Automated depression level estimation: a study on discourse structure, input representation and clinical reliability

Navneet Agarwal

Normandie Université

UNIVERSITÉ CAEN NORMANDIE

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité **INFORMATIQUE**

Préparée au sein de l'**Université de Caen Normandie**

Automated Depression Level Estimation: A Study on Discourse Structure, Input Representation and Clinical Reliability

Présentée et soutenue par

NAVNEET AGARWAL

Thèse soutenue le 27/06/2024

devant le jury composé de :

M. GAEL DIAS	Professeur des universités - Université de Caen Normandie (UCN)	Directeur de thèse
M. MAXIME AMBLARD	Professeur des universités - Université de Lorraine	Président du jury
MME MUNMUM DE CHOUDHURY	Professeur - Institut de Technologie de Géorgie	Membre du jury
MME NATALIA GRABAR	Chargé de recherche au CNRS - UNIVERSITE LILLE 3 CHARLES DE GAULLE	Membre du jury
M. RAMA KRISHNA SAI SUBRAHMANYAM GORTHI	Professeur - Indian Institute of Technology	Membre du jury
MME KAIRIT SIRTS	Maître de conférences - Université de Tartu - Estonie	Membre du jury
M. ANTOINE DOUCET	Professeur des universités - UNIVERSITE LA ROCHELLE	Rapporteur du jury

Thèse dirigée par **GAEL DIAS** (Groupe de recherche en informatique, image et instrumentation de Caen)

MIIS

GREYC
Laboratoire de recherche en sciences du numérique

Résumé en Français

Étant donné l'impact négatif de la dépression dans les sociétés modernes, d'importantes initiatives de recherche ont été entreprises pour définir des systèmes de quantification automatisée de la dépression. S'appuyant sur la vaste littérature dans ce domaine basée sur des entretiens cliniques, cette thèse soulève trois questions majeures relqtivement peu explorées malgré leur pertinence: (1) le rôle de la structure du discours dans l'analyse de la santé mentale, (2) la pertinence de la représentation de l'entrée pour les capacités prédictives des modèles de réseaux neuronaux, et (3) l'importance de l'expertise médicale dans le domaine de la détection automatisée de la dépression.

La nature dyadique des entretiens patient-thérapeute garantit la présence d'une structure sous-jacente complexe au sein du discours. Néanmoins, la plupart des recherches négligent d'exploiter cette connaissance et traitent l'entrée comme une simple séquence de phrases, contraignant ainsi le modèle à comprendre les subtilités de la conversation à partir d'une séquence non structurée de phrases. Dans cette thèse, nous établissons d'abord l'importance des questions des thérapeutes dans l'entrée du modèle neuronal, avant de montrer qu'une combinaison séquentielle de l'entrée du patient et du thérapeute est une stratégie sous-optimale. En conséquence, des architectures Multi-vues sont proposées comme moyen d'incorporer la structure du discours dans le processus d'apprentissage des modèles neuronaux. Les résultats expérimentaux montrent les avantages des architectures Multi-vues proposées, validant la pertinence de conserver la structure du discours dans le processus d'entraînement du modèle. Des expériences sont menées avec deux stratégies d'encodage de texte différentes, l'encodage de texte hiérarchique et l'encodage de texte basé sur Sentence Transformer, pour établir davantage l'efficacité des architectures proposées dans le contexte de la tâche de classification binaire.

Ayant établi la nécessité de conserver la structure du discours dans le processus d'apprentissage, ainsi que les limites de l'encodage de texte séquentiel à cet égard, nous explorons davantage les représentations textuelles basées sur les graphes. Les

graphes fournissent non seulement une structure de données adaptée pour coder la structure complexe et non linéaire de la conversation, mais ouvrent également la possibilité de mettre en évidence des traits spécifiques des transcriptions d'entrée. La recherche menée dans ce contexte met en lumière l'impact des représentations d'entrée non seulement dans la définition des capacités d'apprentissage du modèle, mais aussi dans la compréhension de leur processus prédictif. Les graphes de similarité de phrases et les graphes de corrélation des mots sont utilisés pour illustrer la capacité des représentations graphiques à fournir des perspectives variées de la même entrée, mettant en évidence des informations qui peuvent non seulement améliorer les performances prédictives des modèles, mais qui peuvent également être pertinentes pour les professionnels de la santé. Le concept de multi-vues est également incorporé dans les deux structures de graphes pour mettre davantage en évidence les différences de perspectives entre le patient et le thérapeute au sein du même entretien. De plus, il est démontré que la visualisation des structures de graphes proposées peut fournir des informations précieuses indiquant des changements subtils dans le comportement du patient et du thérapeute, ce qui suggère l'état mental du patient.

Enfin, nous mettons en évidence le manque d'implication des professionnels de santé dans le contexte de la détection automatisée de la dépression basée sur les entretiens cliniques. Étant donné la nature interdisciplinaire de la tâche, la participation active des cliniciens peut jouer un rôle vital dans l'amélioration des capacités d'apprentissage des modèles d'IA et dans le renforcement de leur fiabilité et de leur acceptation en tant qu'outils prédictifs au sein des systèmes de santé. Dans le cadre de cette thèse, la tâche d'annotation clinique d'un corpus d'entretiens d'analyse de la détresse mentale a été entreprise pour fournir une ressource permettant de mener des recherches interdisciplinaires dans ce domaine. Des expériences sont définies pour étudier l'intégration des annotations cliniques dans les modèles neuronaux appliqués à la tâche de prédiction au niveau des symptômes. De plus, les modèles proposés sont analysés dans le contexte des annotations cliniques afin de les analogiser avec le processus prédictif et les tendances psychologiques des professionnels de la santé, une étape vers leur établissement en tant qu'outils cliniques fiables.

Tous les modèles présentés dans cette thèse sont également comparés aux initiatives récentes dans le domaine, ceux-ci offrant les mielleures performances sur le corpus de base utilizé. Nous fournissons également des annotations cliniques dudit ensemble de données pour encourager la recherche multidisciplinaire dans le domaine. La thèse se conclut par une description de nos initiatives de recherche en cours et futures visant à améliorer davantage le travail présenté dans cette dissertation.

Abstract

Given the severe and widespread impact of depression, significant research initiatives have been undertaken to define systems for automated depression assessment. Building upon the extensive literature concerning automated depression detection based on clinical interviews, this thesis raises three major questions. The research presented in this dissertation revolves around the following questions that remain relatively unexplored despite their relevance; (1) the role of discourse structure in mental health analysis, (2) the relevance of input representation towards the predictive abilities of neural network models, and (3) the importance of domain expertise in automated depression detection.

The dyadic nature of patient-therapist interviews ensures the presence of a complex underlying structure within the discourse. Nevertheless, most researchers fail to exploit this knowledge and treat the input as a sequence of sentences, forcing the model to understand the intricacies of the conversation from an unstructured sequence of sentences. Within this thesis, we first establish the importance of therapist questions within the neural network model's input, before showing that a sequential combination of patient and therapist input is a sub-optimal strategy. Consequently, Multi-view architectures are proposed as a means of incorporating the discourse structure within the learning process of neural networks. Experimental results show the advantages of the proposed multi-view architectures, validating the relevance of retaining discourse structure within the model's training process. Experiments are conducted with two different text encoding strategies, hierarchical text encoding and Sentence Transformer based text encoding, to further establish the effectiveness of the proposed architectures in the context of binary classification task within the depression estimation umbrella.

Having established the need to retain the discourse structure within the learning process, and the limitations of sequential text encoding in doing so, we further explore graph based text representations. Graphs not only provide a more optimal

data structure for encoding the complex non-linear structure of conversation, but also open up the possibility to highlight specific traits of the input transcripts. The research conducted in this context highlights the impact of input representations not only in defining the learning abilities of the model, but also in understanding their predictive process. Sentence Similarity Graphs and Keyword Correlation Graphs are used to exemplify the ability of graphical representations to provide varying perspectives of the same input, highlighting information that can not only improve the predictive performance of the models but can also be relevant for medical professionals. Multi-view concept is also incorporated within the two graph structures to further highlight the difference in the perspectives of the patient and the therapist within the same interview. Furthermore, it is shown that visualization of the proposed graph structures can provide valuable insights indicative of subtle changes in patient and therapist's behavior, hinting towards the mental state of the patient.

Finally, we highlight the lack of involvement of medical professionals within the context of automated depression detection based on clinical interviews. Given the interdisciplinary nature of the task, the active participation of clinicians can play a vital role in not only improving the learning ability of the automated models, but also bolstering their reliability and acceptance as predictive tools that can be deployed in healthcare systems. As part of this thesis, clinical annotations of the Distress Analysis Interview Corpus - Wizard of Oz [38] (DAIC-WOZ) dataset were performed to provide a resource for conducting interdisciplinary research in this field. Experiments are defined to study the integration of the clinical annotations within the neural network models applied to symptom-level prediction task within the automated depression detection domain. Furthermore, the proposed models are analyzed in the context of the clinical annotations in order to analogize their predictive process and psychological tendencies with those of medical professionals, a step towards establishing them as reliable clinical tools.

All the neural network models, architectures, and research presented in this thesis are also compared against recent initiatives in the field, with our proposed models providing new state-of-the-art performance evaluated on the test set of the DAIC-WOZ dataset. We also provide clinical annotations of the said dataset to encourage multi-disciplinary research in the field. The thesis concludes with a description of our ongoing and future research initiatives aimed at further improving the work presented within this dissertation.

Acknowledgment

First and foremost, I would like to express my deepest gratitude to my supervisor Dr. Gaël Dias. His confidence in my abilities and guidance have been invaluable throughout my research. I appreciate all his contributions, help and support during the course of my PhD. I consider myself fortunate to have had him as my supervisor, and cannot thank him enough for his mentorship and encouragement. Thanks to Dr. Sonia Dollfus, Dr. Fabrice Maurel, Dr. Alexis Lechervy and Dr. Youssef Chahir for all their research inputs, valuable discussions and guidance.

My sincere thanks to Dr. Kairit Sirts and Dr. Maxim Amblard for agreeing to be the reviewer and member of the jury for the defense. I would also like to thank Dr. Munmun De Choudhury, Dr. Natalia Grabar, and Dr. Rama Krishna Gorthi for agreeing to review this work and being members of the jury.

This thesis would not be possible without the financial support from the FHU A^2M^2P project, GREYC-CNRS UMR 6072 Laboratory and MIIS doctoral school. Thanks to the administrative staff within GREYC and MIIS doctoral school for their help with all the documents and missions. Special thanks to all the permanent and non-permanent members of GREYC for their direct and indirect contributions.

Thanks to Kirill Milintsevich, Lucie Metivier, and Dr. Maud Rotharmel for their contributions to my research. I would like to extend my special gratitude to Dr. Parameshwari, Guillaume, Julien, Stevan, Saurabh, Valentin, and Nilesh for their constant motivation and support in my professional and personal life. I would also like to thank the Indian community in Caen for all the memorable moments.

Finally, I'm grateful to my family Sanjeev Agarwal, Manjari Agarwal, Prateek and Niketa Agarwal. I would not be here without their love, support and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Mental health represents an integral part of an individual's ability to think, emote, interact with others, earn a living, and enjoy life in general. Consequently, mental health underpins the core human values of independent thought and action, happiness and friendship, and plays a vital role in defining the quality of our life. All over the world, mental, neurological, and substance disorders are common, affecting every community and age group across all income countries. In many Western countries, mental disorders are the leading cause of disability, responsible for 30-40% of chronic sick leaves and costing almost 3% of GDP[1]. In particular, approximately 25% of the population is affected by them in the European region[2]. According to global health estimates for the European region in 2019, the number of people with mental health conditions (including depression, anxiety disorders and psychosis in adults, as well as developmental and behavioral disorders in children and adolescents) stood at over 125 million, equivalent to 13% of the population. Furthermore, mental health conditions account for 15% of all years lived with a disability in this region. Additionally, it is estimated that 119,000 lives were lost due to suicide in this region in 2019 alone, representing an unacceptable figure including an increasing number of young people[3].

Depression is one of the most prevalent mental disorders that affects millions of people worldwide. According to statistics from the World Health Organization

[1] https://www.who.int/europe/health-topics/mental-health#tab=tab_2

[2] World Health Organization, "The European Mental Health Action Plan 2013–2020," 2015. [Online]. Available: https://bit.ly/2UvIQi6

[3] Global health estimates: life expectancy and leading causes of death and disability. In: The Global Health Observatory [online database]. Geneva: World Health Organization; 2019 (https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates, accessed 16 August 2021)

(WHO)[4], approximately 5% of adults worldwide experienced depression in 2019, amounting to an estimated 280 million people. The lifetime prevalence reports show high variance, with 3% reported in Japan to 17% in the United States. Within the US alone, depression affects more than 27 million people and is believed to be the cause of 30,000 suicides each year [21, 36, 54]. In North America, the probability of having a major depressive episode within a period of one year is 3–5% for males and 8–10% for females [3, 22].

Similarly to other health aspects, mental health and well-being of a person are also affected by a wide range of individual, social, and environmental factors, including poverty and deprivation; debt and unemployment; and violence and conflict. As demonstrated by the economic recession following the financial crisis in 2008 and by the SARS-CoV-2 (COVID-19) pandemic starting in 2020, the mental health of both individuals and populations can also be undermined by macroeconomic forces or by emergency public health measures taken to contain disease outbreaks. Within France, in particular, COVID-19 and the resulting social and economic conditions had a significant impact on the mental health of the population. According to the Weekly Epidemiological Bulletin published by the French public health agency Santé Publique France (SPF)[5], 13.3% of people aged 18-75 experienced a depressive episode during 2021, up 36% from 2017. The increase is mainly observed among young adults (18-24 years), with 20.8% of this age group affected in 2021, compared with 11.7% four years earlier, an increase of nearly 80%. Young women, in particular, are more impacted (26.5%) than young men (15.2%)[6].

Despite their prevalence and impact on society, global provisions and services for identifying, supporting, and treating mental disorders of this nature and at this magnitude have been considered insufficient [22, 24]. These mental disorders form 14% of the global burden of disease, and yet most of the affected people - up to 75% in low-income countries - lack access to necessary treatments and services[7]. Although most governments in the world (almost 87%) offer some form of primary healthcare services within the mental health domain, significant portion either do not have specific programs or lack the budget specifically identified for mental health [24]. The situation is further aggravated due to a lack of a reliable clinical test for the

[4]https://www.who.int/news-room/fact-sheets/detail/depression

[5]https://www.santepubliquefrance.fr/revues/beh/bulletin-epidemiologique-hebdomadaire

[6]https://www.lemonde.fr/en/science/article/2023/02/14/one-in-five-young-french-people-has-a-depressive-disorder_6015640_10.html#.

[7]https://www.who.int/teams/mental-health-and-substance-use/treatment-care/mental-health-gap-action-programme

diagnosis of most forms of mental illness; with the diagnosis typically based on the patient's self-reported behavior, inputs from friends and family, or a mental status examination by medical professionals which is subjective in nature.

The surge in the number of people seeking professional help for mental disorders has led to a significant burden on the current healthcare system, with the demand reaching beyond the capacity of the available resources in some cases. In the Ain region of France for example, the supply of psychiatric care is half the national average with only 9 psychiatrists for 100,000 inhabitants[8]. The deteriorating global situation has prompted both governmental and non-governmental organizations to take action, aiming to reduce the social and financial impacts of these mental disorders. World Health Organization (WHO), for instance, proposed the Mental Health Gap Action Programme (mhGAP) that aims to scale up services for mental, neurological, and substance use disorders, especially for low- and middle-income countries. The program asserts that with proper care, psychosocial assistance, and medication, even in regions with scarce resources tens of millions could be treated for depression, schizophrenia, and epilepsy, helping them lead normal lives.

1.1 Artificial Intelligence and Mental Health

Advancements in the field of computer science, powered by progress in the domains of Machine Learning (ML) and Artificial Intelligence (AI), have allowed researchers to develop digital tools that can support healthcare systems in managing this growing demand while also extending their reach. Digital technologies have a far greater reach than any healthcare system in the world, consequently, the integration of medical services and digital tools can provide ample benefits, especially in the context of reach and accessibility of these services. In recent years, steps have been taken by various governments to digitize their healthcare systems. The travel restrictions and lockdowns implemented during the COVID-19 pandemic fast-tracked this integration process in many countries and regions in order to address not just the growing demand but also the difficulties faced by patients in accessing these services. These initiatives include basic measures like digitization of health records for easier access, and also AI-based solutions for both personal and clinical use cases. One of the major advancements in this field has been the advent of smart wearable devices like watches and bands. These devices are not only capable of tracking an individual's vitals, but the data from their sensors can also be used to predict conditions like stress, anxiety,

[8] https://www.leprogres.fr/ain-01/2019/03/06/le-departement-en-penurie-de-psychiatres

high/low heart rate, etc., in real-time. Progress in the field of Natural Language Processing (NLP), in particular, has allowed for the development of systems capable of understanding human behavior to a great extent through language. In addition to various other applications, these tools have also been applied to the mental health analysis field, aiming to not only reduce stress on the healthcare systems but also exploit digital tools and technologies to extend the reach of medical services.

Automated mental health assessment has been a major research focus in recent years, enabling real-time analysis of a much larger population, carried out according to patient's convenience. This includes a wide range of services ranging from anxiety and stress detection using devices like smartwatches and bands, to the possibility of automated depression estimation based on linguistic features learned from patients' interactions with psychiatrists or virtual agents. Such virtual agents, accessed through mobile phones or computers, open up the possibility of catering to the needs of a larger population as compared to in-person interviews which are the common practice within the healthcare systems.

1.2 AI and Clinical Depression Estimation

Although the research in the field of Automated Depression Detection (ADD) relies heavily on social media based datasets, this thesis is focused only on depression estimation based on clinical interviews. Within the healthcare systems, patient-therapist interviews are the common practice for assessing a patient's mental health, during which, medical professionals actively try to uncover verbal and non-verbal indicators of the patient's health. Therefore, in contrast to social media posts, these interviews are more detailed and abundant in pertinent information for assessing depression. Clinical mental health assessment depends significantly on the patient's personal life, consequently, such patient-therapist interviews contain highly sensitive information, making it extremely difficult to collect and distribute such datasets. To the best of our knowledge there are only two such datasets, the General Psychotherapy Corpus (GPC)[9] and the Distress Analysis Interview Corpus - Wizard of Oz [38] (DAIC-WOZ) dataset, with DAIC-WOZ being the only publicly available clinical dataset[10], making it the gold standard for automated depression estimation based on clinical interviews. The DAIC-WOZ dataset comprises interviews collected in the Southern California region, focusing mostly on army veterans. It contains data for visual,

[9] http://alexanderstreet.com

[10] Despite multiple attempts and emails, we were not able to gain access to the General Psychotherapy Corpus (GPC) used by Xezonaki et al. [92].

speech, and text modalities, thus encoding both linguistic and non-linguistic features. Each participant within the study was also asked to fill out a self-assessment form, the results of which act as ground truth within our research. Unlike other medical conditions, Major Depressive Disorder (MDD) lack clinical tests for their evaluation and quantification. For clinical assessment, medical professionals rely on their knowledge, training, and experience to assess a patient's mental health. Although grounded in medical knowledge and expertise, these evaluations are subjective, with different clinicians providing varying assessments of a patient's mental health. In addition to these clinical evaluations, various self-assessment tools have also been defined for a more standardized evaluation of an individual's mental health. These tools represent a patient's self-evaluation in the context of pre-defined markers, with most implementations scoring individual symptoms on a Likert scale ranging from 0 to 3. Over the years, different self-assessment tools have been proposed with the Beck Depression Inventory [9] (BDI) containing 21 items, the Center for Epidemiologic Studies Depression Scale [70] (CES-D) containing 20 questions, and the Patient Health Questionnaire [49] (PHQ-9) with 9 symptoms being the most widely used questionnaires. The DAIC-WOZ dataset, in particular, uses the Patient Health Questionnaire-8 [50] (PHQ-8) self-assessment tool for generating the ground-truth assessments of individual patients. PHQ-8 is a variant of the original PHQ-9 questionnaire formed by simply removing the question related to suicide and self-harm. The PHQ-8 questionnaire scores the individuals based on eight markers, with their sum acting as the final evaluation of the patient. The symptoms considered within this questionnaire are: loss of interest, sudden change in appetite, sleeping habits, lack of concentration, feeling of depression, feelings of failure and low self-worth, lack of movement or hyperactivity, and lack of energy.

Recently, there has been a surge in research focused on automating depression detection based on clinical interviews. Researchers have been leveraging inputs from various modalities to explore how different aspects of depression manifest in patients. DAIC-WOZ dataset provides data from all three modalities, visual, speech, and text, allowing researchers to study changes in body language, speech patterns, and language use of the participants. Among other effects, depression is known to impact the use of language within the patients, with observed differences in language use between depressed and non-depressed individuals reported in various psychological studies [76, 11, 71]. Seeking to exploit this, researchers have been using language as a differentiating factor between depressed and non-depressed individuals. Within the context of DAIC-WOZ dataset, different strategies have been proposed for depression

estimation based on language, which consists of inferring the screening tool score (PHQ-8 score) based on transcribed clinical interviews. Multi-modal architectures combine inputs from different modalities combining linguistic features with those from speech and visual modalities [73, 69]. Multi-task architectures simultaneously learn related tasks, relying on their similarities for improved model performance and robustness [69, 68]. Gender-aware models explore the impact of gender on depression estimation [7, 61]. Models based on hierarchical text encoding process input text at different granularity levels [55, 92, 56], while attention models integrate external knowledge from mental health lexicons [92], and feature-based solutions compute multiple multi-modal characteristics [20]. Graph neural networks are used to not only highlight the non-linear structures within the individual transcripts at both word and sentence levels but also study the interactions between the important words within the corpus and the interview transcripts [43, 59, 17]. Symptom-based models treat depression estimation as an extension of the symptom prediction problem [56] and train models to predict individual symptoms rather than the final PHQ-8 scores. Domain-specific language models have been built [45] and large language models have been prefix-tuned to automate depression estimation [51].

1.3 Research Hypothesis

Building upon these recent initiatives, this work also focuses on automated depression estimation based on linguistic information extracted from transcribed clinical interviews. In particular, we work on the binary classification problem within the ADD domain. Most recent research initiatives within this context focus on defining complex neural architectures for processing the input transcripts, while failing to account for other key aspects of the data. This thesis raises the following research questions that are missing from the literature:

Question 1: *What is the relevance of discourse structure in understanding the interview transcripts?*

Given the aim of assessing the mental health of patients, clinicians usually tend to base their evaluations only on inputs from the patient. This belief has also percolated into computational research with researchers discarding therapist input from the transcript and only using patient utterances to train their models [55]. Although the relevance of therapist utterances for mental health assessment has been verified by Xezonaki et al. [92] (for the GPC dataset), they still consider the transcript as an unstructured sequence of sentences. The dyadic nature of the conversation implies

the existence of an underlying discourse structure that is not accounted for in recent initiatives. Within this work, it is argued that incorporating the said structure into the training process can improve the model's performance by removing irrelevant interactions from the input stream and training the model on more refined data.

Question 2: *What role do input representations play in the ADD domain?*

Conversations are inherently non-linear structures made up of complex interactions including frequent use of past utterances as context. This is especially true for people suffering from mental disorders who often have trouble forming coherent sentences. Despite this, most research initiatives in the ADD domain use a sequential encoding of the input transcript, forcing the model to learn the underlying linguistic complexities from an unstructured sequence of sentences. This dissertation emphasizes the sub-optimal nature of sequential text representations and argues in favor of graph-based encoding of the input transcripts allowing better representation of the non-linear interactions within the discourse for improved predictive performance of neural network models. This hypothesis is based on the success of graphical representations in various text classification tasks as well as depression estimation tasks [59, 43, 17]. Furthermore, it is shown that depending on the definition of the graph structures, different graphical representations can provide different understandings of the same input, which is not always possible with sequential models. Finally, this work also explores the possibility of using these graphical representations as a means of understanding behavioral changes from both patient and therapist input, generating insights that can be useful for medical professionals, and using their visualizations as a quick visual synopsis for clinicians. We claim that, compared to sequential encoding, graphs are better suited for representing patient-therapist interviews not just for improving the predictive performance of the models but also as a possible source of insights and indicators of behavioral changes.

