M., & Allen, N. B. (2015). Detection of depression in adolescents based on statistical modeling of emotional influences in parent-adolescent conversations. *2015*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 987–991.

Suk, H.-I., Lee, S.-W., Shen, D., Initiative, A. D. N., et al. (2014). Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage*, *101*, 569–582.

Sun, B., Zhang, Y., He, J., Yu, L., Xu, Q., Li, D., & Wang, Z. (2017). A random forest regression method with selected-text feature for depression assessment. *Proceedings of the 7th annual workshop on Audio/Visual emotion challenge*, 61–68.

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of suicide ideation in social media forums using deep learning. *Algorithms*, *13*(1), 7.

Tang, H., Mi, Y., Xue, F., & Cao, Y. (2020). An integration model based on graph convolutional network for text classification. *IEEE Access*, *8*, 148865–148876.

Terechshenko, Z., Linder, F., Padmakumar, V., Liu, M., Nagler, J., Tucker, J. A., & Bonneau, R. (2020). A comparison of methods in political science text classification: Transfer learning language models for politics. *Available at SSRN 3724644*.

Thombs, B. D., Benedetti, A., Kloda, L. A., Levis, B., Nicolau, I., Cuijpers, P., Gilbody, S., Ioannidis, J. P., McMillan, D., Patten, S. B., et al. (2014). The diagnostic accuracy of the patient health questionnaire-2 (PHQ-2), patient health questionnaire-8 (PHQ-8), and patient health questionnaire-9 (PHQ-9) for detecting major depression: Protocol for a systematic review and individual patient data meta-analyses. *Systematic reviews*, *3*(1), 1–16.

Trevino, A. C., Quatieri, T. F., & Malyska, N. (2011). Phonologically-based biomarkers for major depressive disorder. *EURASIP Journal on Advances in Signal Processing*, *2011*(1), 1–18.

Trifan, A., Antunes, R., Matos, S., & Oliveira, J. L. (2020). Understanding depression from psycholinguistic patterns in social media texts. *European Conference on Information Retrieval*, 402–409.

Tsakalidis, A., Liakata, M., Damoulas, T., & Cristea, A. I. (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*, 407–423.

Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., & Ohsaki, H. (2015). Recognizing depression from twitter activity. *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 3187–3196.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, *37*, 141–188.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., & Pantic, M. (2016). Avec 2016: Depression, mood, and emotion recognition workshop and challenge. *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 3–10.

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., & Pantic, M. (2014). Avec 2014: 3d dimensional affect and depression recognition challenge. *Proceedings of the 4th international workshop on audio/visual emotion challenge*, 3–10.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., & Pantic, M. (2013). Avec 2013: The continuous audio/visual emotion and depression recognition challenge. *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 3–10.

van Borkulo, C., Boschloo, L., Borsboom, D., Penninx, B. W., Waldorp, L. J., & Schoevers, R. A. (2015). Association of symptom network structure with the course of depression. *JAMA psychiatry*, *72*(12), 1219–1226.

van der Maaten, L. (2014). Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research*, *15*, 3221–3245.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Van Loo, H. M., De Jonge, P., Romeijn, J.-W., Kessler, R. C., & Schoevers, R. A. (2012). Data-driven subtypes of major depressive disorder: A systematic review. *BMC medicine*, *10*(1), 1–12.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wang, G., Li, C., Wang, W., Zhang, Y., Shen, D., Zhang, X., Henao, R., & Carin, L. (2018). Joint embedding of words and labels for text classification. *arXiv preprint arXiv:1805.04174*.

Wang, H., & Li, F. (2022). A text classification method based on LSTM and graph attention network. *Connection Science*, *34*(1), 2466–2480.

Wang, K., Zhang, Y., Yang, D., Song, L., & Qin, T. (2021). GNN is a counter? revisiting GNN for question answering. *arXiv preprint arXiv:2110.03192.*

Wang, N., Luo, F., Peddagangireddy, V., Subbalakshmi, K. P., & Chandramouli, R. (2020). Personalized Early Stage Alzheimer's Disease detection: A case study of President Reagan's Speeches. *arXiv preprint arXiv:2005.12385.*

Weeds, J., & Weir, D. (2005). Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, *31*(4), 439–475.

Wichers, M., Groot, P. C., Psychosystems, E., Group, E., et al. (2016). Critical slowing down as a personalized early warning signal for depression. *Psychotherapy and psychosomatics*, *85*(2), 114–116.

Williamson, J. R., Quatieri, T. F., Helfer, B. S., Horwitz, R., Yu, B., & Mehta, D. D. (2013). Vocal biomarkers of depression based on motor incoordination. *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 41–48.

Williamson, J. R., Schwarzentruber, A., Kung, H., Godoy, E., Khorrami, P., Dagli, C., Cha, M., Gwon, Y., & Quatieri, T. F. (2016). Detecting depression using vocal, facial and semantic communication cues. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 11–18.

World Health Organization. (2017). *Depression and other common mental disorders: Global health estimates* (tech. rep.). World Health Organization.

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). How powerful are graph neural networks? *International Conference on Learning Representations.*

Yang, L., Jiang, D., & Sahli, H. (2020). Feature augmenting networks for improving depression severity estimation from speech signals. *IEEE Access*, *8*, 24033–24045.

Yang, Y., Fairbairn, C., & Cohn, J. F. (2012). Detecting depression severity from vocal prosody. *IEEE transactions on affective computing*, *4*(2), 142–150.

Yao, L., Mao, C., & Luo, Y. (2019). Graph convolutional networks for text classification. *Proceedings of the AAAI conference on artificial intelligence*, *33*(01), 7370–7377.

Yasunaga, M., Ren, H., Bosselut, A., Liang, P., & Leskovec, J. (2021). QA-GNN: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378.*

Yenduri, G., Rajakumar, B., Praghash, K., & Binu, D. (2021). Heuristic-assisted BERT for twitter sentiment analysis. *International Journal of Computational Intelligence and Applications*, *20*(03), 2150015.

Ying, R., You, J., Morris, C., Ren, X., Hamilton, W. L., & Leskovec, J. (2018). Hierarchical graph representation learning with differentiable pooling. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 4805–4815.

Young, J. E. (1999a). *Cognitive therapy for personality disorders: A schema-focused approach.* Professional Resource Press/Professional Resource Exchange.

Young, J. E. (1999b). *Cognitive therapy for personality disorders: A schema-focused approach.* Professional Resource Press/Professional Resource Exchange.

Young, J. E., & Lindemann, M. D. (1992). An integrative schema-focused model for personality disorders. *Journal of Cognitive Psychotherapy*, *6*(1), 11.

Zhang, T., Gong, X., & Chen, C. P. (2021). BMT-Net: Broad multitask transformer network for sentiment analysis. *IEEE Transactions on Cybernetics.*

Zhang, Y., Wang, Y., Wang, X., Zou, B., & Xie, H. (2020). Text-based decision fusion model for detecting depression. *2020 2nd Symposium on Signal Processing Systems*, 101–106.

Zhang, Z., Cui, P., & Zhu, W. (2020). Deep learning on graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering.*

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.

Zhu, Y., Liang, X., Batsis, J. A., & Roth, R. M. (2021). Exploring deep transfer learning techniques for alzheimer's dementia detection. *Frontiers in computer science*, 22.

Zimmermann, J., Brockmeyer, T., Hunn, M., Schauenburg, H., & Wolf, M. (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*(2), 384–391.