Question 3: *How reliable are neural network predictions?*

One major hindrance in the use of neural networks as predictive models in the medical domain is their "black box" nature. Since their predictive process cannot be explained reliably, their integration into the healthcare system has been difficult. This problem is further aggravated by the lack of medical professionals in the learning process of these models, further raising questions regarding the trustworthiness of their predictions. The integration of domain expertise into the learning process of neural network models can not only improve their predictive performance by allowing them to focus on relevant information, but also strengthen their reliability and

trustworthiness as predictive tools within the medical domain. This thesis examines the behavior of neural network models in the context of clinical annotations and analogizes their psychological tendencies.

1.4 Thesis Structure and Contributions

The remainder of the thesis is divided into four main chapters (chapter 2, chapter 3, chapter 4, and chapter 5). Chapter 2 details the literature review and presents the recent research initiatives in the ADD domain based on both clinical and social media based datasets. It also provides a detailed description of the clinical interview based Distress Analysis Interview Corpus - Wizard of Oz [38] (DAIC-WOZ) dataset used in the experiments discussed within this dissertation. This is followed by a discussion on recent research initiatives in the context of the DAIC-WOZ dataset. Finally, this chapter details some of the more relevant initiatives that have influenced the research defined in this thesis.

Chapter 3 presents research conducted in the context of the first question. To this end, multi-view architectures have been proposed that account for the dyadic nature of the discourse and divide the input transcript into two views, patient view and therapist view, restricting the number of noisy interactions encountered by the model and controling the discourse symmetry. The two views are processed independently and co-dependently, focusing on the relevant sentence-level interactions within the discourse. These include interactions within the set of questions/answers (Intra-view interactions) and those between the corresponding questions and answers (Inter-view interactions). The validity of multi-view architecture is verified using two different text encoding methods, hierarchical models and Sentence Transformer based pre-trained models. Publications based on this research direction are as follows:

- *Agarwal, N., Dias, G. & Dollfus, S. Agent-based Splitting of Patient-Therapist Interviews for Depression Estimation. Workshop on Participatory Approach to AI for Mental Health (PAI4MH) associated to 36th Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA.*

- *Agarwal, N., Dias, G, & Dollfus, S. Analysing Relevance of Discourse Structure for Improved Mental Health Estimation. 9th Workshop on Computational Linguistics and Clinical Psychology (CLPSYCH) associated to 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL). St Julians, Malta.*

Chapter 4 presents research conducted in the context of the second question. It details the different graph-based representations explored during the course of this thesis, highlighting different aspects of the data at both transcript and corpus levels. It further compares the predictive performance of graph-based models against sequential text encoding based models. The multi-view concept is also extended to graphical representations, further highlighting the difference in perspective between the patient and the therapist, further reinforcing the model-agnostic nature of the multi-view concept. Finally, this chapter also exemplifies insight generation in the context of the two graph structures considered in this study. Publication based on this research direction is as follows:

- *Agarwal, N., Dias, G. & Dollfus, S. Multi-view Graph-based Interview Representation to Improve Depression Level Estimation. Brain Informatics, 2024.*

Chapter 5 presents research conducted in the context of the third and final question. To this end, firstly, clinical annotation of the DAIC-WOZ dataset is carried out and the model's predictive tendencies are compared with the psychiatrist evaluations. This chapter provides the details of the annotation process and a basic analysis of the annotations received from the clinicians. The annotations are then incorporated into the learning process as markings for fine-tuning the neural network models. This is followed by comparing said annotations with both, a baseline model trained without clinical input and a model fine-tuned using the clinical annotations. Finally, the model's behavior and predictive tendencies are analogized with those of medical professionals, strengthening the reliability of neural network models. Publication based on this research is as follows:

- *Agarwal, N., Milintsevich, K., Métivier, L., Rotharmel, M., Dias, G., & Dollfus, S. Analyzing Symptom-based Depression Level Estimation through the Prism of Psychiatric Expertise. Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING). Torino, Italia.*

Chapter 2

Related work

In recent years there has been considerable research in the field of automated mental health assessment. Advances in AI and machine learning techniques have allowed the development of automated models for evaluating the mental state of people based on inputs from different sources, thus aiming to alleviate pressure on the healthcare systems. The severe impact depression has on a person's mental and physical health, combined with its widespread impact on the world population has driven significant research into the field of automated depression estimation. Advances in the field of deep learning and Natural Language Processing (NLP) have further provided researchers with the tools and technologies to train more complicated neural network models capable of learning the input much better.

Within the Automated Depression Detection (ADD) field, researchers have studied various aspects of the problem statement including data sources, data modalities, neural network definitions, and training strategies. In terms of source of data, research initiatives are mainly categorized into two groups: (1) research based on social media posts [15, 37, 75, 80, 91, 100], and (2) research based on clinical interviews [20, 55, 56, 61, 92]. Social media based datasets typically comprise posts scraped from online portals like X (formally Twitter) [80, 15] and Reddit [100, 66], and generally include individual posts from depressed (self-diagnosed) and non-depressed individuals. On the other hand, clinical datasets contain actual patient-therapist interviews aimed at the mental health assessment of an individual and are therefore more descriptive and informative. This highlights a fundamental difference between the two categories of datasets and defines their use cases. Although clinical interview based datasets provide a more accurate and realistic representation of the clinical mental health assessment process, social media datasets are better suited for monitoring and

real-time interventions. Moreover, these datasets, clinical interview based datasets in particular, can contain information encoded in multiple modalities (audio, video, and text) allowing models to not only understand the language used by the patient, but also learn other relevant traits like the patient's facial expressions, and tone. Researchers have explored ADD tasks based on individual data modalities [56, 92] and also combinations of them [59, 20, 69]. Despite the availability of multiple modalities, this research focuses only on textual data. Incorporating other modalities within the proposed architectures is left as a future endeavor. Another widely exploited aspect of the dataset is the possibility to define different tasks within the ADD umbrella. Depression estimation can be studied as a regression problem predicting the final PHQ-8 scores [56, 59, 68], a multi-class classification problem to understand depression severity [68], or as a binary classification task to differentiate between depressed and non-depressed patients [92, 55]. Multi-task learning frameworks have also been defined for learning more robust models by combining tasks within the ADD domain with other tasks within the domain [59, 69], as well as other behavioral assessment tasks [68]. As is the trend with most applications of NLP, different neural network definitions have also been explored in the context of ADD including models based on Recurrent Neural Network (RNN) [92, 55, 78], attention mechanisms [92], and transformer-based language models [56]. This chapter not only describes the Distress Analysis Interview Corpus - Wizard of Oz [38] (DAIC-WOZ) dataset used in our experiments, but also discusses some of the research initiatives in the context of both social media based and clinical interview based datasets.

2.1 Social Media and Digital Mental Health Evaluation

Among the two different sources of data available, social media posts are possibly the most popular one. Social media has provided a platform for people to connect with individuals worldwide, express their emotions and feelings, and communicate with others having similar issues, more often than not sharing experiences, and coping mechanisms, which somewhat alleviates feelings of isolation and the stigma surrounding mental health issues. Within the healthcare setting, medical professionals rely on Electronic Health Records (EHR) and clinical interviews for a patient's mental health assessment. Despite their clinical validity and usefulness, these input sources suffer from two major constraints. The presence of sensitive personal information of the patient within these interviews raises serious concerns over the privacy and confidentiality of the data, making access to such datasets extremely difficult. Secondly,

these records only hold information based on patient's occasional meetings with the healthcare providers, consequently, changes in their health and well-being may not be recorded immediately, thus preventing possible real-time interventions. The growing number of users on social media platforms, combined with the long hours people spend on them means that social media based data is abundant compared to clinical interviews. Although social media posts are seldom as informative as clinical interviews, they promise great benefits in identifying risky behavior, assessing developing conditions, providing timely interventions to at-risk people, and reaching populations not within the reach of current clinical setups. In fact, such approaches have already been applied within platforms like Facebook for suicide prevention efforts[1][2]

Owing to the relative abundance of social media based data and their ability to support real-time tracking of a person's mental well-being, various studies have been conducted on Automated Depression Detection based on social media posts. The initiatives in this research direction mainly focus on social media portals and apps like X (formerly Twitter) [80, 15], Reddit [37, 100], Instagram [75], and Facebook [91, 29]. Gkotsis et al. [37] focus on a more fundamental problem within this research direction and define neural network and deep learning based approaches for identifying posts related to mental health on the Reddit portal. Given the huge quantities of data available on social media platforms, it becomes imperative to recognize and filter mental-health related posts to have more refined training data. They propose two sub-tasks within their research: firstly, a binary classification of posts as being mental-health related or not, and secondly, the classification of each post into one of eleven mental-health themes considered in their study. Rather than providing final predictions, their research is suited for use within the data processing step to filter out noisy posts from the input before training dedicated ADD models. While a majority of the research focuses on user-generated posts for predicting Major Depressive Disorder (MDD), Ricard et al. [75] propose an exciting approach and incorporate community-generated information within the model input. Specifically, they supplement user-generated content, i.e. content created by the user, like their posts and pictures, with community-generated content, i.e. content generated by a community of friends and followers like a post's "likes"/comments, friends' "wall" posts, and followers. This community information, although not generated by the individual users themselves, contains information about them, and friend pair's bi-

[1] https://engineering.fb.com/2018/02/21/ml-applications/under-the-hood-suicide-prevention-tools-powered-by-ai/

[2] https://www.theverge.com/2017/11/28/16709224/facebook-suicidal-thoughts-ai-help

directional engagement on the social media platform. They hypothesize and show that word-based user-generated and community-generated content contain complementary information indicative of an individual's MDD status.

Various other research works have also been published for automatic depression estimation based on social media posts. These include studies based on statistical and machine learning based approaches [15, 21, 22] that utilize hand-crafted features for the task. Various deep learning based approaches have also been explored in recent years owing to their success in other NLP tasks. Some researchers apply straightforward deep learning architectures for different mental health assessment tasks [81, 89]. Others have incorporated more advanced techniques including Convolutional Neural Networks (CNNs) [37] and Recurrent Neural Networks (RNNs) [78] based approaches. Advanced transformer-based language models have also been used within this field in order to leverage their excellent language understanding [46], with more recent works also incorporating Large Language Models (LLM) into their research [96]. Chancellor et al. [18] provide a more detailed survey of the research done within the field of social media based mental health assessment.

2.2 Clinical Mental Health Assessment

Contrary to social media datasets that contain short texts, clinical datasets are made up of more descriptive patient-therapist interviews. These interviews are longer, dyadic conversations wherein the medical professionals actively try to uncover verbal and non-verbal clues about a patient's mental health. As such, compared to social media posts, these interviews provide a more detailed, informative and structured input for training automated depression estimation models that can be deployed within the healthcare system. Unfortunately, the nature of the task guarantees the presence of sensitive information within the interview transcripts, resulting in major concerns surrounding the collection and distribution of such datasets. To the best of our knowledge, there are only two clinical datasets available at the moment: the General Psychotherapy Corpus (GPC)[3] and the Distress Analysis Interview Corpus - Wizard of Oz [38] (DAIC-WOZ) dataset, with DAIC-WOZ being the only publicly available dataset in this domain. The research discussed in this dissertation is completely focused on the DAIC-WOZ dataset, the gold standard for automated depression estimation based on clinical interviews. Subsequent parts of this chapter discuss the DAIC-WOZ dataset and present the recent literature on automatic de-

[3] http://alexanderstreet.com

pression estimation based on clinical interviews, with a specific focus on works based on the DAIC-WOZ dataset.

2.2.1 Distress Analysis Interview Corpus - Wizard of Oz

In the context of clinical interview based automated depression estimation, Distress Analysis Interview Corpus - Wizard of Oz [38] is the most widely used public dataset. It is part of a larger corpus, the Distress Analysis Interview Corpus (DAIC) [38] which is a multi-modal collection of semi-structured clinical interviews. The interviews are designed to simulate standard protocols for identifying people at risk for conditions such as Major Depressive Disorder (MDD) and Post-Traumatic Stress Disorder (PTSD). These interviews were collected at the University of Southern California with a larger goal of developing a computer agent that interviews participants to identify verbal and non-verbal signs of mental illness [25]. Participants were drawn from two distinct populations living in the Greater Los Angeles metropolitan area - veterans of the U.S. armed forces and the general public - and are coded for depression, PTSD, and anxiety based on accepted psychiatric questionnaires. The DAIC corpus contains four types of interviews:

Face-to-face interviews were conducted by human interviewer.

Teleconference interviews, conducted by human interviewers over a teleconferencing system.

Wizard-of-Oz interviews were conducted by a virtual agent named Ellie, controlled by a human interviewer from another room.

Automated interviews, conducted by the virtual agent operating in a fully automated mode.

Research presented in this thesis only uses the Wizard-of-Oz part of the corpus which is publicly available. Ellie's behavior in the Wizard-of-Oz collection was controlled by two wizards, responsible for the non-verbal behavior (e.g. nods and facial expressions) and verbal utterances (the wizards were the interviewers from the face-to-face and teleconference interviews). Within this setting, Ellie had a fixed set of utterances containing pre-recorded audio of the wizard that controlled Ellie's verbal behavior and pre-animated gestures and facial expressions based on those typically employed during the face-to-face interviews.

These interviews have been transcribed and annotated for variety of verbal and non-verbal features. In addition to the transcripts, the dataset also includes the

Depression severity	Data split		
	Train	Val.	Test
No symptoms [0..4]	47	17	22
Mild [5..9]	29	6	11
Non-depressed Total	76	23	33
Moderate [10..14]	20	5	5
Moderately severe [15..19]	7	6	7
Severe [20..24]	4	1	2
Depressed Total	31	12	14
Total	107	35	47

Table 2.1: Number of interviews for each depressive class severity in the DAIC-WOZ dataset, distributed by train, validation, and test sets.

corresponding visual and audio features extracted from the interview recordings, although this research solely utilizes the textual features. In addition to the interviews, the participants were also asked to fill out self-assessment Patient Health Questionnaire-8 (PHQ-8) forms. The results from these questionnaires are used as ground truth within this research initiative, which is in line with the standard practice in the field. Depression severity was assessed based on the PHQ-8 depression scale, with a score of 10 acting as the threshold to differentiate between depressed and non-depressed classes within the binary classification task. The dataset is divided into training, development, and test sets containing 107, 35, and 47 interviews respectively. Furthermore, the data shows a bias towards lower PHQ-8 scores with almost 70% data points belonging to the negative class in case of binary classification and only 6 instances with severe depression (PHQ-8 score > 17). Table 2.1 gives a detailed class distribution within the DAIC-WOZ dataset, while figure 2.1 provides a sample excerpt from the dataset.

2.2.2 AI and Clinical Depression Estimation

Within the context of the DAIC-WOZ dataset, various architectures and strategies have been proposed throughout the literature exploiting different aspects of the data for patient's mental health assessment.

A promising research area is to leverage inputs from different modalities into a single learning model [73, 69, 59, 20]. This stems from the fact that clinicians also rely on multi-modal features like facial expressions, posture, and speech characteristics, for making their final assessment. The availability of data from multiple modalities within the DAIC-WOZ dataset has allowed researchers to train more robust models by combining information from multiple modalities. Qureshi et al. [69] explore the

ELLIE: Who's someone that's been a positive influence in your life?
PARTICIPANT: Uh my father.
ELLIE: Can you tell me about that?
PARTICIPANT: Yeah, he is a uh
PARTICIPANT: He's a very he's a man of few words
PARTICIPANT: And uh he's very calm
PARTICIPANT: Slow to anger
PARTICIPANT: And um very warm very loving man
PARTICIPANT: Responsible
PARTICIPANT: And uh he's a gentleman has a great sense of style and he's a great cook
ELLIE: Uh huh
ELLIE: What are you most proud of in your life?

Figure 2.1: Sample excerpt from the wizard-of-oz interviews.

possibility of combining audio, visual, and textual input features into a single architecture using attention fusion networks. Ray et al. [73] present a similar framework that invokes attention mechanisms at several layers to identify and extract important features from different modalities. The network uses several low-level and mid-level features from audio, visual, and textual modalities of the participants' inputs. Niu et al. [59] also incorporate patient audio into the learning process, and use an early fusion strategy to combine audio and textual features within the input. An interesting trend seen within these research initiatives is the constant presence of text modality within the input configurations, showcasing how important a role language plays in understanding a patient's mental health.

Another interesting approach aims to combine different tasks that share some common traits, thus following the multi-task paradigm. Qureshi et al. [69] and Niu et al. [59], for example, train their models on DAIC-WOZ dataset for both regression and classification tasks. Moving further in this multi-task learning paradigm, Qureshi et al. [68] propose to simultaneously learn both depression level estimation and emotion recognition on the basis that depression is a disorder of impaired emotion regulation. They show that this combination provides improvements in performance for the multi-class emotion classification problem as well as the regression of the PHQ-8 score. Exploring a different research direction, Qureshi et al. [61] study the impact of gender on depression estimation and build four different gender-aware models that show steady improvements over gender-agnostic models. In particular, an adversarial multi-task architecture provides the best results overall. Along the same line, Bailey et al. [7] study gender bias from audio features as compared to

Qureshi et al. [61], who targeted textual information. Their research findings show that deep learning models based on raw audio are more robust to gender bias than the ones based on other common hand-crafted features, such as mel-spectrogram.

Building on the success of hierarchical models for document classification, different studies [55, 92] propose to encode patient-therapist interviews using hierarchical structures as text encoding frameworks, showing boosts in performance. They propose a two-stage hierarchy that allows the model to encode both word-level and sentence-level information. Their hierarchical models are further augmented with attention mechanisms [6] to identify salient words and sentences within the input transcripts. Xezonaki et al. [92] further extend their proposal and integrate affective information (emotion, sentiment, valence, and psycho-linguistic annotations) from existing lexicons in the form of specific embeddings. Their models aim to leverage the affective context of depression language by fusing these specific embeddings with word-level features. Milintsevich et al. [56] also define a hierarchical architecture, although, they employ more advanced models from SentenceTransformer (S-RoBERTa)[4] to learn word-level features for a more contextualized understanding of the transcript. They treat the binary classification task as an extension of symptom profile prediction problem and train a multi-target hierarchical regression model to predict individual depression symptoms from clinical interview transcripts.

Although most strategies rely on deep learning architectures, a different research direction is proposed by Dai et al. [20], who build a topic-wise feature vector based on a context-aware analysis over different modalities (audio, video, and text). They show the effectiveness of these hand-crafted features by using them as inputs for training Support Vector Machine (SVM). The success of graph-based approaches in different linguistic tasks has also prompted their use within the ADD domain [59, 43, 17]. Niu et al. [59] use graph structures within their architecture to grasp relational contextual information from audio and text modality using their proposed hierarchical context-aware model (HCAG) that captures and integrates contextual information among relational interview questions at word and question-answer pair levels. Hong et al. [43] and Burdisso et al. [17] explore word-word interactions in the context of both transcript-level graphs [43] and a global corpus-level graph encoding word-word and word-transcript interactions [17]. The remainder of this chapter provides a detailed discussion of some of the recent initiatives in the field that have influenced the research discussed in the subsequent chapters.

[4] http://huggingface.co/sentence-transformers/all-distilroberta-v1.

2.3 Hierarchical text encoding

Hierarchical models have been proposed for document classification tasks, in order to leverage the hierarchies existing in the document structure and construct a document-level representation based on intermediate word-level and turn/sentence-level representations [85]. These models have further been augmented with attention mechanism [6, 86] in order to extract salient words and sentences in the document for a more refined understanding [97]. Studies have shown that depression has an impact on language use with observed differences between depressed and non-depressed individuals [72, 11, 76]. Seeking to exploit this fact, various studies have focused on mental health estimation through the analysis of linguistic information from transcribed clinical interviews, using a hierarchical model of text encoding [55, 92, 56].

Within the context of clinical depression estimation, psychiatrists assess a patient's mental health based on a sequence of sentences, which are themselves sequences of words. This hierarchical structure of depression assessment interviews has led to the success of hierarchical text encoding based research within the ADD field. Within such studies, most models encode hierarchies within the documents in a bottom-up manner defining a two-stage hierarchical network. The first stage of the hierarchy encodes word-level features and generates turn/sentence-level representations of text. The second stage of the network focuses on these learned turn/sentence level encodings to generate a document-level representation of the input transcript. Within this basic definition, different configurations of hierarchical models have been defined and applied to different tasks within the ADD umbrella. Ragolta et al. [55] and Xezonaki et al. [92] define RNN based hierarchical architectures and apply them to classification task within the ADD domain. They rely on pre-trained word-embedding models to generate the input word sequences. These pre-trained models are chosen for generating input word embeddings since they are trained on a much larger corpus, thus ensuring the stability of the generated word representations. Both these research initiatives, Ragolta et al. [55] and Xezonaki et al. [92], use pre-trained Global Vectors [64] (GloVe) embeddings for encoding input word representations, in particular the 300D embeddings trained on the Common Crawl corpus are used. These word-level embeddings are combined using RNN based encoders, specifically a bi-directional Gated Recurrent Unit (GRU), to generate learned sentence-level representations that act as inputs for the next stage within the hierarchy. An attention mechanism is also applied on top of the GRU layers to account for the varying importance of words in a sentence. The advantage of attention models in this context

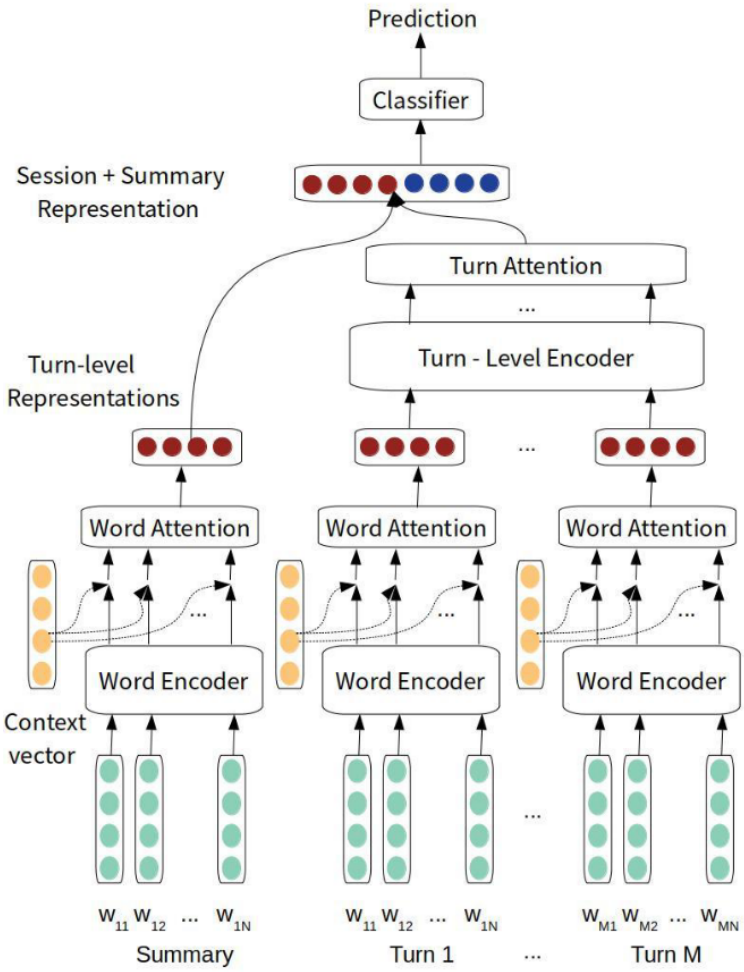


Figure 2.2: Overview of Hierarchical Model with Attentional Conditioning from [92].

is verified by Ragolta et al. [55], who compare models trained with and without attention mechanism, showing clear advantages of using attention mechanism in hierarchical models. Ragolta et al. [55] also conduct an ablation study and define three different attention configurations: a naive approach, local attention-based approach, and contextual attention approach (detailed in [55]), while Xezonaki et al. [92] employ attention mechanism defined by Bahdanau et al. [6] within their neural network definition. The learned sequence of sentence representations is then fed through a similar combination of GRU and attention mechanism to learn the document-level representation of the transcript. These document-level representations are finally passed through a classifier network to generate the final predictions. Figure 2.2 shows an overview of the model defined by Xezonaki et al. [92]. Building upon the

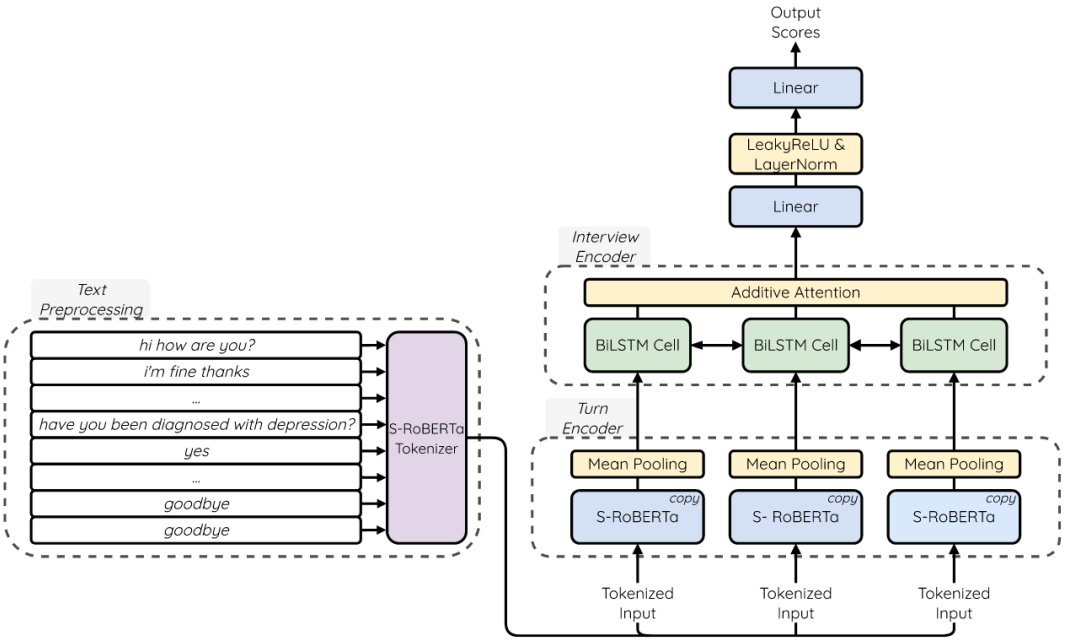


Figure 2.3: Hierarchical architecture used by Milintsevich et al. [56]. On turn level, the same instance of the S-RoBERTa model is used to encode each turn.

work done by Rude et al. [76], who emphasize the role of affective content as a distinguishing factor between depressed and non-depressed language, Xezonaki et al. [92] further incorporate external linguistic knowledge in their model. These features (context vectors) are constructed based on various emotional, sentiment, valence, and psycho-linguistic annotations, and concatenated to the input word representations. In the context of General Psychotherapy Corpus (GPC), they also combine the learned session/document representation with the summary representation vector, which is the learned representation of the transcript summaries available in GPC dataset. Within their experiments with the DAIC-WOZ dataset, this summary representations are not used[5], although the context vectors are still incorporated into the model. An interpretation of this hierarchical model is later used for text encoding within research on Multi-view architectures discussed in chapter 3.

Milintsevich et al. [56] propose a similar hierarchical model for symptom-based prediction of depression. Compared to previous works [55, 92], Milintsevich et al. [56] employ more advanced NLP techniques in their architecture (figure 2.3). The word-level encoder defined in their architecture uses a distilled RoBERTa-based pre-

[5]DAIC-WOZ dataset does not provide transcript summaries as part of the data.

trained model (S-RoBERTa) from the SentenceTransformer module[6]. The sentence-level encoder is defined using a single bi-directional LSTM (BiLSTM) with additive attention, in contrast to bi-directional GRUs used in [92]. They focus on regression task within the ADD domain and define depression estimation as an extension of the symptom estimation problem. As such, the model adopts a prediction head that produces eight symptom-level regression outputs effectively making it a multi-target regression model. An updated version of this architecture is employed for symptom prediction in research discussed in chapter 5.

2.4 Automated depression Estimation and GNN's

Neural network architectures like GRU, and Long Short Term Memory (LSTM) have been the most popular architectural choices within the NLP domain. RNN architectures, although effective on NLP tasks, were originally defined for processing sequential time-series data. Consequently, they rely on a sequential encoding of the input text which in itself provides a limited understanding of the non-linear linguistic information. Although text is typically represented as a sequence of tokens, there is a rich variety of NLP problems that can be best expressed using a graph structure. Recent years have seen a surge of interest in applying and developing different Graph Neural Network (GNN) based approaches to the NLP domain. These models have achieved considerable success in many NLP tasks ranging from classification tasks like sentence classification [42, 12], semantic role labeling [53, 39], and relation extraction [67, 79], to generation tasks like machine translation [8, 10], question generation [63, 77], and summarization [32, 99]. Even within the context of automated mental health assessment, some researchers have explored the possibility of incorporating GNNs within their learning paradigm [2, 93, 59, 43, 17].

2.4.1 Hierarchical Text Structure and GNNs

Keeping in line with the success of hierarchical models in mental health assessment tasks, Niu et al. [59] define the Hierarchical Context-Aware Graph (HCAG) model that not only mirrors the hierarchical structure of depression assessment by encoding text at word and Question-Answer Pair (QA-pair) levels but also leverages Graph Attention Network [87] (GAT) to grasp relational contextual information of text/audio modality. In particular, they define the HCAG which can effectively capture and

[6]https://huggingface.co/sentence-transformers/all-distilroberta-v1.

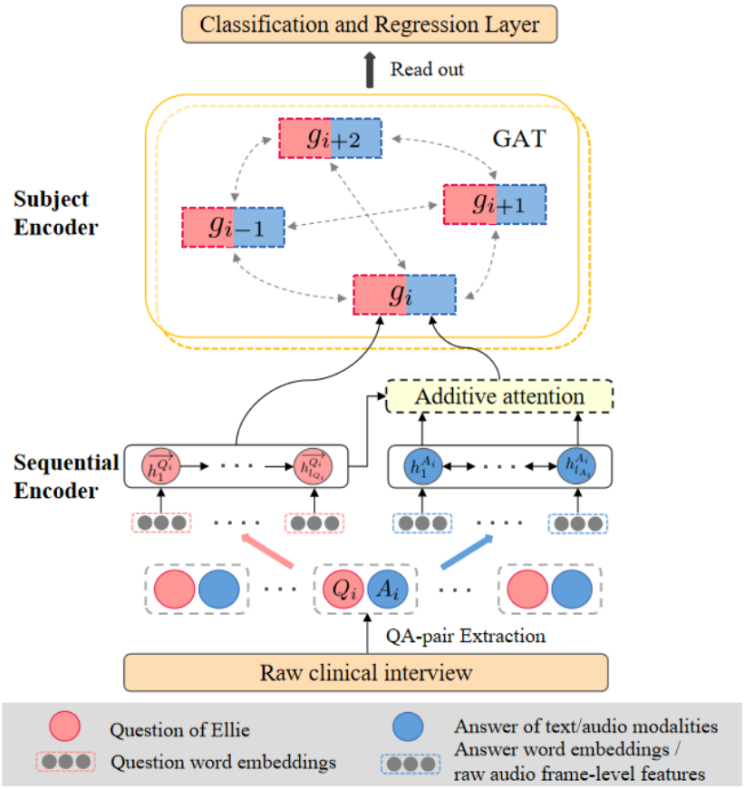


Figure 2.4: An overview of the HCAG architecture taken from [59]. For brevity, self-loop lines are ignored in this figure.

integrate contextual information among relational interview questions by applying GAT networks to text and audio modalities. As the name suggests, HCAG employs a hierarchical representation of the input sequence, encoding it at two levels; word/frame level and question-answer pairs (QA-pairs) level. Medical professionals base their assessment of a patient's mental health on patient-therapist interviews, which are sequences of QA-pairs, which in turn are sequences of words or audio signals. QA-pair encoding layer within HCAG aims to encode the semantic relations of questions and answers. HCAG further uses an attention mechanism to highlight the important behavioral signals that can occur depending on the question types [41]. Finally, a graph neural network based on GAT architecture aggregates the pieces of depressive clues among all QA-pairs. This combined contextual information is then used for a multi-task learning strategy combining classification and regression tasks.

Figure 2.4 illustrates an overview of the HCAG model that uses the early fusion method to combine inputs from both text and audio modalities. Niu et al. [59]

perform manual cleaning of the interview transcripts to remove dialogic feedback from the virtual agent (e.g."that sounds great!") and only the direct queries are preserved (e.g., "how easy is it for you to get a good night's sleep?"). Although this enriches the data used for training the model, such manual pre-processing steps cannot be applied at the inference stage in real-world applications. 300-dimensional GloVe embeddings were used for word-level representations, while for audio modality low-level Mel Frequency Cepstral Coefficients (MFCCs) [58] and extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [30] audio descriptors by openSMILE [31] are concatenated as features. Within their experiments, audio features are only used for the patient responses, while therapist questions are only encoded using pre-trained GloVe embeddings. This is possibly because a virtual therapist was used during dataset recording, which severely limits the expressive power of therapist audio. HCAG model employs different techniques for encoding information at different granularity of the hierarchical structure. The *Sequential Encoder* encodes the questions and answers to generate sentence-level embeddings of the corresponding QA-pairs. Forward Gated Recurrent Unit (GRU) is used to capture the local context information within the questions, while bidirectional GRU layer encodes the answer sequences and concatenates the forward and backward hidden states to get the final representations. During an interview, depending on the question, patients show signs of depression to varying degrees. HCAG model employs an additive attention mechanism to account for this distribution and detect the salient elements in the answer sequence. This mechanism defines the attention scores within the answer word embedding sequence based on both answer word embeddings and the final representation of the corresponding question. Within the second level of the hierarchical structure, subject encoder module captures the contextual information in the interview. For each interview, a graph structure is defined with QA-pairs acting as vertices and edges defined based on a sliding window protocol where each QA-pair is connected to immediate m vertices (QA-pairs)[7]. The model employs GAT networks to process the resulting graph structure and the learned node embeddings are fed into a max aggregator to generate final transcript representation. They further exploit a multi-task learning methodology, combining regression and classification tasks for more robust training of the model. A similar multi-task training is also defined by Qureshi et al. [69].

[7] m is treated as a hyper-parameter and tuned during model training.

2.4.2 Exploring Word-Level Context for Depression Estimation

The ability of graph structures to highlight different features in the input text based on the definition of nodes and edges allows researchers to focus on different aspects of the same input data. Hong et al. [43], for example, model the DAIC-WOZ transcripts as word-word interaction graphs. In contrast to Niu et al. [59] who learn QA-pair level context within transcripts, Hong et al. [43] explore contextual information present in word-level interactions within the transcripts. They emphasize the importance of encoding contextual information about the PHQ-8 topics to determine a patient's mental health and hypothesize that the context of words in a transcript can be used to learn this information. They motivate generating graph-based representations of input transcripts to encode contextual information, assuming that the transcripts contain facts representing depressive symptoms. A novel form of node attributes are also proposed for use within the GNN based model that captures node-specific embeddings for every word in the vocabulary. Word representations are shared globally and are updated according to associations among words in the transcript. The final prediction is made by summarizing representations of all the words in the transcript.

Each transcript is represented using a word-level graph that encodes word-word interactions based on co-occurrence patterns. All unique words appearing in the transcript form nodes of the graph, with edges connecting each node to words appearing within a fixed window on its either side in the transcript. Within NLP domain, learning vector representations that encode the meaning and context of words has been a basic learning task. In their research, Hong et al. [43] defined a novel form of word representations within their graph structure. Each node (word) within the graph is represented using an embedding matrix (rather than a vector), referred to as 'schema' $U_i \in \mathbb{R}^{n \times d}$, which performs the role of recording a global context from interactions between current word and every other word. The j^{th} row of U_i is a vector of length d containing the representation that node v_i has of v_j, with n denoting the total number of unique words (vocabulary size) in the corpus. These schemas allow every word to maintain a dynamic record of the context from the given transcript. Figure 2.5 illustrates an example of the schema update using the Schema-Based Graph Neural Network (SGNN) model defined by Hong et al. [43]. The upper figure shows the initial schema for the word 'hopeless' containing representations for its neighboring words and itself[8]. The figure at the bottom shows the updated schema

[8]For convenience of display window size of 1 is used for displaying associated edges; in actual experiments, the window size is larger.

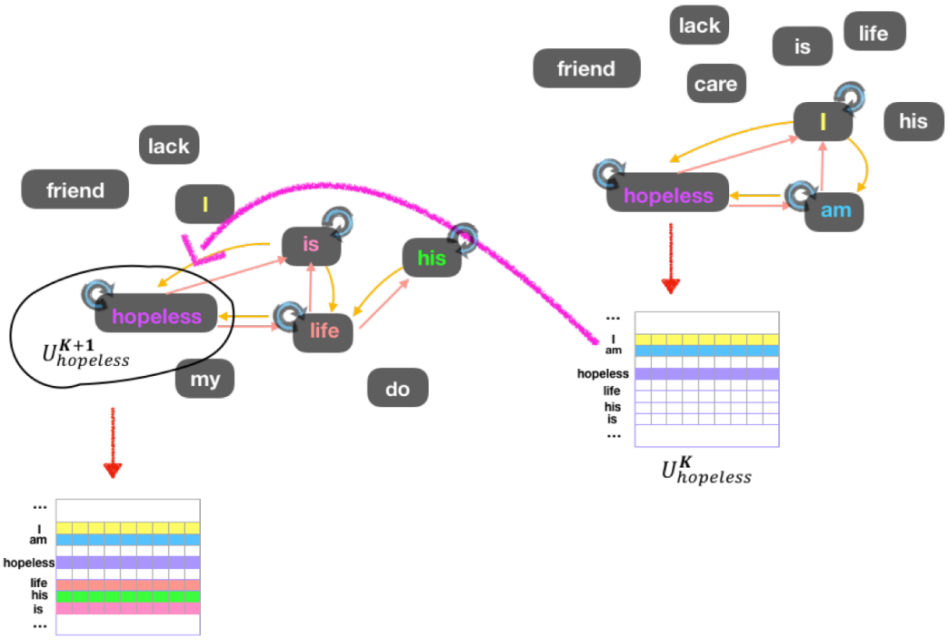


Figure 2.5: An example of schema update taken from Hong et al. [43].

for the same word after learning. The new schema encodes representations for the new neighbors while preserving the information of existing neighbors.

Experiments done by Hong et al. [43] on the DAIC-WOZ dataset show significant performance improvements when using schema-based SGNN models compared to standard GNN architectures. Compared to standard GNNs that use pre-trained embedding models for generating vector representations of nodes, schemas within SGNN employ 2-D node features initialized with random vectors. This increases the expressive power of the message-passing mechanism within SGNN models, thus improving their predictive performance. Furthermore, visualization of transcript-level word clouds (figure 2.6) based on these graph structures can highlight the model's focus, helping us understand the model's learning process to some extent. Compared to the content of transcripts illustrated by word clouds in figure 2.6 (left), word clouds based on content selected by the SGNN model represent a very different view of the data (figure 2.6 (right)).

2.4.3 Node-weighted Graph Convolution Networks

The use of graph-based approaches for document classification has been the focus of various research efforts in the past. Among these studies, Yao et al. [98] proposed a

(a) Transcript A with GT score 16

(b) Transcript A with pred. score 16.46

(c) Transcript B with GT score 19

(d) Transcript A with pred. score 17.30

Figure 2.6: A word cloud depicting words from a transcript on the development set before and after applying the SGNN model. Word clouds on the left depict the most salient words based on the frequency of their occurrences in raw transcripts and those on the right illustrate the most focused content selected by the SGNN model. Graphic taken from [43].

novel neural network method, TextGCN, for text classification that models the whole corpus as a single heterogeneous graph and jointly learns both word and document embeddings using Graph Convolution Network [47] (GCN) based neural network architecture. Figure 2.7 illustrates the schematics of TextGCN taken from [98]. Burdisso et al. [17] further build on this research by adding a simple approach for weighting self-connecting edges and showing its effectiveness on depression detection tasks. Although both, Hong et al. [43] and Burdisso et al. [17], model word-level interactions, the former focus on interactions within individual transcripts and define transcript-level graphs. The latter constructs a corpus-level graph structure encoding both word-word and word-transcript interactions.

In line with the TextGCN definition, Burdisso et al. [17] define a large heterogeneous text graph containing word nodes (V_{words}) and training document nodes (V_{tr_docs}) to capture global word co-occurrence patterns within the entire corpus, as well as the word-transcript interactions. Accordingly, the complete set of nodes is composed as $V = \{V_{tr\_docs}, V_{words}\}$. The corresponding adjacency matrix comprises three major edge categories: (i) word-word edges defined based on Point-wise Mutual Information (PMI) [98, 17], (ii) self-connections for word nodes based on

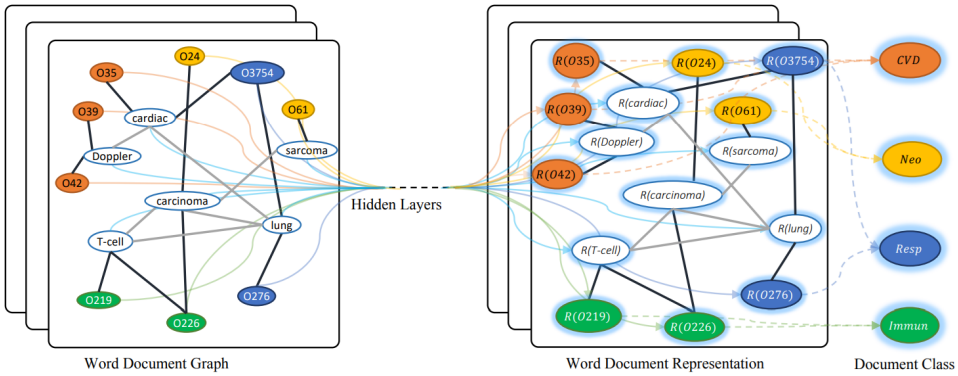
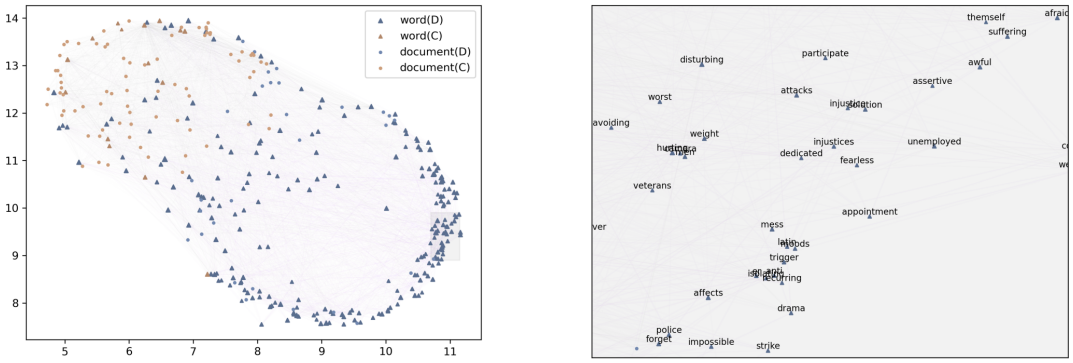


Figure 2.7: Schematic of Text GCN taken from [98]. The example is taken from the Ohsumed corpus. Nodes beginning with 'O' are document nodes, while others are word nodes. Black bold edges are document-word edges and gray thin edges are word-word edges. $R(x)$ means the representation (embedding) of x. Different colors mean different document classes (only four example classes are shown to avoid clutter). CVD: Cardiovascular Diseases, Neo: Neoplasms, Resp: Respiratory Tract Diseases, Immun: Immunologic Diseases.

PageRank (PR) algorithm [16], and (*iii*) word-document edges defined based on Term Frequency - Inverse Document Frequency (TF-IDF) features. Given a graph, PR computes the importance of each node in relation to its role within the overall structure of the graph. Intuitively high PMI values will strongly link word nodes with high semantic correlation, high TF-IDF values will strongly link words to specific documents and high PR values will strongly link a node to itself proportionally to its global structural relevance within the graph. The addition of PageRank-based self-edges in [17] aims to mitigate the limiting assumptions of locality, and the equal importance of self-connections vs. edges to neighboring nodes, in GCNs. This modification to the GCN structure is referred to as w-GCN.

The experiments conducted by Burdisso et al. [17] use the inductive version of GCN as described in [88] instead of the original transductive one [47]. The word node embeddings are defined as one-hot vectors and document node embeddings are defined as TF-IDF representations of the given document with respect to the training set vocabulary. Further implementation details and training process can be referred from [17]. The label information associated with the documents is propagated to the word nodes through the word-document edges, allowing the model to learn relations between the words and output labels (e.g. depressed or control labels), a key aspect favoring the interpretability of the model. Overall results (§4 in [17]) show that the w-GCN approach consistently outperforms the vanilla version,



(a) Overall graph with learned node embeddings

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

Figure 2.8: 2-Dimensional projection of node embeddings learned for DAIC-WOZ taken from [17]. Circles denote documents, triangles words, and colors denote class ([D] - depression, [C] - control). The gray rectangle in (a) indicates the zoomed region shown in (b). Graph edges are also included.

baseline models, and previous approaches. In particular, on the DAIC-WOZ dataset w-GCN evidences the best performance when only the top 250 words are included in the vocabulary. Furthermore, this w-GCN approach can provide interpretability of the learned model. Figure 2.8 shows a 2-dimensional projection of the word and document embeddings learned by the best performing w-GCN models. The figure illustrates how the model can use the graph structure to learn, in the same latent space, word and document embeddings whose distance is influenced by their mutual relations and final predictions. These embeddings allow identifying clusters of strongly related words with high co-occurrence and linked to similar documents in the dataset, i.e., dataset-specific "topics" that could interest the medical experts.

This dissertation also describes research on graph-based representation of the input transcripts, and provides details of experiments and analysis conducted in this context. Chapter 4 presents work in this research direction, showing not only improvements in the predictive abilities of the models but also exemplifying insight generation based on visualization of the graph structures considered in the study. The next chapter in this dissertation presents our research on the relevance of discourse structure within the learning process of neural networks and proposes multi-view architectures within this context.

Chapter 3

Discourse Structure and Text Encoding

In recent years, factors including the global health crisis, increased mental health awareness, and a lack of mental health professionals, have motivated significant research interest in the field of automated mental health assessment. Automated depression estimation, in particular, has been the focus of numerous studies due to its widespread and devastating impact, which can potentially lead to suicide in extreme cases. Throughout the literature, different aspects of the problem have been explored including but not limited to gender bias [7, 61], availability of multi-modal data [73, 69, 59, 20], and integration of lexical knowledge [92]. Various training methodologies like multi-task learning [69, 68], hierarchical text encoding [55, 92, 56], and graphical networks [59, 43, 48] have also been applied to and studied within the context of Automated Depression Detection (ADD) tasks. The research in this field is primarily focused on two major categories of datasets; (1) social media based datasets, and (2) clinical interview based datasets. Datasets based on social media sites comprise of posts by self-diagnosed individuals expressing their emotional and mental state, while clinical datasets include actual patient-therapist interviews aimed at the mental health assessment of an individual. Although social media posts represent the psychological state of an individual, they lack the depth of information found in clinical interviews. This research work focuses on depression estimation based on clinical interviews, with specific interest in the dyadic nature of the conversations. Although, in some cases social media datasets may include conversations between depressed individuals and other participants on the platform (including the general population and medical professionals), these interactions are relatively short

and not aimed towards the person's mental health assessment.

In the clinical setting, patient-therapist interviews are the standard practice for mental health assessment of patients. Within such interviews, the therapist tries to identify verbal and non-verbal signs of mental distress within the patient's behavior. In the past years, some researchers have argued the use of only patient utterances as input within the neural network models [55, 20], centered on the rationale that since we aim to understand the patient's mental health, the assessment should solely be based on their input. This reasoning is further supported by many psychiatrists who focus only on patient utterances during their diagnosis[1]. Patient-therapist interviews are conversations involving two agents, the patient, and the therapist, and this dyadic nature of the discourse inherently places importance on utterances from both agents. This is especially relevant in the case of automated mental health assessment where neural network models require therapist questions in order to contextualize patient responses for a better understanding. One-word responses by the patients perfectly exemplify this co-dependence between the questions and answers within an interview. Let's consider the patient's response "yes", which in the context of the question "Do you want some water?" bears little relevance to the task. However, when the same answer is given in the context of the question "Do you feel depressed?", its relevance is increased significantly. Xezonaki et al. [92] proved the importance of therapist questions in the context of the General Psychotherapy Corpus (GPC) and show improvements in model performance by incorporating both patient and therapist utterances within the input. The validity of retaining therapist questions within the model input is further justified in the context of DAIC-WOZ dataset based on experimental results discussed later in this chapter (§3.3).

Despite the extensive list of research initiatives, ways to express the structure of an input transcript remain a relatively unexplored research direction. In the context of depression estimation based on text modality, most related works either exclusively focus on the patient utterances [55, 20], or treat the overall transcript as a sequence of sentences [92, 56]. Even with the incorporation of therapist information, this latter case disregards the type of individual sentences as questions (therapist utterances) or answers (patient utterances), forcing the model to understand the inter-dependencies within a sequence of unstructured utterances. This chapter underscores not only the significance of including therapist questions in model inputs, but also assert that a sequential combination of patient-therapist utterances is not the most optimal

[1]This fact is also verified and discussed in our work on clinical annotations of the DAIC-WOZ dataset, wherein psychiatrists only annotated patient responses (Chapter 5).

approach for combining the two aspects of input transcripts. It is argued that the dyadic structure of a patient-therapist interview plays an important role in defining its meaning, and needs to be accounted for while processing the interview transcripts. As such, Multi-view architectures are defined that utilize sentence types (questions or answers) in an attempt to encode the said discourse structure into the model's learning process. In particular, the input sequence is divided into two views based on sentence types, allowing us to not only encode the structure but also control discourse symmetry. Experiments show clear advantages of multi-view architectures in the context of both hierarchical text encoding and encodings from transformer based pre-trained models, in particular the sentence-transformer models. Moreover, figures shown in table 3.2 evidence multi-view based architectures out-performing recent research initiatives in the field.

3.1 Multi-view Strategy and its variants

While the objective is to assess a patient's mental health, therapist questions also play an important role in this endeavor and provide relevant knowledge for the final objective. In the context of sequential encoding of input text, the model learns interactions between all possible pairs of sentences. A significant portion of these interactions are not relevant and contribute to noise in the data. A possible example of such noisy interactions would be the interactions between unrelated questions and answers. To put it into perspective, let us assume an interview containing 100 question-answer pairs (200 sentences in total). A sequential encoding of this transcript would force the model to learn from $\binom{200}{2}$=19900 possible interactions, out of which 9900 (approx 50%) interactions are between unrelated questions and answers (interaction between 5^{th} question and 12^{th} answer for instance). Multi-view architectures are defined to not only account for inputs from both patient and therapist but also help maintain discourse structure and symmetry within the learning process. The multi-view architectures are tailored to focus on the 10,000 remaining interactions in the aforementioned example which are more pertinent to the discourse structure. These include interactions within the set of questions/answers, and also the interactions between corresponding questions and answers. This allows for a more efficient training of neural network models based on a refined learning methodology that controls the amount of noise in the data. Multi-view architectures employ sentence types, questions or answers, as a means of encoding the discourse structure for improved predictive performance. The architecture divides the input into two

views based on sentence types, i.e. the therapist view and the patient view, which allows us to control the type of interactions encountered by the model during training and inference. Dedicated sub-networks are used to process the two views both independently and co-dependently, thus learning the different types of interactions within the transcript. This chapter first describes an interpretation of the hierarchical model discussed in chapter 2 (§2.3), which is used as a text encoder in the experiments defined in the following sections. This is followed by a detailed working of the multi-view concept based on the said encoding scheme. Finally, experiments replacing hierarchical text encoding with pre-trained sentence-transformer models are defined to further strengthen the validity of the multi-view concept.

3.1.1 Hierarchical Text Encoding

Hierarchical models treat a patient-therapist interview as a hierarchy of intermediate representations and are widely used within text classification tasks [102, 60, 97]. Inspired by their success in the NLP domain, hierarchical models have also been successfully applied to the ADD field [55, 92]. This thesis also utilizes hierarchical models for text encoding within the research on multi-view architectures. In particular, the hierarchical model defined by Xezonaki et al. [92] is used for text encoding with two main differences: (1) a non-RNN based implementation of hierarchical models is defined by replacing RNN layers with a weighted sum of input-embeddings based on attention mechanism (equ. 3.1). This choice is based on the findings of Mohankumar et al. [57], who show the limitations of attention mechanisms over RNN encodings. These findings are further discussed in the context of DAIC-WOZ dataset in §3.4; (2) context vectors are not included in the attention mechanisms since lexicon-based external knowledge is not used in this work. Figure 3.1 gives an overview of our interpretation of hierarchical text encoding model, where w_{ij} represents the embedding of the j^{th} word of the i^{th} sentence, $W_i = \{w_{i1}, w_{i2}, \cdots, w_{iN}\}$ represents the word encoding sequence for the i^{th} sentence, S_i is the learned representation of the i^{th} sentence, and r is the transcript level representation of the textual input. *Word Attention* and *Sentence Encoder* networks are defined as self-attention networks. Formally, let $[h_1, h_2, ..., h_N]$ be the input of the attention model. The learned representation rep is defined in Equation 3.1, where $g(.)$ is a learnable mapping function, and γ_i is the attention score of the i^{th} input in the sequence. Note that *Word Attention* layer is applied independently to all word sequences W_i, with the i^{th} instance of the *Word Attention* layer using $[w_{ij}, \forall j]$ as the input of the self-attention mechanism giving rise to sentence embedding S_i. The *Sentence Encoder* layer acts on these learned

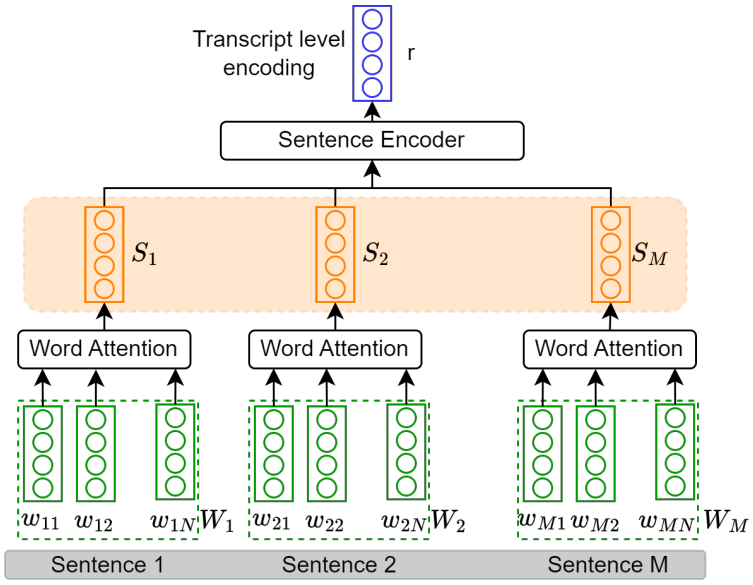


Figure 3.1: Our non-RNN based implementation of the hierarchical model.

sentence embeddings with $[S_i, \forall i]$ acting as its input sequence.

$$\begin{aligned} \alpha_i &= g(h_i) \\ \gamma_i &= \frac{e^{\alpha_i}}{\sum e^{\alpha_i}} \\ rep &= \sum \gamma_i \cdot h_i \end{aligned} \tag{3.1}$$

3.1.2 Hierarchical-Baseline

Based on the above definition of hierarchical architecture, *Hierarchical-Baseline* model is also defined for a fair baseline comparison with the multi-view based definitions. The model utilizes the architecture illustrated in figure 3.1 with the transcript-level representation r used for the final prediction. Specifically, classification layers are added to the architecture, that take transcript-level representation r as input and generate the final model prediction. An ablation study is also conducted to investigate the impact of patient and therapist utterances on the model's learning ability. Consequently, following three input configurations are defined within the context of *Hierarchical-Baseline* configuration:

Patient. This configuration only accounts for the answers given by the patient in the model input while ignoring the questions asked by the therapist (input configuration similarly to the one used by Ragolta et al. [55]).

Therapist. Here we only consider the questions asked by the therapist and remove the patient responses from the model input. This configuration allows us to study the presence of relevant information in therapist utterances.

Patient+Therapist. Both questions and answers are incorporated within the model input. However, they are combined as a sequence of unstructured sentences, thus neglecting their type as question or answer (input configuration similar to the one used by Xezonaki et al. [92]).

3.1.3 Multi-view strategy

Multi-view architectures aim to exploit the discourse structure in order to control the number of noisy interactions encountered by the model during training and inference. As such, the architecture is defined to focus on the relevant interactions within the discourse, i.e. interactions within therapist utterances (questions), interactions within patient utterances (answers), and the interactions between the corresponding questions and answers. It divides the input into two views based on sentence type, patient view and therapist view, thus eliminating all inter-view interactions which are the major source of noise in the data. The remaining relevant interactions are then learned using dedicated sub-networks that allow the model to focus on interaction within the two views. The co-dependency between corresponding questions and answers in the transcript is re-introduced into the model in the form of a cross-attention mechanism between the two views. The transcript-level representations learned by the view-networks are combined and fed through a global network that generates the final predictions.

Figure 3.2 illustrates the proposed multi-view architecture. Within the figure, $\{W_1^Q, W_2^Q, \cdots, W_N^Q\}$ and $\{W_1^A, W_2^A, \cdots, W_N^A\}$ are the corresponding therapist and patient view inputs, with W_i^A representing the word sequence of i^{th} sentence in patient input and W_i^Q the i^{th} word sequence in therapist input. The networks corresponding to the two views, i.e. the *Therapist Network* and the *Patient Network*, are identical instances of the proposed interpretation of the hierarchical model (§3.1.1), and learn transcript level representations of the therapist view (Q) and the patient view (A) from the given word sequences. The *Sentence Encoders* from the hierarchical model are renamed as *View Encoders* in the multi-view architecture, with $\{\alpha_1, \alpha_2, ..., \alpha_N\}$ and $\{\beta_1, \beta_2, ..., \beta_N\}$ representing the attention scores of the respective sentence embeddings $\{Q_1, Q_2, ..., Q_N\}$ and $\{A_1, A_2, ..., A_N\}$. The learned transcript-level representations of the two views (Q and A) are then combined and passed

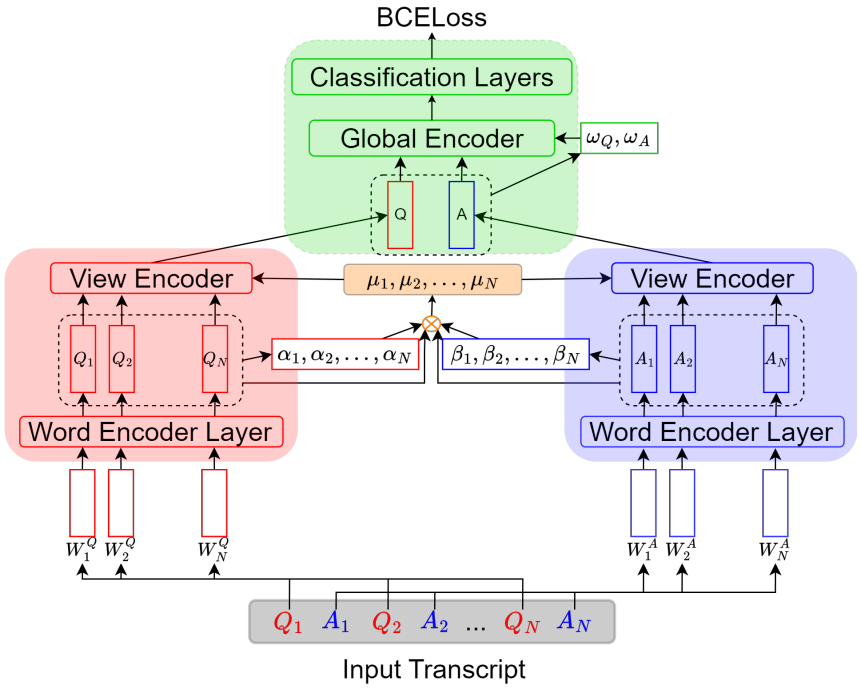


Figure 3.2: Multi-view architecture where the intra-view information is outlined in red and blue, the inter-view linking is painted in orange, and the view fusion network is shown in green.

through a *Global Encoder* layer that combines the two representations before generating the final prediction using classification layers (Multi Layer Perceptron (MLP) layers). Multi-view architectures are further subdivided into two categories based on the interactions learned by the model: (1) intra-view attention based configuration, and (2) inter-view attention based configurations. *Intra-view Attention* strategy only focuses on the interactions existing within the individual views, while the *Inter-view Attention* strategy also models the co-dependency between the corresponding questions and answers.

3.1.4 Multi-view Strategies with Intra-view Attention

This configuration focuses only on the interaction within the two views (intra-view attention). As such, the two views are treated independently of each other using dedicated sub-networks (highlighted using blue and red colors in figure 3.2). The underlying idea is to process the views independently before fusing the learned embeddings at the transcript level to generate the final prediction. This configuration allows us to study the impact of noisy inter-view interactions and the relative con-

tributions of the two views in the decision-making process, while allowing the model to learn the individual perspectives of the two agents involved in the conversation. The two sub-networks are defined as identical instances of the hierarchical model (§3.1.1) and generate transcript-level view representations Q and A. Both view-level attention layers (*View Encoders*) are defined using a self-attention mechanism and combine sentence-level features within the respective views. These embeddings are then fused using the *Global Encoder* layer before feeding it to the classification layers (MLPs) for the final prediction. The global attention layer (*Global Encoder*) is also defined as a self-attention model aimed at fusing transcript-level view representations Q and A. This layer can be seen as an aggregator of all the information contained in a transcript, i.e. questions and answers. Within this context, the following three configurations can be defined for an ablation study, where the self-attention layers are the adjustment variables.

View-Global Attention. Within this configuration both the *View Encoder* and the *Global Encoder* layers are defined using self-attention mechanism (Equation 3.1). The two *View Encoders* pay attention to the corresponding sentence encodings, whereas, the *Global Encoder* learns the relative importance of the two views.

Global Attention. In this configuration, *View Encoders* are replaced by a simple averaging operation instead of a self-attention layer[2] resulting in equal importance to all sentences within the input. The *Global Encoder* remains the same as in the *View-Global Attention* model and employs self-attention mechanism.

View Attention. In this configuration, the *Global Encoder* is replaced by a simple concatenation of the patient representation A and the therapist representation Q[3], while the *View Encoders* remain the same as in the *View-Global Attention* model and employ self-attention mechanism.

3.1.5 Multi-view Strategies with Inter-view Attention

Within the context of intra-view attention models, questions and answers are treated independently, and their co-dependency is not tackled. This is done to avoid noisy interactions between the two views. However, the coherent structure of a dialogue plays an essential role in the global understanding of the message conveyed by the patient. Let's consider a typical patient response "yes", which in itself does not hold much meaning or relevance. However, in context of the corresponding question,

[2]No attention information is acquired at the sentence level.
[3]There is no attention information at transcript level.

the importance of this answer can vary significantly as exemplified earlier[4]. Tackling this co-dependency between questions and answers is of the utmost importance for the learning process. As a consequence, multi-view architectures, with inter-view attention, are proposed that use a shared attention mechanism (highlighted with orange color in figure 3.2) to model the inter-dependencies between the two views. The shared attention mechanism transfers attention scores from one view to another, following the cross-attention paradigm [84]. Formally, attention scores $\{\mu_1, \mu_2, ..., \mu_N\}$ are shared between the two *View Encoders*, and are the result of function $\mu_i = f(\alpha_i, \beta_i)$ that combines the individual view attention scores α and β. Experiments are conducted with five different instantiations of the function f including both unbalanced definition (functions that define final score based on only one of the input scores) and balanced definitions (function definitions that account for both input scores at each step).

Patient. Unbalanced definition focusing only on the patient's attention score.
$f(\alpha_i, \beta_i) = \alpha_i, 1 \leq i \leq N$.

Therapist. Unbalanced definition focusing only on the therapist's attention score.
$f(\alpha_i, \beta_i) = \beta_i, 1 \leq i \leq N$.

Max. Unbalanced definition that favours the view with higher attention score.
$f(\alpha_i, \beta_i) = max(\alpha_i, \beta_i), 1 \leq i \leq N$.

Mean. Balanced definition of f that pays equal attention to both views.
$f(\alpha_i, \beta_i) = (\alpha_i + \beta_i)/2, 1 \leq i \leq N$.

Learnable. A balanced definition where $f(.,.)$ is defined as self-attention (equ. 3.1) acting on combined patient and therapist inputs $h_i = (A_i \oplus Q_i), 1 \leq i \leq N$. Within this definition, f acts on the corresponding sentences embeddings rather than the attention scores.

3.1.6 BERT-based Concept

The advent of transformer architecture has resulted in many NLP models capable of generating rich embeddings at both word and sentence levels. These pre-trained models provide encodings that are highly contextualized, capturing the true meaning of the textual input. While the initial experiments with multi-view architectures

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

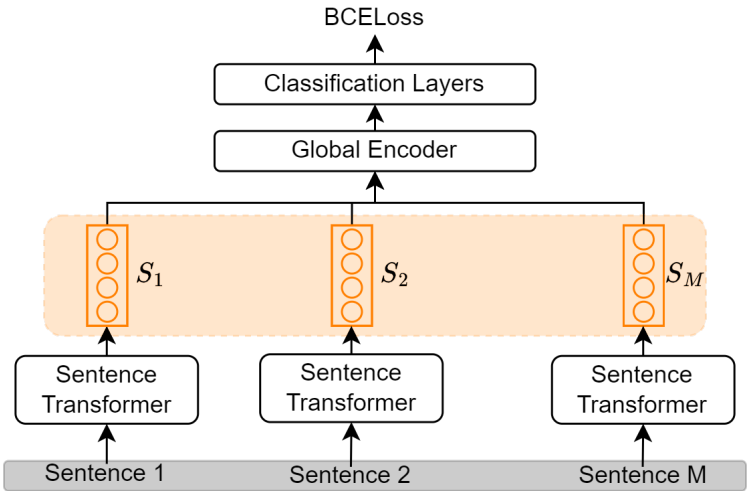


Figure 3.3: Transformer based baseline architecture employing sentence-transformers for text encoding.

are rooted in hierarchical text encoding scheme, research has also been expanded to incorporate transformer-based text encoding, integrating the sentence-transformers [74] into the model. In particular, the hierarchical models are replaced with pre-trained sentence-transformer models to directly generate sentence-level embeddings $\{Q_1, Q_2, \cdots, Q_N\}$ and $\{A_1, A_2, \cdots, A_N\}$. Furthermore, all encoder layers, *Global Encoder* and *View Transformer* (*View Encoder* in hierarchical setting), are defined using transformer-based Multi-head Attention Networks [86] in place of self-attention model used in hierarchical architecture. Cross attention within view transformer layers is also defined based on the multi-head attention mechanism with patient and therapist inputs playing the corresponding roles of query, key, or value based on the situation (highlighted with corresponding colors in figure 3.4). Within this setting, we only consider the best-performing multi-view configuration among the different ablation studies performed with hierarchical text encodings. Thus, only *ST-Baseline* model (based on *Hierarchical-Baseline (Patient+Therapist)* configuration) and *ST-MV* model (based on Inter-view Attention (Mean) configuration) are implemented within the context of this experiment. Figure 3.3 illustrates the proposed *ST-Baseline* configuration that acts as the baseline architecture within this experiment, while figure 3.4 illustrates the multi-view based *ST-MV* model.

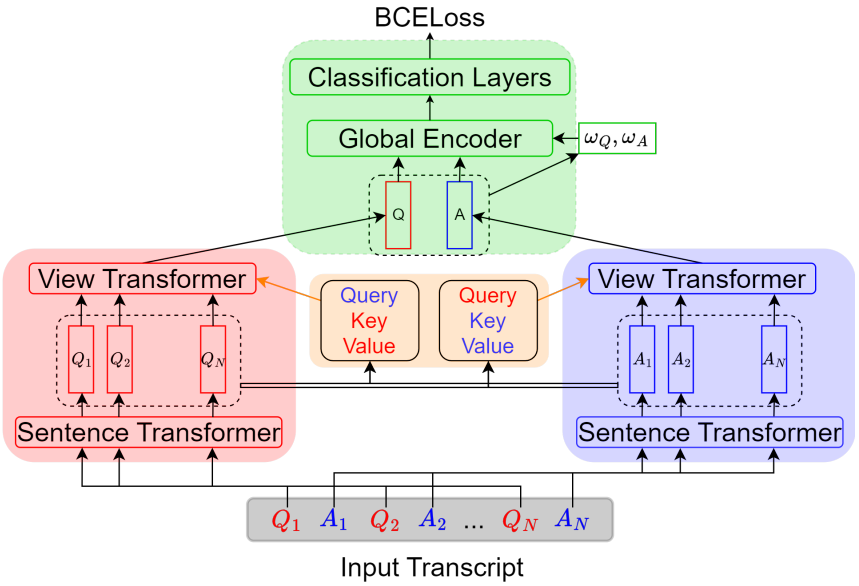


Figure 3.4: Transformer based multi-view architecture using sentence-transformer as text encoder.

3.2 Implementation Details

In hierarchical text encoding setting, we use pre-trained GloVe embeddings (300D) [64] for word encodings[5]. Adam optimizer is utilized with a learning rate of $5 * 10^{-4}$ and the binary cross-entropy (BCELoss) is the final loss function. Dropout is also applied with the probability of 40%. Within the experiment with sentence-transformer based text encoding, we use *all-mpnet-base-v2* model for generating sentence-level text embeddings. Adam optimizer with weighted binary cross entropy loss (BCELoss) is used during training to account for class imbalance in data, along with a learning rate of $5 * 10^{-4}$, and dropout of 0.4. The cross-attention mechanism in view transformer layers uses transformer based multi-head attention mechanism with 2 attention heads, whereas, Global Encoder layer employs 4 heads.

3.3 Results

Experiments are conducted on the Distress Analysis Interview Corpus - Wizard of Oz [38] (DAIC-WOZ) dataset that contains interviews between patients and a virtual therapist in a wizard-of-oz setting. The best model is chosen based on macro

[5]https://github.com/stanfordnlp/GloVe

Architectures	macro F1		UAR		Accuracy		macro Precision	
	(Dev)	Test	(Dev)	Test	(Dev)	Test	(Dev)	Test
Hierarchical Baseline								
Patient	(0.6413)	0.6429	(0.6369)	0.6361	(0.6969)	0.7608	(0.6725)	0.6584
Therapist	(0.8253)	0.5818	(0.8095)	0.5803	(0.8484)	0.6521	(0.8611)	0.6184
Patient+Therapist	(0.7555)	0.6053	(0.7440)	0.6004	(0.7878)	0.6739	(0.7847)	0.6250
MV-Intra-Att.								
View-Global Attention	(0.6944)	0.6811	(0.6845)	0.6674	(0.7575)	0.7391	(0.7870)	0.7252
Global Attention	(0.6857)	0.7116	(0.6785)	0.7075	(0.7272)	0.7173	(0.7083)	0.6887
View Attention	(0.6944)	0.6919	(0.6845)	0.6919	(0.7575)	0.6739	(0.7870)	0.6919
MV-Inter-Att.								
Patient	(0.5460)	0.5719	(0.5476)	0.5736	(0.6060)	0.6956	(0.5555)	0.5709
Therapist	(0.7664)	0.5710	(0.7619)	0.5691	(0.7878)	0.6304	(0.7727)	0.5759
Max	(0.6616)	0.5801	(0.6845)	0.5982	(0.6666)	0.6304	(0.6709)	0.5846
Mean	(0.6857)	0.7319	(0.6785)	0.7232	(0.7272)	0.7173	(0.7083)	0.7450
Learnable	(0.6434)	0.6043	(0.6428)	0.6093	(0.7272)	0.4782	(0.7571)	0.6020
Sentence-transformer based configurations								
ST-Baseline	(0.79)	0.75	(0.78)	0.75	(0.82)	0.80	(0.82)	0.77
ST-MV	(0.77)	**0.80**	(0.76)	**0.83**	(0.80)	**0.82**	(0.79)	**0.79**

Table 3.1: Overall results over the DAIC-WOZ dataset. UAR stands for Unweighted Average Recall.

F1 scores calculated over the development set and performance is evaluated over the test set. Table 3.1 gives detailed results of the experiments, and figures show that multi-view architectures provide a better way of combining inputs from patient-therapist interviews as compared to sequential encoding. Multi-view models outperform the corresponding baseline architectures for both hierarchical and sentence-transformer based text encodings. In particular, hierarchical encoding based multi-view architecture *MV-Inter-Att. (Mean)* shows improvements of 13.84% on macro F1 score, 13.69% on Unweighted Average Recall (UAR), and 13.15% on macro Precision compared to the corresponding baseline configuration (*Hierarchical Baseline (Patient)*). Furthermore, the sentence-transformer based *ST-MV* model also outperformed the corresponding baseline (*ST-Baseline*) showing improvements of 6.6% on macro F1 score and 10.6% on UAR. As expected, the multi-view model with sentence-transformer based text encoding, *ST-MV*, evidenced the best-performing results of all configurations considered in our experiments, outperforming all the hierarchical encoding based models, as well as the sentence-transformer based baseline model (*ST-Baseline*).

The validation of pertinent information within therapist questions is confirmed by the outcomes achieved for the *Hierarchical Baseline (Therapist)* model, which exclusively employs therapist utterances as input. Additionally, comparing results

for different *Hierarchical-Baseline* models, we can argue that combining questions and answers as a sequence of sentences does not provide improvements over using just the patient's utterances as input (*Hierarchical Baseline (Patient+Therapist)* vs. *Hierarchical Baseline (Patient)*). We believe that the lack of structural information in the former input configuration plays an important role in restricting the learning ability of the baseline model. This is dealt with by the multi-view architecture definition, that incorporates the discourse structure in the learning process.

Comparing results obtained by the multi-view strategies with intra-view attention (*MV-Intra-Att.*) against the *Hierarchical-Baseline* models, we can assess that multi-view architectures are a better alternative for processing patient-therapist interviews. Indeed, all *MV-Intra-Att.* architectures provide significant performance improvements over the *Hierarchical Baseline* models for all 4 evaluation metrics considered. This highlights the significance of retaining structural information of a dialogue during training, rather than processing an unstructured sequence of sentences. In particular, multi-view architectures utilize the discourse structure and disregard the inter-view interactions within the discourse, thus reducing the number of noisy interactions and allowing the model to focus on relevant information for more efficient training.

Further results support our argument of co-dependence between questions and answers with the inter-view attention based multi-view model, *MV-Inter-Attention (Mean)*, outperforming all other architectures including the corresponding *Hierarchical-Baseline* and *MV-Intra-Att.* configurations. This proves that despite inter-view interactions being a major source of noise in the data, co-dependency between the corresponding questions and answers needs to be retained. However, this improvement does not stand for all cross-attention functions considered in this study. Indeed, we observe that results obtained with non-balanced attention functions (i.e. Patient, Therapist, Max) are lower compared to (1) the balanced attention functions (i.e. Mean, Learnable), and (2) all other configurations (i.e. *Hierarchical-Baselines* and *MV-Intra-Att.*). Within non-balanced definitions of function f, attention scores are transferred from one view to the other based on the hypothesis that only one of the two views drives the learning process. As such, these models represent the extreme case of cross-attention, where questions' (resp. answers) importance is directly based on corresponding answers' (resp. questions) importance while neglecting their own attention score. Results prove that both views, patient and therapist, play a role in defining their importance, and selecting either one as the sole criterion for importance can be counterproductive. Both models based on balanced definitions of $f(.)$

Architectures	Modality	macro F1		UAR	
		(Dev)	Test	(Dev)	Test
Raw Audio [7]	Audio	(0.66)	-	-	-
SVM:m-M&S [20]	All	(0.96)	0.67	-	-
HCAG [59]	Text + Audio	(0.92)	-	(0.92)	-
HCAN [55]	Text	(0.51)	0.63	(0.54)	0.66
HLGAN [55]	Text	(0.60)	0.35	(0.60)	0.33
HAN [92]	Text	(0.46)	0.62	(0.48)	0.63
HAN+L [92]	Text	(0.62)	0.70	(0.63)	0.70
HCAG+T [59]	Text	(0.77)	-	(0.82)	-
Symptom prediction [56]	Text	(0.80)	0.74	-	-
ST-Baseline	Text	(0.79)	0.75	(0.78)	0.75
ST-MV	Text	(0.77)	**0.80**	(0.76)	**0.83**

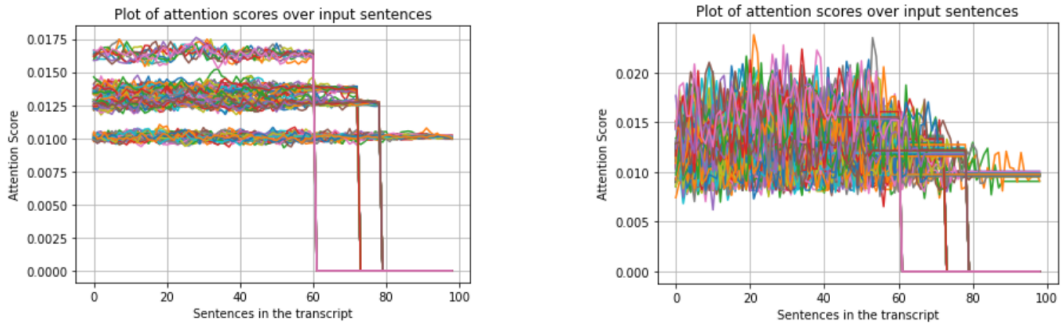
Table 3.2: Comparison of best Multi-view models against recent initiatives.

performed better than their non-balanced counterparts, with the best performance evidenced by the *Mean* configuration. We expected the *MV-Inter-Attention (Learnable)* model to perform on par with the *MV-Inter-Attention (Mean)* architecture, if not better. We believe that the small size of the dataset played an important role in restricting the model's ability to learn a more complex attention function, leading to reduced predictive performance.

Table 3.2, compares the best-performing model (*Linear-MV*) against recent research initiatives over the DAIC-WOZ dataset. Results show *Linear-MV* model successfully outperforming recent initiatives with comparable setups (HAN [92], HCAN [55]) as well as those relying on external knowledge (HAN+L [92]) or different modalities (SVM:m-M&S [20]). Note that the reported results are taken directly from the original papers and some related work surprisingly do not evidence results over the test split, such as HCAG and HCAG+T [59], although they perform highly on the development set.

3.4 LSTM vs. No-LSTM: An Analysis

Recurrent neural networks such as LSTM have traditionally been used as a way to encode text sequences [101]. Within the context of long text encoding, they form the core representation modelling of intermediate layers of hierarchical architectures [13, 92]. However, recent research [57] have shown limitations of using RNNs, especially when combined with attention models. Within this context, Mohankumar et al. [57] show that learned representations of LSTM have high conicity across time steps, which can lead to unreliable attention scores. High conicity refers to the fact that



(a) Model with LSTM.

(b) Model without LSTM.

Figure 3.5: Plots of attention scores for the training data. Each color represents one interview transcripts. Values are given for 4 different batches of 32 interviews.

element-wise intermediate representations are similar to each other, thus limiting diversity in representation expressiveness.

Similar findings have been evidenced in our experiments as we found that the LSTM-based intermediate representations of the hierarchical architecture have high conicity, thus leading to low variation in attention scores as illustrated in Figure 3.5a. Oppositely, we observed that our non-RNN based implementation of hierarchical models (§3.1.1), where attention scores are learned over the concatenation of embeddings, provides high variation in attention scores as shown in Figure 3.5b. As a consequence, best performance results for the classification task have been obtained with the non-RNN-based implementation.

3.5 Conclusion

This chapter presents initial research on understanding the importance of discourse structure within automated depression estimation tasks. We not only validate the importance of retaining the therapist's questions within model input in the context of ADD, but also show that a sequential combination of the two input streams, patient utterances and therapist utterances, is not ideal. *Multi-view* architectures are proposed for automated depression estimation, that treat patient-therapist interviews as a combination of two views (therapist questions and patient responses) rather than a single unstructured document. The underlying idea is not only to incorporate utterances from both agents within the model input but also to retain the discourse structure within the learning process for improved results. In particular, the presented multi-view approach allows models to handle discourse structure as well as

symmetry, thus outperforming models trained on simple sequential encoding of text.

Our experimental results showcase the importance of removing noisy interactions from the learning process of the neural network models. *MV-Intra-Att.* configuration perfectly exemplifies this with all configurations within this setting outperforming the corresponding baseline (*Hierarchical Baseline*) architectures. *MV-Intra-Att.* models disregard the inter-view interactions within the discourse and focus only on the intra-view relations. The success of *MV-Intra-Att.* configuration over the baseline models confirms the presence of irrelevant information in an unstructured sequence of sentences, consequently proving the relevance of incorporating discourse structure in the model training process.

Further improvements evidenced by the *MV-Inter-Att. (Mean)* model validate the co-dependency between the corresponding questions and answers in an interview. In particular, *MV-Inter-Att. (Mean)* configuration evidenced the best results, in the context of hierarchical text encoding, by taking into account both intra- and inter-view interactions. Other configurations examined within the *MV-Inter-Att.* category, non-balanced attention based configurations (patient, therapist, and max) in particular, perform poorly in comparison to the other definitions. This further strengthens our claim that both views play an important role in the decision-making process, and choosing either one as the sole criterion for defining sentence importance can be counter-productive.

Results on the DAIC-WOZ dataset show that the multi-view architecture steadily outperforms corresponding baseline architectures for both hierarchical and sentence-transformer based text encoding. They further show improvements over recent research initiatives in the field and not only provide improvements over models with comparable architectures (HAN [92], HCAN [55]), but also those using external knowledge (HAN+L [92]) or multiple modalities (SVM:m-M&S [20]). We plan to continue exploring the importance of discourse structure within the automated depression estimation paradigm and further investigate ways of encoding the input text for more efficient training. Another research direction can be to investigate the integration of the multi-view concept with external knowledge and multiple modalities.

Building on the need to incorporate discourse structures into the learning process of neural networks, the next chapter presents research done on graph based input representations of the interview transcripts. The study not only employs graph structures to improve the representation of non-linear interactions within interviews but also showcases insight generation through their visualization.

Chapter 4

Input Representations and Insight Generation

Conversations are the most prevalent form of human communication used to convey a wide variety of information and emotions. These dialogues represent highly complicated structures made up of complex interactions between various parts of the discourse. Dyadic conversations, in particular, embody intricate interactions wherein both participants not only articulate their thoughts but are also required to contextualize their responses according to other person's utterances. This is particularly true in the case of patient-therapist interviews where the meaning of the patient responses needs to be studied in the context of the corresponding therapist questions. Patient responses in turn influence the questions asked by the therapist to uncover signs of mental distress in their behavior. Within the context of Automated Depression Detection (ADD), researchers have been exploring novel techniques for better encoding patient-therapist interviews. However, their focus has mostly been towards defining neural architectures for improved predictive performance within the various sub-tasks under the ADD umbrella. Within such studies, defining better neural architectures takes center stage, while the input transcript is treated as an unstructured sequence of sentences. Their work reflects the ideology that given a complex enough architecture and training process, a neural network model can learn the intricacies of patient-therapist discourse (and language in general) from the given unstructured sequence of sentences. This strategy is perfectly exemplified by the recent success of Large Language Models (LLM) like GPT-3 [35], InstructGPT [62], GPT-4 [1] that are trained using enormous amounts of textual data scraped from the internet. These transformer [86] based models rely purely on their architectural complexity and ad-

vanced training methodologies for understanding language from sequentially encoded text. These models are capable of comprehending the intricacies of human language and showcase exceptional learning abilities. Although this represents a valid research paradigm, we propose a slightly different approach and explore the impact of input representation on models' predictive capabilities. We hypothesize that correct encoding of the input text can highlight latent features within the input, allowing a more efficient training of the neural network models with constrained architectural complexity.

Input representation plays a major role in defining the learning abilities of the neural network architectures. The initial representation of the input not only controls the features learned by the neural network during training but can also help encode the information in a way so as to facilitate the learning process. Appropriate representations bring out latent attributes within the input, potentially allowing more efficient training of the model with constrained architectural complexity. In the context of ADD research, many researchers utilize linear data structures to encode the input text, thereby hindering the model's ability to discern the structure and interactions within the discourse. Compared to sequential representations, graphs are a better choice of data structure for representing the inherent non-linear interactions that form the basis of human conversations. Graphs are discrete data structures, composed of nodes and edges, that model complex non-linear interactions within a given input. Furthermore, based on the definition of nodes and their interactions (edges), different graph structures can be used to highlight different aspects of the same input, thus providing different perspectives on the data.

Graph-based text representations have been explored in the literature with Graph Neural Network (GNN) being applied to a variety of Natural Language Processing (NLP) tasks ranging from classification tasks like sentence classification [42, 12], semantic role labeling [53, 39], and relation extraction [67, 79], to generation tasks like machine translation [8, 10], question generation [63, 77], and summarization [32, 99]. Building upon the success of graph neural networks and their use within NLP applications, progress has been made towards incorporating them within the mental health analysis domain [2, 93, 59, 43, 17]. Research based on clinical interviews, in particular, has seen researchers incorporate graph structures within their models to focus on different characteristics of the input data. Niu et al. [59] use graph structures within their architectures to grasp relational contextual information from both audio and text modalities. They propose Hierarchical Context-Aware Model [59] (HCAG) which can effectively capture and integrate contextual information among relational

interview questions at both word/frame and Question-Answer Pair (QA-pair) levels. HCAG models aim to mirror the sequential and hierarchical structure of a depression interview assessment and utilize graph structures to model QA-pair level interactions within individual interview transcripts. A detailed explanation of the model can be found in §2.4.1. While Niu et al. [59] focus on QA-pair level interactions, other researchers have explored word-word associations for depression estimation [43, 17]. Hong et al. [43] hypothesize that the context of words in a transcript can provide valuable knowledge for the mental health assessment of patients. As such, they focus on word-level interactions and define transcript-level graphs where each node represents a word and connections are based on co-occurrence patterns of words. They further propose a novel form of node attribute that captures node-specific embeddings for every word in the vocabulary. This provides a global representation at each node, coupled with node-level updates according to associations among words in the transcript. Along the same research direction, Burdisso et al. [17] also explore depression estimation based on word-level interactions. Contrary to Hong et al. [43] who define transcript-level graph structures, Burdisso et al. [17] construct a corpus-level graph that combines both word and transcript-level representations into a single graph. The proposed method aims to mitigate limiting assumptions of locality and the equal importance of self-connections vs. edges to neighboring nodes in GCN. Their graph structure simultaneously models word-word interactions and word-transcript interactions, allowing the network to learn transcript representations in the context of a corpus-level word interaction graph.

Within this section, initial experiments along a similar research direction are described, wherein we explore graph-based representations of patient-therapist interviews and study their impact on the ADD task. The primary focus is directed towards Sentence Similarity Graphs (SSG) and Keyword Correlation Graphs [19] (KCG), which highlight distinctive features across varying levels of granularity within the data. These graph structures not only highlight complex interactions within the discourse but also provide a perspective that does not exist within sequential data structures. Additionally, we also integrate the multi-view concept defined in chapter 3 within the graph representations. This extends the multi-view concept beyond its neural architectural definition and applies it directly to input representations, thus reinforcing its validity as a more generic concept. Moreover, this integration allows us to treat the inputs from the two views independently, further highlighting the difference in perspectives of the two agents. Finally, within this chapter, we also demonstrate that the visual representation of the graph structures considered in this

study can serve as a rapid visual synopsis of the discourse, while providing valuable insights that can be helpful to healthcare experts in their prognosis. Automated models for depression estimation, when deployed in the clinical setting, are meant to assist medical professionals in their diagnostic process rather than providing a final assessment themselves. The relevance of an automated model's final predictions within the mental health domain weakens in the absence of a comprehensible explanation of the underlying process. As such, it is imperative for AI models to provide knowledge or explanation regarding their decision-making process along with the final assessment. Although they don't constitute explanations, insights generated from graph visualizations can highlight the important aspects of the input text. Additionally, since these insights are generated directly from the input and not based on complex neural network interpretations, they are more reliable compared to attention models whose trustworthiness is debatable [44, 90].

4.1 Graph Structures and Learning Models

Graphs are discrete data structures that provide intuitive representations capable of not only capturing the non-linear connections within conversations but also defining intricate input representations that don't exist in a sequential setting. Graphs are defined as a collection of entities (nodes), and connections (edges) representing interactions between said entities. This generic interpretation of graphs does not include the definition of nodes (entities) or their interactions, allowing a plethora of graphical representations for a given textual input with each showcasing a different understanding of the text. As part of our initial study, we focus on static graph definitions that generate the graphical transcript representation at the pre-processing stage. This enables us to separate the generation of input representations from the final training of the neural network model, thereby guaranteeing that our input representations are not reliant on intricate model interpretations. Conceptually, a static graph incorporates different domain/external knowledge hidden in the original text sequence, which augments the raw text with rich structured information. Within this research we have chosen to work with two categories of graph structures; (1) Similarity based structure, *Sentence Similarity Graphs*, which is one of the most basic and widely used graphical representation of text and highlights local sentence-level interactions within transcripts; and (2) Topical graph structures, Keyword Correlation Graphs [19], that integrate corpus-level topical information within the graphical representation of individual transcripts. Multi-view concept is also integrated within the graph

definitions to further highlight the different perspectives and their interactions within the input discourse. In particular, we not only use the multi-view concept within our network definitions but also apply it to input graph representations. Based on the experimental results discussed in chapter 3 table 3.1 *MV-Inter-Att. (Mean)* configuration of multi-view architecture is used within the neural networks defined in this chapter. This configuration of multi-view architectures also inspires the multi-view based graph representations that divide the input into corresponding views and encode both intra-view and inter-view interactions within the discourse.

4.1.1 Sentence Similarity Graphs

Sentence Similarity Graphs are the most basic and most widely used graphical representations of text that highlight sentence-level interactions within the input. Individual sentences form the nodes of the graph and edges are defined based on cosine similarity between corresponding node embeddings (sentence representations). This graph structure focuses on encoding local sentence-level interactions within the discourse. Within this context, we define two configurations that explore both the generic definition of SSG and the multi-view infused interpretation. Figure 4.1 provides an overview of the architectures used within this setting.

Similarity-Baseline. We start with the generic definition of sentence similarity graphs where all sentences are treated equally irrespective of their identity (patient or therapist input). Edges are defined between all possible node pairs based on cosine similarity between corresponding node embeddings. A similarity threshold is applied to introduce sparsity into the graph structure and the value is treated as a hyper-parameter during training. GNN architectures are used for processing the input graph structure and the resulting node embeddings are passed through transformer-based multi-head attention [86] and classification layers to generate the final predictions. The resulting model is illustrated in figure 4.1(a).

Similarity-MV. Keeping in line with the multi-view idea, the transcript is divided into patient and therapist inputs. Individual sentence similarity graphs are defined for the two views (highlighted with red and blue colors in figure 4.1(b)) based on the definition used above. These individual graphs encode the intra-view interactions while cross-connections (highlighted in orange), represent edges between the corresponding questions and answers and model inter-view interactions. GNNs are used to process the resulting graph structure and the learned sentence embeddings are again divided into corresponding views and used as sentence-level input within the

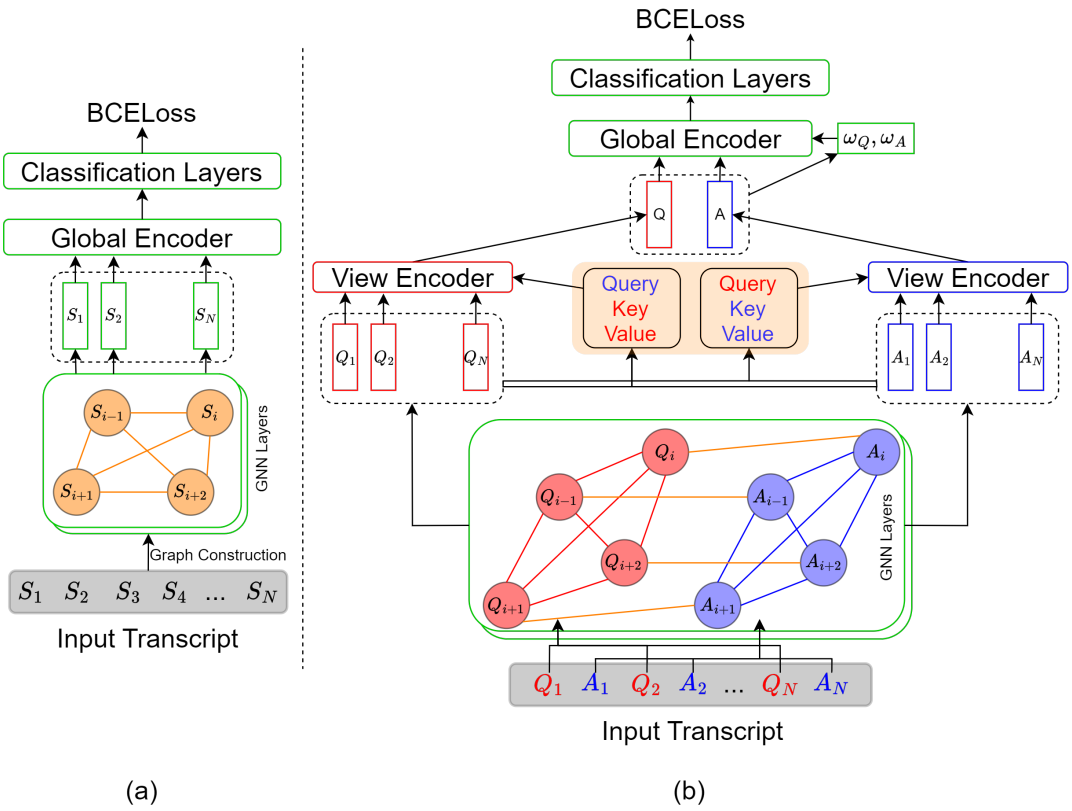


Figure 4.1: Overview of (a) Similarity-Baseline and (2) Similarity-MV architectures. Input color coding, red: therapist view, blue: patient view, orange: global nodes/cross connections, green: global network.

multi-view part of the neural architecture. The integration of multi-view concept with sentence similarity graph definition also allows us to address a limitation of the multi-view models defined in chapter 3. Corresponding questions and answers within dyadic interviews rely on each other to convey their true meaning. Although the shared-attention mechanism defined in chapter 3 accounts for the co-dependency between the attention scores of questions and answers, it does not encode their contextual co-dependency. Inter-view connections within *Similarity-MV* structure allow exchange of actual information between corresponding questions and answers within GNN layers, thus learning their contextualized meaning. *Similarity-MV* structures not only provide a different perspective of the input transcript but also incorporate the discourse structure better than multi-view architectures alone.

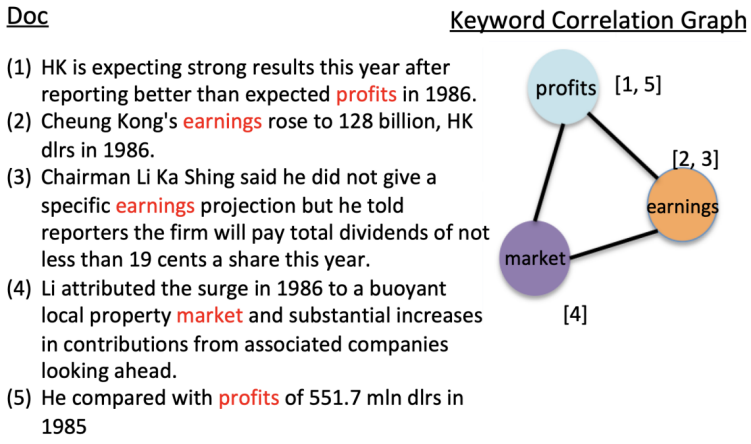


Figure 4.2: Graphic showcasing a document, its keywords (red) and KCG representation taken from [19]. Example adapted from the Reuters dataset [52].

4.1.2 Keyword Correlation Graphs

The Original Definition of KCG

Document clustering requires a deep understanding of the complex structure of long text; in particular, the intra-sentential (local) and inter-sentential (global) features which are not fully captured by most representation learning models. Most methods model text as bag-of-words or as sequences of variable-length units, and are ineffective in capturing global features. Chiu et al. [19] propose Keyword Correlation Graph structures that represent documents as weighted graphs of topical keywords and integrate global information into the input using learned topical knowledge. Within the KCG definition, topic modeling algorithms learn a set of global topics that are used for extracting important keywords within each document. These topic models are trained on the entire training data, and hence encode corpus-level understanding of the text which is then incorporated within the input graph structure of individual transcripts. Each node within the graph represents a keyword, with sentences in the document assigned to the node they are most related to. The edges between the nodes indicate their correlation strength calculated based on pair-wise cosine similarity between corresponding sentence sets. The construction process for KCG structures can be divided into 4 major steps:

1. Non-negative Matrix Factorization [33] (NMF) based topic modeling is used to learn the set of relevant topics from the training dataset. This topical representation of the documents is then used to extract the most important

keywords from each document, representing nodes of resulting KCG structure.

2. Each sentence in the document is then mapped to the keyword (node) it is most related to, thus generating

   sentence sets for each node (example in figure 4.2). TF-IDF features of sentences, calculated in the context of corresponding document keywords, are used to define the importance of nodes in individual sentences.

3. Sentence Transformers (SBERT) [74] are employed for generating vector representations of individual sentences. Each keyword node is then represented as the average of its sentence set embeddings.

4. Edges between nodes are defined based on pair-wise cosine similarity between the corresponding sentence embedding sets.

In their research, Chiu et al. [19] also conduct ablation studies and explore multiple definitions of various aspects within the proposed graph structure. The configuration defined within this section represents the best-performing definition from these experiments. Firstly, different text embeddings were explored including GloVe [64], Embeddings from Language Models [65] (ELMo), Bidirectional Encoder Representations from Transformers [26] (BERT) and Sentence-BERT (SBERT) [74] with SBERT out-performing all other methods. Further, two definitions of word-word interactions were explored: (1) based on word co-occurrence within a fixed window, and (2) based on pair-wise cosine similarity between corresponding sentence sets. Edges defined based on sentence similarity performed better than the other definition, possibly because text embeddings (like SBERT) already encode the local semantic relations between adjacent words and sentences, thus negating the impact of word co-occurrence edges. Given that SBERT is specifically trained for generating state-of-the-art contextualized sentence encodings, this observation is not surprising. Overall they define a KCG structure that can be used to highlight global corpus-level topical information and integrate it within individual document representations. They further define a Multi-task Graph Auto-Encoder based model and use it to showcase the effectiveness of KCG representations in document clustering tasks.

KCG Structure in Context of Proposed Research

SSG structures, although good at representing local sentence-level interactions within a transcript are ineffective in capturing global features. Within the context of patient-

therapist interviews for mental health assessment, topics discussed within each transcript (like work, family, children, living situation, etc.) belong to a larger finite set of topics shared across all interviews. This collection of topics constitutes a crucial set of information that mirrors a clinician's viewpoint on what is essential for depression assessment, drawing from their knowledge and experience. We propose to use KCG structures in order to combine this corpus-level knowledge with transcript level representations and define each interview as a graph of important topical keywords representative of the transcript. Due to the semi-structured nature of interviews, psychiatrists typically discuss every relevant aspect of a person's life within each interview. Consequently, to attain more distinct and differentiating topics, we focus on topical analysis at sentence-level, and train our topic models on the collection of sentences instead of using transcript-level text. As such, NMF topic model is trained on the collection of individual sentences within the training set with each sentence treated as an individual document within this step. These models capture overarching topical knowledge at the corpus level which is employed to deduce the significance of words within each transcript, with the top 50 keywords[1] being utilized as nodes within the graph structures. These keywords are representative of the most relevant topics discussed within individual transcripts, with the entire graph encoding interactions between the various topics. Within our experiments, node features are learned from the corresponding sentence set embeddings during training for a task-specific representation, rather than applying an averaging operation over pre-trained sentence embeddings. *Node Encoder* layer, defined using transformer based multi-head attention architecture [86], is used to combine sentence embeddings within corresponding sentence sets to generate learned node representations. Keyword interactions are defined using average pairwise cosine similarity between their corresponding sentence set embeddings. As in the previous case of SSG structures, multi-view inspired KCG structures are also defined, that learn independent topic models for the two views, highlighting the difference in perspective of the two agents. Figure 4.3 shows an overview of different configurations used within this context.

KCG-Baseline. For the baseline configuration we treat all sentences equally and train a single topic model on collection of all sentences within the training set. This global topic model represents combined topical knowledge from both therapist and patient inputs and is used to generate a single KCG structure for each

[1]This choice is based on results from [19], and has not been tuned to a task-specific value in our current experiments.

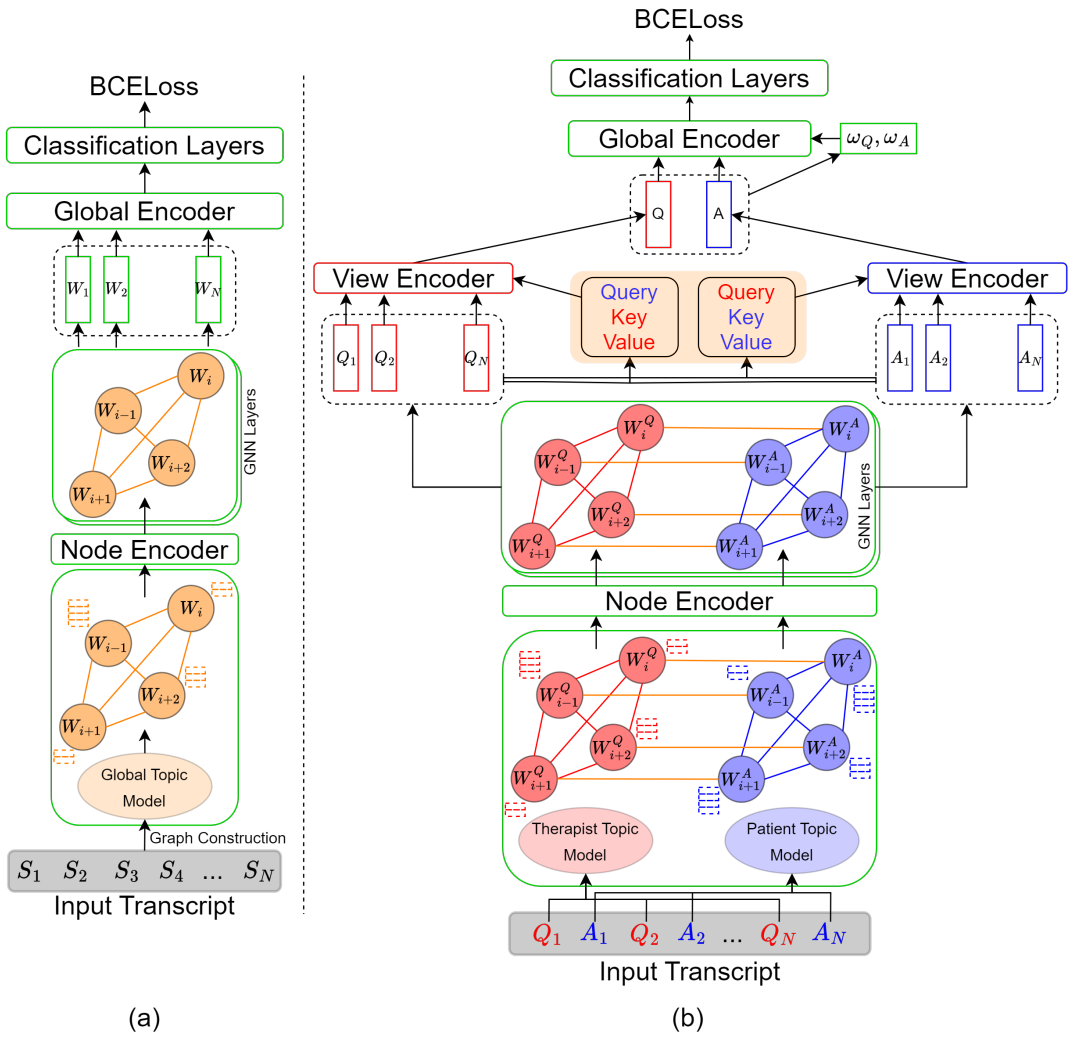


Figure 4.3: Overview of (a) KCG-Baseline and (2) KCG-MV configurations. Input color coding, red: therapist view, blue: patient view, orange: global nodes/cross connections, green: global layers acting on combined input.

transcript. *Node Encoder* layer learns node embeddings by combining sentence encodings from the corresponding sentence sets of nodes. GNN based architecture similar to *Similarity-Baseline* configuration is used to further process this graph and generate final prediction. An overview of the model is illustrated in figure 4.3(a).

KCG-MV. Although topics are shared across interviews, within this configuration we explore the possibility that the two views of the data represent complementary topical information. Although both patient and therapist are part of the same conversation, they have different purposes during the interview. A therapist participates

in an interview with the aim of assessing the patient's mental health, and usually relies on a pre-defined line of questioning and topics that need to be discussed during the interview. However, patients do not have such well defined objectives and mostly focus on responding to the therapist's questions. As such, topics learned from only therapist questions represent aspects of patient life relevant towards mental health assessment, while topics learned from patient inputs might be better suited for understanding their emotions and feelings. Within this setting, transcripts are divided into patient and therapist inputs and individual NMF based topic models are trained for each one. Training independent topic models for both views allows us to encode the difference in their perspective to a higher degree. Cross connections within this configuration are used to study patients feeling with regard to different aspects of their life and are defined based on the presence of corresponding question and answer in the sentence sets of the nodes. A neural network architecture similar to *Similarity-MV* is used to process the resulting graph. Figure 4.3(b) shows an overview of the architecture.

4.2 Experimental Setups

Sentence-transformers [74], all-mpnet-base-v2 in particular, are used for generating sentence-level text encodings within the experiments. Weighted binary cross entropy loss (BCELoss) is used to account for the class imbalance, coupled with Adam optimizer with learning rate of $5 * 10^{-4}$ during training. Similarity threshold has been applied to introduce sparsity into SSG structures, and the value is treated as hyper-parameter during training. The GNN layers used within all the model definitions are based on GCN. Other GNN definitions were experimented with but GCN was chosen since it provided more stable training[2]. All encoder layers, *Global Encoder*, *View Encoder* and *Node Encoder*, employ self-attention based on Multihead Attention Networks [86]. Cross-attention at *View Encoder* level within the multi-view architecture is also defined using multi-head attention networks with inputs from the two views playing respective roles among query, key and value as per requirement. Pytorch and Pytorch Geometric [34] frameworks are used for network definition and training.

[2]This is further discussed in §4.3.2

4.3 Results and Analysis

Configuration	macro F1	UAR	Accuracy	macro Precision
ST-Baseline	(0.79±0.04)	(0.78±0.03)	(0.82±0.03)	(0.82±0.04)
ST-MV	(0.77±0.02)	(0.76±0.03)	(0.80±0.02)	(0.79±0.02)
Similarity-Baseline	(0.71±0.00)	(0.70±0.00)	(0.76±0.00)	(0.75±0.00)
Similarity-MV	(0.76±0.0)	(0.74±0.0)	(0.79±0.0)	(0.78±0.0)
KCG-Baseline	(0.67±0.03)	(0.66±0.03)	(0.73±0.02)	(0.73±0.04)
KCG-MV	(0.66±0.02)	(0.65±0.02)	(0.72±0.01)	(0.72±0.02)

Table 4.1: Overall results over the development set of DAIC-WOZ dataset. UAR stands for Unweighted Average Recall. The best model is chosen based on F1(macro) values over the development set.

Configuration	macro F1	UAR	Accuracy	macro Precision
ST-Baseline	0.75±0.04	0.75±0.04	0.80±0.03	0.77±0.04
ST-MV	0.80±0.02	**0.83±0.02**	0.82±0.02	0.79±0.02
Similarity-Baseline	0.77±0.03	0.77±0.04	0.81±0.02	0.78±0.03
Similarity-MV	**0.81±0.01**	0.82±0.01	**0.83±0.01**	**0.80±0.01**
KCG-Baseline	0.68±0.01	0.69±0.01	0.72±0.01	0.68±0.01
KCG-MV	0.76±0.03	0.74±0.03	0.81±0.02	0.80±0.02

Table 4.2: Overall results over the test set of DAIC-WOZ dataset. UAR stands for Unweighted Average Recall. The best results over the test set are highlighted.

Tables 4.1 and 4.2 provide detailed results for all configurations considered in this study. Models are chosen based on macro F1 scores on the development set and performance on both the development (table 4.1) and the test set (table 4.2) are reported. Mean and standard deviation are calculated over 5 random initializations of the models. In order to establish a sequential baseline, we also include results from multi-view experiments (chapter 3) with comparable neural network architectures applied to linear input configuration (*ST-MV* and *ST-Baseline*). Figures prove that graph-based representation of transcripts provides better and more stable performance compared to sequential representation. In particular, *Similarity-MV* representation evidences best-performing results on the test set for 3 out of 4 evaluation metrics outperforming all other configurations considered in our research. Moreover, for both SSG and KCG based input representations, multi-view infused configurations outperform their corresponding baseline models for all evaluation metrics.

Architectures	Modality	macro F1		UAR	
		(Dev)	Test	(Dev)	Test
Raw Audio [7]	A	(0.66)	-	-	-
SVM:m-M&S [20]	T+V+A	(0.96)	0.67	-	-
HCAG [59]	T+A	(0.92)	-	(0.92)	-
HCAN [55]	T	(0.51)	0.63	(0.54)	0.66
HLGAN [55]	T	(0.60)	0.35	(0.60)	0.33
HAN [92]	T	(0.46)	0.62	(0.48)	0.63
HAN+L [92]	T	(0.62)	0.70	(0.63)	0.70
HCAG+T [59]	T	(0.77)	-	(0.82)	-
Symptom Pred. [56]	T	(0.72)	0.74	-	-
ST-MV	T	(0.66)	0.80	(0.65)	**0.83**
Similarity-MV	T	**(0.76)**	**0.81**	**(0.74)**	0.82

Table 4.3: State-of-the-art results on DAIC-WOZ. T, V and A stand for Text, Visual and Audio modalities respectively. Note that the reported results are taken directly from the original papers, and some related work surprisingly do not evidence results over the test split, such as HCAG and HCAG+T [59], although they perform highly on the development set.

From Table 4.3, we further show that our best-performing model (*Similarity-MV*) provides new state-of-the-art results over the DAIC-WOZ dataset, outperforming recent initiatives including those relying on external knowledge (HAN+L [92]), different modalities (SVM:m-M&S [20]) or multi-target learning (Symptom Prediction [56]). Figure 4.4 also proves that graph-based models not only provide state-of-the-art results, but also have a stable learning curve, which is a desirable property for applications in the medical domain.

4.3.1 Comparing Sequential and Graphical Representations

Comparing sequential and Sentence Similarity based models, we see clear improvements with graphical representations for both *Baseline* and *MV* configurations. Specifically, *Similarity-Baseline* outperforms *ST-Baseline* by 2.6% on macro F1 score while *Similarity-MV* outperforms *ST-MV* by 1.2% for the same metric. Overall, *Similarity-Baseline* outperforms *ST-Baseline* for 4 out of 4 metrics while *Similarity-MV* evidences better results than *ST-MV* for 3 out of 4 metrics. The results underscore the advantage of graphical models in not only improving predictive performance but also demonstrating greater stability (see figure 4.4) compared to sequential models. This makes them more reliable for applications in the medical domain.

Both KCG based input representations (KCG-Baseline and KCG-MV) performed poorly compared to sequential and SSG based approaches. As mentioned earlier,

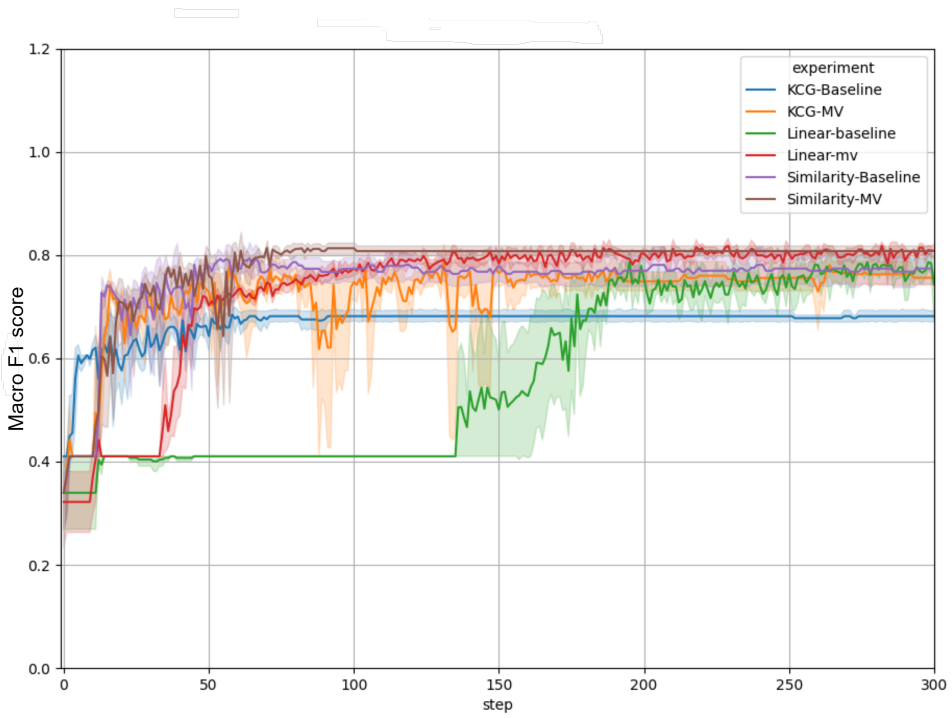


Figure 4.4: Plot of F1-score (macro) against epochs for different configurations on test set.

questions and answers within mental health assessment interviews rely on each other for contextual information that plays a vital role in defining their meaning. Although the structural definition of KCG allows them to model corpus level knowledge that is missing from SSG structure into the input graphs, they lack in their ability to represent local sentence level interactions. This lack of local information significantly restricts the learning ability of the model, especially for this data where the two views (questions and answers) share a strong contextual bond. This is evident from the results (table 4.2) where KCG based configurations systematically under perform compared to other models, thus showing the strong co-dependence between corresponding questions and answers in defining their meaning. Its important to note that within this part of the research, KCG representations are used only to exemplify different possibilities within graphical input representations and showcase their ability to highlight various aspects of the interview. Results shown for KCG based inputs use same configurations as defined by Chiu et al. [19] and have not been tuned to task specific values. These models contain various aspects, ranging from graph parameters like number of key-words, and definition of cross-connections, to

topic model parameters like number of topics, text pre-processing methods, etc., that need to be tuned to data specific configuration. Both views in KCG-MV configuration can further have their own set of hyper-parameters, thus increasing the number of possibilities. Although we expect the results to improve with configurations tuned specifically for mental health assessment (and DAIC-WOZ dataset specifically), it requires detailed research into the various aspects of KCG definition which is part of our ongoing and future research efforts.

4.3.2 Comparison of Baselines With Multi-view Graph Structures

Within our experiments, multi-view based representations steadily outperform the corresponding baseline configurations for both sentence similarity graphs and keyword correlation graphs. In particular, *Similarity-MV* configuration outperforms the *Similarity-Baseline* configuration by 5% on macro F1, 6% on UAR, 2% on Accuracy and 2% on macro Precision. SSG structures focus on local sentence level interactions where the context shared between the corresponding questions and answers plays an important role. This strong dependency between the two views restricts the graphs ability to highlight the distinctions in their individual perspective, thus restricting the advantage of incorporating multi-view concept within this framework. Yet, their capability to capture interactions at the sentence level enables them to learn sentence representations that are contextualized within their surroundings, resulting in a more robust understanding of the interview.

Compared to sentence similarity graphs, KCG representations show greater improvements when combined with multi-view architectures. *KCG-MV* model outperforms *KCG-Baseline* by 11% on macro F1, 7% on UAR, 12% on Accuracy and 17.6% on macro Precision. This jump can be attributed to the fact that KCG representations focus on keyword interactions rather than sentence interactions, allowing the model to better integrate view specific features into the graph structure. Since Keyword Correlation Graphs represent both views as interactions between topical keywords rather than sentences (question-answer interactions in particular), they avoid contextual dependency between the views and learn a more independent representations of the two views. Additionally, the training of dedicated topic models for the two views enhances the independent encoding of view perspectives. This can be verified by studying topics learned within the two configurations (figure 4.6) with a detailed discussion in §4.4.2. This property of Keyword Correlation Graphs is aligned with the multi-view idea resulting in significant performance gain by integrating the two approaches.

Finally, experiments have been carried out with different dense input representations including hierarchical models and BERT-based input embeddings. Among the different combinations and configurations, sentence transformer based embeddings evidence the best and most stable results. Different implementations for GNN's were also explored including GAT [87], GIN [94], and GraphSAGE [40], with all configurations providing similar performance. This behaviour is inline with the findings of Dwivedi et al. [28] who show that different implementations of GNN's evidence similar predictive performance when applied to small datasets. Within our experiments, GCN based implementation of GNN were chosen since they provided most stable results across different initialisation.

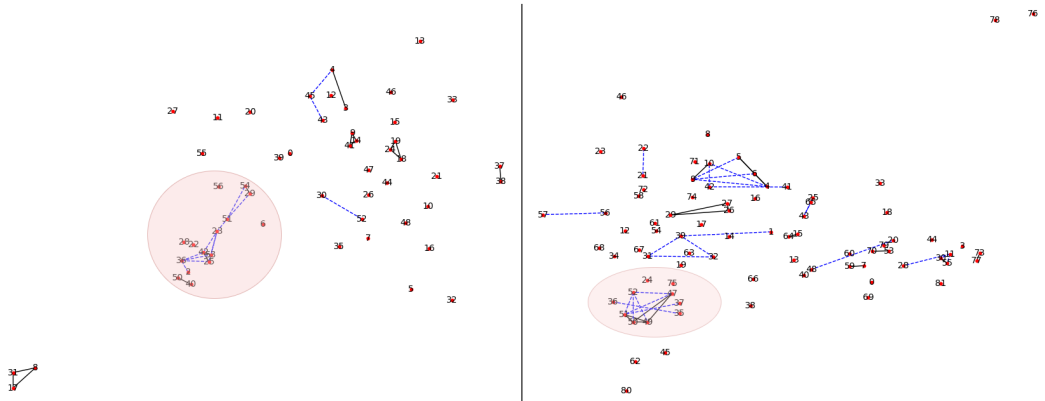
4.4 Visualization and Insights

Although current deep learning models provide excellent results, explaining their predictions is still a challenging task. Attention scores are widely used as a tool to justify model predictions, however validity of these explanations is debatable [90, 44]. In healthcare applications, despite their high performance, there is a reluctance to adopt black box neural network models. Instead, medical professionals are more inclined towards models that provide justification for their predictions rather than focusing solely of performance.

Our research not only aims to show the advantages of using graph-based interview representations towards predictive ability of the models, we also motivate the notion that input representations themselves can be used for insight generation. The aim is not to provide explanation of model predictions, but rather to use visualizations of the input graphs as a quick visual summary of the transcripts to be used by medical professionals. These visualizations can highlight information within the transcripts that might be relevant for healthcare professionals, and present it in an easy to comprehend manner. We explore this possibility in the context of SSG and KCG structures, and present our findings.

4.4.1 Sentence Similarity and Therapist Behaviour

Figure 4.5 shows visualization of sentence similarity graphs based on therapist inputs for patients with different PHQ-8 scores. Each node in the graph represents a question and the numbers are their corresponding position in input sequence. Clusters highlighted in red comprise of conversation fillers and one word responses used by the therapist that can be ignored for this analysis. Comparing the remaining clusters, we



(a) PHQ-8 score 0 (Therapist input)

(b) PHQ-8 score 18 (Therapist input)

Figure 4.5: Sentence similarity graphs based on therapist inputs for different PHQ scores. Blue dashed line represent weak correlations, while black solid line represent strong correlation.

find more descriptive graphs for people with high depression scores as compared to patients not suffering from depression. People with depression can have a tendency to be more reserved and usually give short and precise answers, forcing the therapist to ask more detailed questions. This is evident from the presence of elaborate clusters within figure 4.5(b), where each cluster represents therapist questions regarding relevant aspects of a patients life including work, relationships, children, etc. A contrasting view is observed for patients without depression, figure 4.5(a), where we see a significant lack of clusters within the graph. This is usually due to presence of more detailed answers by the patient, allowing the therapist to avoid detailed questions and rely on conversation fillers to sustain the interaction. These visualizations of sentence similarity graphs highlight the subtle differences in therapist's behaviour when interacting with patients having different severity of depression, which in turn can be an indicator of patients mental health.

4.4.2 KCG Structures and Global Viewpoints

Within the context of clinical interviews, patient and therapist have different motives for attending the interview and consequently their respective interventions can provide complementary information about the same discourse. Topic models trained within the KCG definition encapsulate corpus-level knowledge of the input text. Consequently, view-specific topic models more effectively emphasize the difference in perspective between the two agents engaged in an interview compared to a topic

(a) Combined input (b) Patient input (c) Therapist input

Figure 4.6: Topics learned with different inputs.

model trained on combined input. Figure 4.6 shows the topics learned within the KCG-Baseline (combined input) and KCG-MV (patient input and therapist input) configurations. In order to understand patients' mental health, it is desirable to study the emotions and feelings associated with important aspects of their lives. This involves learning both, the global set of the relevant topics (representing important aspects of a persons life) and patient attitude towards them. Comparing the different topic models in figure 4.6, we clearly see that topics based on therapist inputs (figure 4.6(c)) are better suited for representing the various aspects of a person's life that have relevance in depression estimation. We see distinct topics representing sleep (topic 1), family (topic 2), positive influence (topic 0), military service (topic 5), change in behavior (topic 8), p_t_s_d and past diagnoses (topic 9), which correlate with information desired by medical professionals. Within such interviews, therapists usually have a methodical approach towards the interview trajectory, as reflected in the clearly defined topics derived from their inputs. Conversely, patients have a slightly less pronounced role in defining the structure of discourse, primarily responding to topics chosen by the therapist, figure 4.6(b). Although topics learned on the combined input, figure 4.6(a), contain information on both views, they don't provide a complete knowledge of either and lack specific topics representing impor-

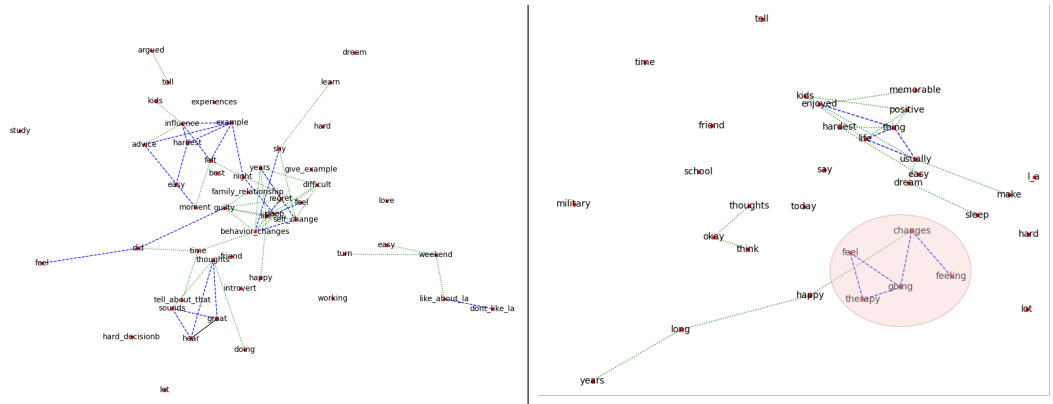
tant characteristics within each view.

4.4.3 KCG Structures and Transcript Level Visualization

Another interesting trait of KCG representations can be seen in the visualizations of individual transcript level graphs. Since these graphs are defined in terms of interactions between most relevant keywords within individual interviews, their visualization can act as a topical summary of the transcript. Within the context of structured interviews, most global topics are discussed in each transcript albeit with varying importance depending on each patient's situation. KCG structures utilize this fact by selecting the most important keywords within the transcript in order to highlight the subtle patterns that can be indicative of patients mental health. An example is shown in figure 4.7 that compares graph visualizations for two patients with different depression scores. For patient with high depression score, keywords like therapy, changes and feeling are clustered together (highlighted in red) while being absent from the graph of patients with low depression scores. Further analysis of the entire dataset reveals a pattern where keywords like "depression", "p_t_s_d", and "therapy" frequently appear as clusters in graphs of patients with high depression scores (score $\geq$ 18) while generally being absent from graphs of non-depressed patients. This highlights the fact that although topics like "depression" and "p_t_s_d" are discussed in most interviews, they are more relevant in the context of patients with high depression as compared to those without depression. Consequently, their presence in the graph can be indicative of depressive tendencies and can be easily highlighted within KCG visualizations. These visualizations illustrate the relative importance of different topics discussed within the interview, which in turn can be an indication of a patient's mental health.

4.5 Conclusion

In this chapter, we argue that the correct representation of the input can not only play a vital role in defining the learning abilities of neural network models, but also highlight desirable features within the transcripts. In particular, this research initiatives is aimed at three major goals in context of automated depression estimation based on DAIC-WOZ dataset: (1) investigate graph based representation of patient-therapist interviews for improved performance based on a more efficient and informative data representation, (2) study the incorporation of multi-view concept within the graph definitions to better highlight the difference in perspective between



(a) PHQ-8 score 0

(b) PHQ-8 score 20

Figure 4.7: KCG representations of transcripts with different PHQ-8 scores.

patient and therapist, and (3) explore the use of graph visualizations as a visual synopsis to be used by the medical professionals. We directed our focus on static graph structures and investigated our objectives within the context of Sentence Similarity Graphs and Keyword Correlation Graph structures.

Experimental results show that graph based representation, *Similarity-MV* in particular, evidences best-performing results, out-performing the corresponding sequential configuration *ST-MV*. In our experiments, graph-based models outperformed the corresponding linear configurations for both baseline and multi-view based definitions. This shows that graph based text encoding are better at representing the inherent non-linear interactions within conversations as compared to the widely used sequential encoding of input text. The *Similarity-MV* structure defined in this chapter also provides a more robust way to incorporate the discourse structure within the learning process as compared to multi-view models discussed in chapter 3. *Similarity-MV* allow the exchange of contextual information through graph edges rather than relying on shared-attention mechanism to encode inter-view interactions as was the case in previous definitions. Our experiments further demonstrate the ability of graphical representations to focus on different aspects of the input based on the definition of nodes and edges in the graph. Within this context, SSG structure is used to focus on local sentence-level interactions within the transcript and encode discourse structure into the input representations. KCG based structures on the other hand focus on global understanding of the dataset and encode corpus-level topical knowledge into individual transcript representations.

We also experimented with the integration of multi-view concept with graphical

models both at neural architecture level and graph definition. Figures in table 4.2 show that multi-view based configurations out-perform the baseline models for both graph structures considered in our study. *KCG-MV* based architectures show considerable gains in predictive performance, compared to KCG-Baseline, across all evaluation metrics considered in this study. This stems from the fact that KCG structures focus on topical information within the data rather than sentence-level interactions, allowing a greater degree of independence between the two view representations. On the other hand, sentence similarity graphs are not able to fully utilize the multi-view concept due to their high reliance on sentence-level interactions (question-answer interactions in particular). Nevertheless, our results conclusively establish multi-view as a more generic concept than the neural architectural definition given in chapter 3 and support the idea of it being a model-agnostic concept.

Finally, we demonstrated how visualizations of the input graphs defined in our experiments can provide valuable insights into the decision making process of AI models. Our analysis exemplifies these insights in context of Sentence Similarity Graphs and Keyword Correlation Graphs. We demonstrate how sentence similarity graphs based only on therapist utterances can point towards subtle changes in therapist behaviour, which in-turn can be an indication of patient's mental health. KCG structures, on the other hand, can be used as a quick visual summary of a transcript's topical distribution. Although the semi-structured nature of these interviews means that most topics appear in all transcripts, their relative importance varies on a case-by-case basis and can be used as an indication of patient's mental health. For example, questions related to depression and p_t_s_d are asked in almost all interviews but keywords depression and p_t_s_d only appear in graphs of depressed individuals. The ability of our models to not only provide a good predictive performance but also insights into their decision-making process. This makes them ideal for use in clinical settings where medical professionals favor explanation and insights over the sheer predictive prowess of the model. We also compare our models against recent research initiatives in the field (table 4.3), and show that *similarity-MV* configuration evidenced the best results on the test set of DAIC-WOZ dataset. Our model was able to outperform even the initiatives using external knowledge (HAN+L [92]) and multiple modalities (SVM:m-M&S [20]).

We plan to continue in this research direction and further explore the graphical representation of patient-therapist interviews to generate relevant insights for medical professionals. More specifically, we intend to delve into a more comprehensive investigation of KCG definitions to enhance the learned topics and fine-tune the

hyper-parameters for improved representations and predictive performance. We also plan to work in close collaboration with clinicians to understand their views and diagnostic processes. This would allow us to define graph structures capable of generating more valuable knowledge that can be incorporated in their decision-making process.

The next chapter discusses our initial efforts to incorporate medical expertise into the learning process of our neural network models. To this end, clinical annotation of the DAIC-WOZ dataset is endeavoured, and neural networks are trained with and without the inclusion of domain expertise. The proposed models are further analyzed in the context of these annotations to analogize the psychological tendencies of the models with those of trained medical professionals to strengthen their reliability and trustworthiness as predictive tools within the healthcare system.

Chapter 5

Depression Estimation and Psychiatric Expertise

Valid and reliable tools for automated mental health assessment from text and speech can prove groundbreaking. These systems can provide new opportunities for early detection and intervention, with the active involvement of medical professionals, covering a wider population. Their integration with mobile devices and computers via the internet can effectively make mental health assessment and care accessible to the majority of the world population, thus removing geographical constraints. Advances in AI, particularly its applications in the mental health domain, have led to significant progress in this direction with deep learning models reporting compelling accuracy rates. Given the widespread impact and heavy toll of depression, it is not surprising that automated depression estimation has been the focus of significant research initiatives [56, 68, 92, 59, 43, 20, 55]. Many of these modeling efforts are inspired by a long-term common vision of an end-to-end, automated system which can even be deployed in clinical settings.

Despite the extensive list of research initiatives undertaken in recent years, all these studies treat automated depression estimation as a purely computational problem with little to no involvement of medical professionals within their research paradigm. Given the multi-disciplinary nature of the task, omitting clinicians from the learning process implies disregarding a substantial source of domain expertise. This lack of consideration not only limits the learning capabilities of neural network models but also impedes their acceptance as predictive tools within the healthcare domain. Indeed, the absence of medical professionals within the research process means computational models for depression estimation are often disconnected from

the lived experience and siloed from the larger debates on how to characterize and classify mental health.

One major factor contributing towards this absence of domain knowledge within the learning process is the lack of relevant resources and availability of information. In the context of clinical interviews for depression estimation, the final assessment requires the therapist to discuss personal details of a patient's life. This creates confidentiality and privacy concerns resulting in an acute lack of publicly available datasets for depression estimation. The situation gets worse when looking for medical annotations for these datasets. Although DAIC-WOZ is the only publicly available dataset for depression estimation based on clinical interviews, clinical annotations for the dataset do not exist. Koehler et al. [4] undertake annotation of a subset of the DAIC dataset, however they employ crowd-workers for the process rather than medical professionals. In the context of social media based datasets, Yadav et al. [95] employ native English speakers from multiple disciplines for annotating their X (formerly Twitter) based dataset. This absence of medical expertise among the annotators undermines the credibility of these annotations.

In an attempt to encourage multi-disciplinary research within this field based on reliable and quality medical knowledge, we carried out clinical annotation of the DAIC-WOZ dataset. Our endeavor employs only medical professionals for the annotation process to maintain the reliability of the markings. We further analyze our neural network architecture in the context of these expert annotations and analogize the psychological tendencies of medical professionals against the proposed model in an attempt to validate its reliability as a predictive tool in clinical settings. This chapter first defines our annotation protocols and neural network architectures and then presents the results of our analysis.

5.1 Psychiatrist Annotations and Protocols

In recent years, there has been significant research interest in gathering and assimilation of domain expertise into automated models to exploit expert knowledge within the training process of neural network architectures. Within the context of automated depression estimation, Arseniev et al. [4] investigate the disconnect between computational models and clinical depression diagnostics. They experimented with DAIC-WOZ interviews and obtained layperson annotations of participants' mental health. In particular, crowd workers were employed to read excerpts of de-identified and transcribed interview data in order to evaluate their mental health. The ratings

symbolize how likely they thought a speaker had depression based on transcribed utterances with responses selected from "very likely", "likely", "unlikely", "very unlikely" or that there was "no evidence" either way for depression. Yadav et al. [95] explore a similar research path but in the context of social media data. They propose a novel multi-task learning framework to accurately identify depressive symptoms from tweets using the auxiliary task of figurative usage detection. Moreover, they created a dataset containing 12,155 tweets, including 3738 tweets posted by 205 self-reported depressed users over 2 weeks (the remaining 8417 tweets form the control group). The 3738 depression tweets were manually annotated using the PHQ-9 questionnaire [49] based symptom categories: lack of interest, feeling down, sleeping disorder, lack of energy, eating disorder, low self-esteem, concentration problem, hyper/low activity, and self-harm. Additionally, these tweets were also labeled with the figurative classes: *metaphor* and *sarcasm*. The annotation process was carried out by four native English speakers from multiple disciplines who independently annotated the tweets into 9 categories of PHQ-9. They were also asked to identify tweets having figurative language such as sarcasm and metaphor. For reference, the annotators were provided with the definitions and samples of annotated tweets from each of the 9 categories of PHQ-9 as well as figurative language. Conflicting annotations were resolved using a majority voting strategy and ones voted evenly were resolved by a psychiatrist. From the final gold standard data, 100 annotated tweets were randomly selected from each of the symptom categories, including the non-depressive ones, and verified by the psychiatrist.

In an attempt to reintroduce domain expertise into the learning process, we carry out the clinical annotation of the publicly available DAIC-WOZ dataset. Our aim behind this research is twofold: (1) to study the integration of medical knowledge within our neural network models and (2) to provide a publicly available resource to encourage multi-disciplinary research in the field of ADD.

In contrast to previous works that use crowd workers or native English speakers as annotators, we employed mental health professionals for the annotations, even though this decision prolonged the process. Given their academic and professional background in medicine, and mental health in particular, we consider their annotations to be more reliable and informative as compared to those from crowd-sourced annotators without a medical background. In particular, three psychiatrists from public hospitals were employed for the annotation process: one Ph.D. student, one junior doctor, and one senior doctor. Furthermore, the annotation process was divided into two major tasks: (1) span-based annotation of the transcripts and (2)

PHQ-8 scoring based on interview transcripts.

5.1.1 Span-based Annotations

This task consists of highlighting information within transcripts that influences a psychiatrist's decision during an interview. Since it is a subjective task that lacks a definitive right or wrong answer, a common consensus on the importance of various utterances within the transcripts does not exist. Even within the field of medicine, professionals do not universally agree on the significance of various pieces of information, and subtle differences in opinion exist between psychiatrists based on their knowledge and experience. As such, after various meetings and discussions with the psychiatrists, it was agreed that the medical annotators should have complete freedom to annotate the transcripts without any constraints in order to capture their true judgment. As a consequence, we forgo defining detailed annotation protocols and rely on the annotator's judgment as experts in the field for the reliability of their annotations. However, they were encouraged not only to identify information that suggests the presence of depression but also to pinpoint clues that indicate its absence, although no distinction was made between the two categories of markings. Unfortunately, at this stage of our research, only one annotator per transcript could be assigned due to the workload experienced by the annotators, particularly due to the radical increase in demand for mental care after the COVID pandemic coupled with the shortage of mental health professionals. The inherent lack of consensus within this subjective task combined with the lack of multiple annotators per transcript eliminates the need for inter-annotator agreements. In case multiple annotators are assigned per transcript, a simple union of annotated spans would be used to capture knowledge from all assigned annotators. Despite assigning only one annotator per transcript, the current annotation process lasted nearly 5 months and we anticipate this time frame to scale linearly with the increase in the number of annotators per transcript. Nevertheless, our choice of annotators ensures the quality of annotations given their extensive education and training in the mental health domain.

For the annotation purpose, an online tool based on the doccano[1] project was designed, which was hosted on servers from the heroku platform[2] enabling the entire annotation process to take place remotely for the convenience of the psychiatrists. The tool was designed to allow the psychiatrists to annotate any span of text (word,

[1] https://github.com/doccano/doccano
[2] https://www.heroku.com/

Span Level	Non-Depressed	Depressed
Word	467 (3.53)	227 (3.98)
Phrase	4101 (31.06)	1913 (33.56)
Sentence	0	0
Multi-sentences	77 (0.58)	42 (0.73)
Total	4645 (35.18)	2182 (38.28)

Table 5.1: Number of annotations for different levels of annotation spans. Figures in brackets indicate the average number of annotations per transcript.

phrase, sentence, etc.) within the transcript and assign a label of importance to each span: highly important, important (default), or minimally important. Upon analysis, it was found that these labels did not provide any valuable information with more than 99% of the spans marked with the default label (important), and were therefore not used for any further analysis. The annotation process gave rise to an average of 36.12 annotations per transcript (35.18 for the non-depressed class and 38.28 for the depressed class) with a mean length of 7.45 words (7.74 for the non-depressed class and 7.17 for the depressed class). The distribution of the annotations by patient class and span level is given in table 5.1. Interestingly, complete sentences were not annotated by any of the psychiatrists, who mostly followed an ngram-based strategy, with a small number of annotations focusing on multiple sentences. Furthermore, none of the psychiatrists highlighted questions within the dataset with all the annotations contained within patient responses. This behavior by medical professionals is inline with the use of only patient responses for prediction in initial ADD research (mentioned in chapter 3), and discussed further in §5.4.

5.1.2 PHQ-8 Scoring

This task involves the annotators completing the self-assessment Patient Health Questionnaire-8 (PHQ-8) on behalf of each patient only based on their interview transcripts. Although the PHQ-8 screening tool is widely used as a measure of depression and has been found to be precise [82], it relies on the subjective assessment by the patient of his/her condition outside the context of the interview. As such, an interview transcript might not contain enough information to accurately express the intensity of individual symptoms. Furthermore, since the interviews are conducted with the aim of depression estimation and not specifically for fulfilling the PHQ-8 questionnaire, information on some symptoms might be missing altogether within individual transcripts depending on the questions asked during the interview. In order to verify these propositions, we asked the clinicians to fulfill the PHQ-8

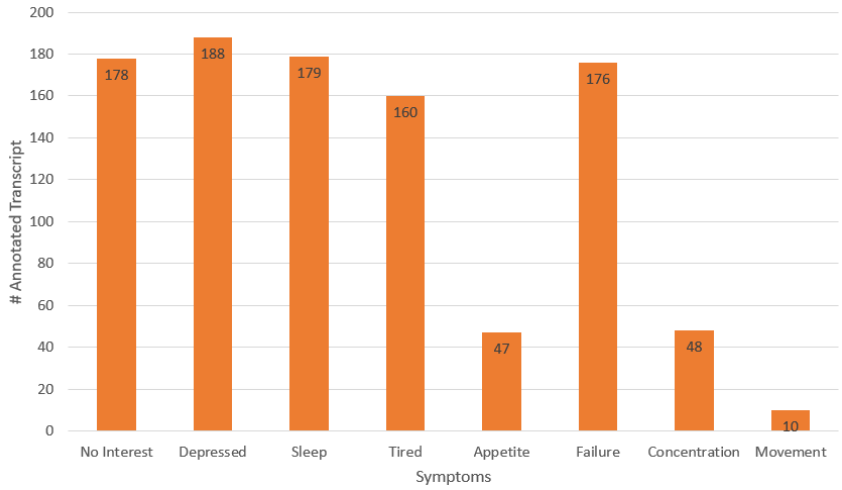


Figure 5.1: Number of transcripts scored for each PHQ-8 symptom out of the 189 interviews of the DAIC-WOZ.

questionnaires on behalf of each patient based on their understanding of the given transcripts. This task involves evaluating each of the 8 symptoms within the PHQ-8 questionnaire on a Likert scale ranging from 0 to 3. The statistics of this task, illustrated in figure 5.1, show that 5 out of 8 symptoms (i.e. loss of interest, feeling of depression, sleeping habits, feeling of tiredness, and feeling of failure) are steadily mentioned in most transcripts, while 3 of them (i.e. loss of appetite, lack of concentration and lack of movement) could not be measured reliably by the psychiatrists. This confirms our claims regarding the lack of symptom-level information within individual interviews. This annotation task also acts as a human baseline, that defines an achievable learning goal for correctly inferring PHQ-8 scores for each symptom based on information present within the transcripts.

5.2 Learning Model and External knowledge integration

5.2.1 Neural Network Architecture

Since our research involves symptom-level analysis of depression, we extend the work done by Milintsevich et al. [56] and define a transformer-based hierarchical model that produces eight regression outputs for the eight symptom scores of PHQ-8. The proposed architecture is based on the model defined by Milintsevich et al. [56], which has been updated to access sentence-level attention and to take advantage of recent sentence representation models. An overview of the model is shown in figure 5.2. In

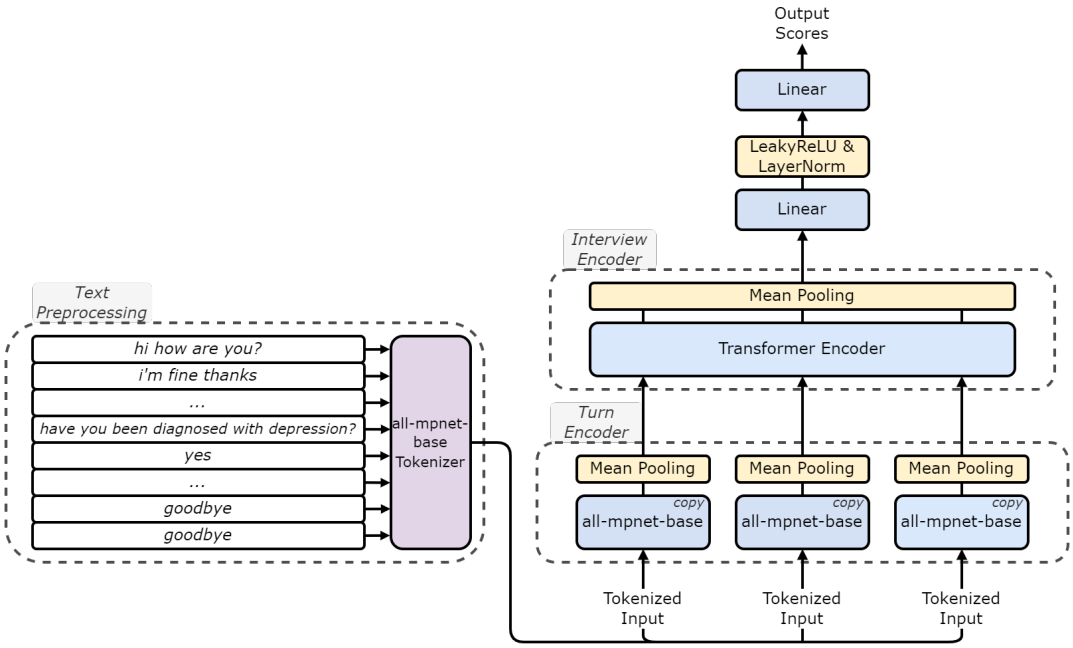


Figure 5.2: Hierarchical neural network architecture for symptom-based predictions.

particular, the architecture has undergone two significant alterations compared to the definition in §3.2 of [56]: (1) the BiLSTM cells are replaced by a transformer-based encoder at the interview level (Interview Encoder), and (2) the pre-trained *Turn Encoder* is based on the all-mpnet-base model[3] in place of *S-RoBERTa*[4], both using a contrastive learning objective [74].

The model consists of two encoders: the *Turn Encoder* ($\boldsymbol{Enc^{turn}}$) that encodes each sentence, and the *Interview Encoder* ($\boldsymbol{Enc^{int}}$) that encodes sentence-level representations into an interview-level embedding. Consider an interview transcript $D = \{t_1, ..., t_{n-1}, t_n\}$ where $t_i = \{w_1^i, ..., w_{m-1}^i, w_m^i\}$ are the dialogue turns and w_j^i is the j*th* token in turn t_i. First Enc^{turn} encodes the token sequence of each turn t_i to generate sentence level embeddings h_i^{turn} (equation 5.1). These sentence embeddings are then processed by Enc^{int} at the next level of hierarchy to produce interview level representation h^{int} (equation 5.2). The interview level embedding is then passed through a feed-forward network that maps it to a prediction vector $m = [m_1, m_2, ..., m_8]$, where each predicted label $m_k \in [0, 3]$ represents a symptom score for the corresponding question in the PHQ-8 questionnaire. The feed-forward classifier contains two linear layers with LeakyReLU activation and a LayerNorm [5]

[3]https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[4]https://huggingface.co/sentence-transformers/all-distilroberta-v1

ELLIE: *how close are you to your family*
PARTICIPANT: *@@ very close @@ even though i don't live with them @@ i try to see them as much as possible @@*
ELLIE: *mhm*
ELLIE: *how do you like your living situation*
PARTICIPANT: *uh it's ok*

Figure 5.3: Example of annotation marking for training Marked-up model.

in-between. The interview encoder, Enc^{int}, contains 4 layers containing 12 attention heads each with an intermediate size of 1536 and a hidden size of 768. This model acts as the base architecture for the different experiments and model configurations explored within our research and is referred to as the *Baseline model* within this chapter.

$$\boldsymbol{h_i^{turn}} = \boldsymbol{Enc^{turn}}(t_i) \ for \ i = 1, ..., |D| \tag{5.1}$$

$$\boldsymbol{h^{int}} = \boldsymbol{Enc^{int}}(\{h_1^{turn}, ..., h_{|D|}^{turn}\}) \tag{5.2}$$

5.2.2 External Knowledge Integration

In our effort to reintroduce domain expertise into depression estimation tasks, we incorporate psychiatrist annotations into the learning process of our neural network model. The proposed research is aligned with the approach taken by Soares et al. [83] and Boualili et al. [14], and introduce special markers into the input text to directly highlight clinical annotations within the transcripts. The underlying idea is that explicitly marking spans in the input text may allow the model to carefully identify the annotations and make a more informed prediction. Consequently, all annotations provided by the psychiatrists are encompassed in between the @@ markers within the transcripts, giving rise to a marked-up corpus (example in figure 5.3). We use the Baseline architecture defined earlier and fine-tune it using the marked-up corpus. Specifically, the pre-trained *all-mpnet-base* model is fine-tuned by unfreezing only the final layer. The resulting model is referred to as the *Marked-up model* within this chapter.

Model	MAE	
	Dev.	Test
SOTA		
ASP MT. DLC+DLR+EIR [68]		3.69
HCAG-T [59]	3.73	-
SGNN [43]	3.76	-
Symptom prediction [56]	3.61	3.78
Dual encoder (warm start) [51]	2.76	3.80
Our Configurations		
Baseline model	4.08	**3.52**
Marked-up model	3.49	3.60

Table 5.2: Comparison of overall model performance against current state-of-the-art results. The results are averaged over 5 random initializations.

5.3 Results and Analysis

Table 5.2 provides overall results for the various model configurations considered in the experiments and puts them into perspective by comparing them against current state-of-the-art results. All these works have used only text modality as input, as is the case in our research. Our models outperform all previous research initiatives, with *Baseline model* providing new state-of-the-art performance on the test set of the DAIC-WOZ on an average taken over 5 runs. It is interesting to notice that the marked-up model does not improve over the baseline model despite containing extra information, although it outperforms all previous research initiatives. This issue is further discussed in detail in §5.5.

Ablation study: Although clinical annotations highlight important information within the transcript, they do not represent exhaustive knowledge sufficient to gauge a patient's mental health. Given the complete set of information required for estimating depression, we seek to understand the role played by our clinical annotations within this set. As such, we conduct an ablation study to analyze the amount of information contained within the annotations by removing parts of interview transcripts and analyzing the change in model performance. Two new input configurations are defined and used with a trained instance of *Baseline model* at the inference stage to generate new predictions over the modified inputs. An instance of *Baseline model* trained on complete interview transcripts is used within the experiments without any further training or fine-tuning to account for modified input configurations. The two input versions in this ablation study are defined as follows:

Ablation configurations	MAE on Test set
Baseline model	**3.52**
$\text{Baseline}_{\text{ann.}}$ inference	4.02
$\text{Baseline}_{\text{non-ann.}}$ inference	3.84

Table 5.3: Ablation study with baseline model for exclusively non-annotated and annotated sentences.

Baseline$_{ann}$ inference: Within this input configuration only question-answer pairs with at least one annotation are retained within the transcripts and the remaining information is discarded.

Baseline$_{non-ann}$ inference: Only question-answer pairs without any annotation are retained within the input transcripts in this input configuration.

Results of the ablation study are shown in table 5.3. As expected, there is a drop in model performance when parts of the interview are removed, thus verifying the importance of both annotated and non-annotated parts of the transcripts toward final predictions. We see a significant drop in performance on removing annotated question-answer pairs from the input transcripts ($Baseline_{non-ann}$ *model*), highlighting the validity of the psychiatrists' annotations. Surprisingly, we also see a drop in performance when only annotated questions-answer pairs are used as inputs ($Baseline_{ann}$ *model*). This behaviour can be attributed to the fact that in this configuration the number of sentences within the interviews is severely reduced and as such the coherence of the discourse is undermined, affecting the performance of the automated models. This shows that although psychiatrists' annotations represent informative parts of the transcript, this knowledge is not disjoint from the remaining discourse.

5.4 Attention and Annotated Spans

Psychiatrist annotations highlight text spans that hold relevance for depression estimation as per clinicians' knowledge and medical guidelines. Given their importance from the medical point of view, we propose to verify whether automated models attend to the same annotated text spans or look for information that complements clinical knowledge. Psychiatrist annotations are analyzed against sentence-level attention scores from the model, sentence being the atomic textual element for this analysis. In particular, we focus on 3 different sentence types: questions (Q), non-annotated turns (N) that contain answers without any annotations, and clinically-

Class	Metric	Q	N	A
Non-depressed	min.	12.84	12.93	13.60
	max.	137.50	136.76	135.35
	med.	42.03	42.10	**42.25**
	avg.	30.85	31.01	**31.25**
Depressed	min.	15.29	15.02	15.37
	max.	103.88	102.83	110.89
	med.	37.96	38.50	**38.82**
	avg.	12.18	12.18	**12.29**

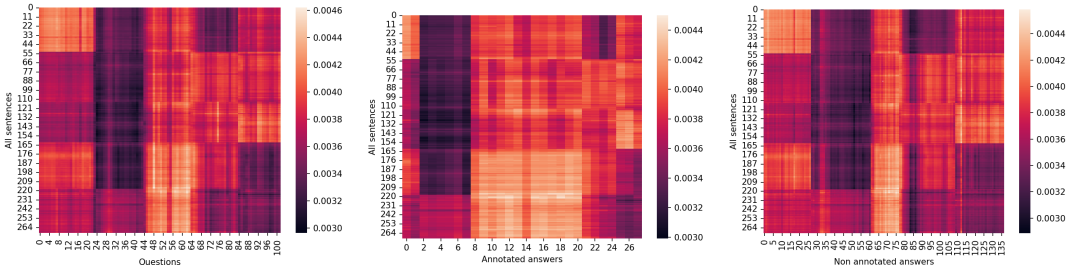
Table 5.4: Sentence-level attention scores calculated over the DAIC-WOZ dataset for **Q**uestions, **N**on-annotated and **A**nnotated turns. Values are with the precision of 10^{-4}. Med. and avg. stand for median and arithmetic mean.

annotated turns (A) that contain patient responses with at least one annotation. Thus, each attention head $H^{s\times s}$ of the interview encoder Enc^{int} is converted into three attention sub-matrices $H^{s\times q}$, $H^{s\times n}$ and $H^{s\times a}$, where s is the number of sentences in a given transcript, q the number of questions, a the number of annotated turns and n the number of non-annotated turns, such that $s = q + n + a$. For each interview, we average the sentence-level attention scores for Q, N, and A sentence types for all attention heads contained in the interview encoder as defined in equation 5.3, where h and l stand for the number of heads and layers respectively.

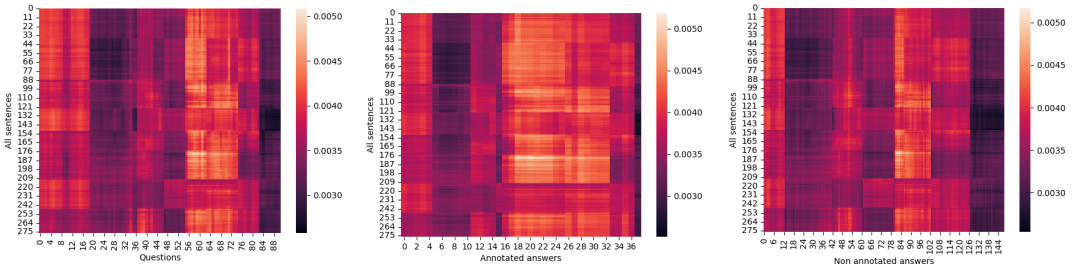
$$\overline{X} = \frac{1}{l.h}\sum_{l,h}\frac{1}{i.j}\sum_{i,j} H_{i,j}^{s\times x}, \forall x \in \{q, n, a\} \tag{5.3}$$

Finally, we average these values over the 189 interviews of the DAIC-WOZ to get the overall picture. Results with the baseline model are given in table 5.4 and show that the transformer-based model focuses more on clinically annotated spans compared to other parts of the transcripts, independently of the patient class. This provides the first evidence that the baseline model targets clinically motivated spans for its decision process without the introduction of any external knowledge or use of specific architectures tuned towards guiding the attention values.

To complement this analysis, figure 5.4 plots the three attention heatmaps $\overline{Q}$, $\overline{A}$ and $\overline{N}$ with brighter regions representing higher attention scores. Plots are provided for a depressed patient as well as a non-depressed patient. This illustration exemplifies overall results and shows that although model attention is distributed over all three categories, clinically-annotated turns receive higher average attention as compared to non-annotated turns and questions. Finally, figure 5.5 illustrates the attention scores in the perspective of the psychiatrists' annotations for the same patients. Following the blue line corresponding to the baseline model, we observe

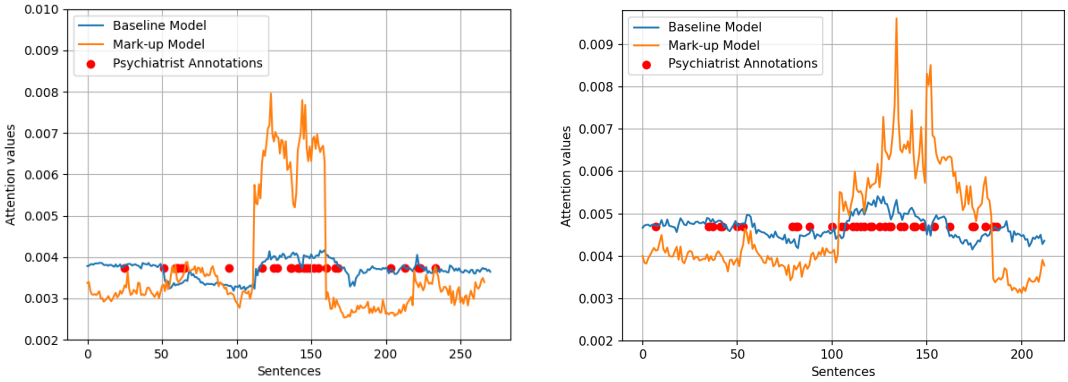


(a) Patient id: 335 (PHQ-8 = 12). Average attention scores (Q, A, N) = (1.994e-05, 2.030e-05, 1.966e-05)



(b) Patient id: 307 (PHQ-8 = 0). Average attention scores (Q, A, N) = (1.955e-05, 2.044e-05, 1.905e-05)

Figure 5.4: Sentence level attention scores from the Baseline model for two different patients.



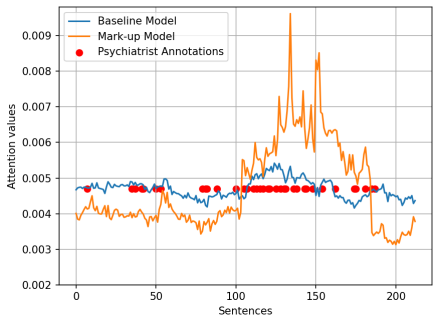
(a) Patient id 335, PHQ-8 score 12

(b) Patient id 307, PHQ-8 score 0

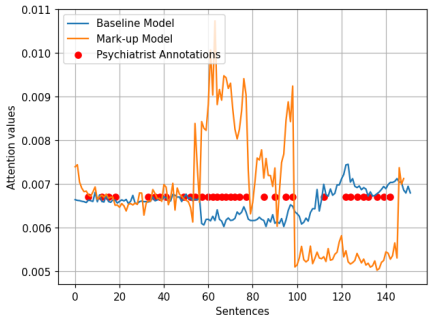
Figure 5.5: Attention scores from baseline and marked-up models plotted against clinical annotations for patients belonging to two classes.

an increase in attention scores in the vicinity of psychiatrist annotations, while the opposite is true in the absence of annotations. These plots represent a general trend observed throughout the dataset with some exceptions.

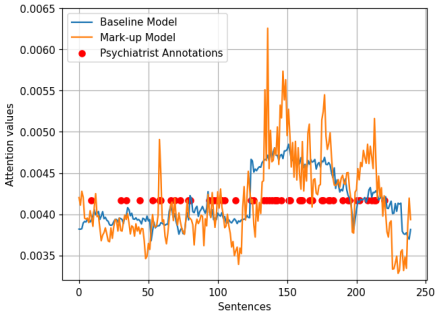
Detailed analysis of the attention scores over the entire dataset revealed a general trend in line with the above findings showing increased attention scores for the Baseline model in the neighborhood of clinically annotated sentences. Figure 5.6



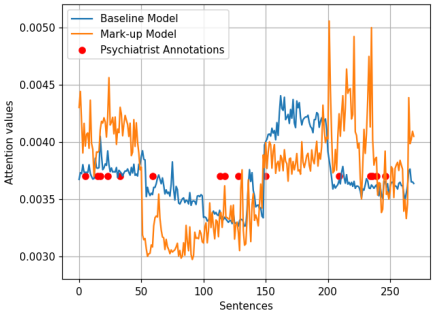
(a) Patient id: 307, PHQ-8 score=0



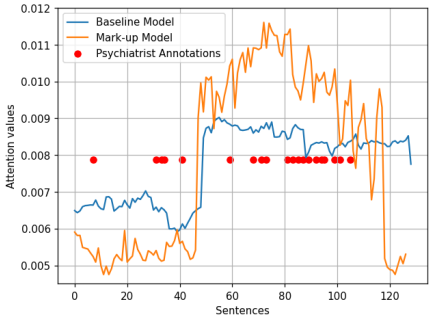
(b) Patient id: 340, PHQ-8 score=1



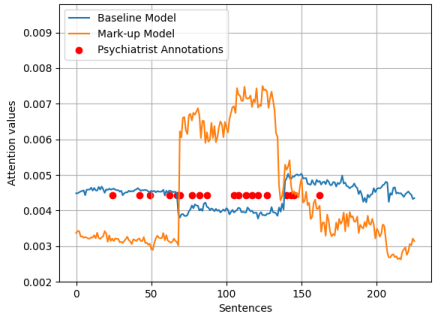
(c) Patient id:336, PHQ-8 score=7



(d) Patient id:410, PHQ-8 score=12



(e) Patient id:362, PHQ-8 score=20



(f) Patient id:311, PHQ-8 score=21

Figure 5.6: Sentence-level attention scores for the Baseline and Marked-up models, with psychiatrist annotations.

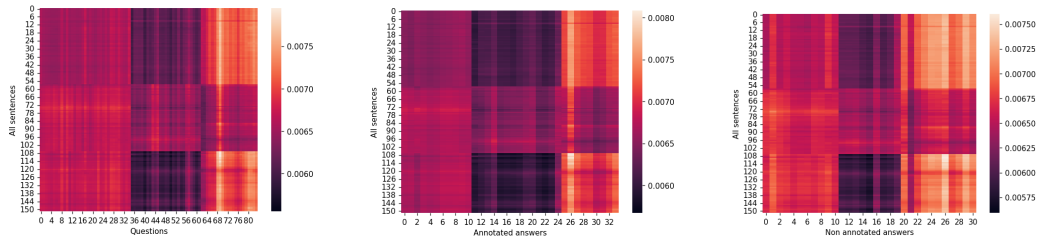
illustrates sentence attention scores from automated models for multiple transcripts within the dataset for the depressed and non-depressed classes. We also include exceptions that deviate from this global pattern (figure 5.6f and figure 5.6b), with low attention values around annotations in some regions. With respect to attention scores for the marked-up model, we steadily notice a peak in values halfway through the transcripts, with lower attention scores before and after the climb. Figure 5.6d represents one rare exception to this rule with high attention at the start and above-average attention scores in regions without annotations. We further plot

the heatmaps in figure 5.7 for some of the same examples, with figure 5.7e verifying exceptional behavior (patient id:311) showing low average attention values towards annotated sentences, although evidencing a highly concentrated attention zone at the end of the plot.

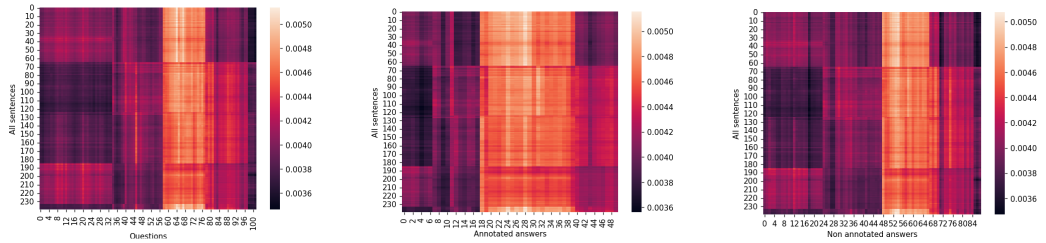
Within the literature, researchers have argued that since the aim is to evaluate patients' mental health, only patients' inputs should be used during the decision-making process [55, 20]. This belief is further reinforced by the fact that our annotators did not annotate any of the questions asked by the psychiatrist within the DAIC-WOZ dataset. It is observed that medical professionals tend to focus only on patient utterances within the interview, while previous research [92] and our experiments (chapter 3) have shown that in the context of neural networks both patient and therapist inputs play an important role. This extends from the fact that models require questions in order to contextualize information within the answers. Figures in table 5.4 support this analysis with questions receiving significant attention scores validating their importance in the decision-making process. Although medical professionals focus more on patient responses, which is in line with neural network behavior (on average answers received more attention compared to questions in table 5.4), they also process questions, albeit subconsciously.

5.5 Performance Analysis and Knowledge Introduction

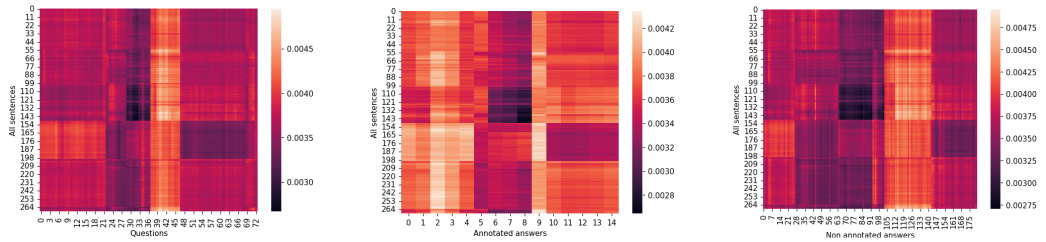
Although the baseline model attends to parts of the interviews that psychiatrists find relevant, we explore the impact of the introduction of clinician expertise directly in the learning process and analyze the performance of the marked-up model. Overall results are illustrated in table 5.5 and do not evidence gains in performance resulting from the knowledge added by the psychiatrist annotations. Indeed, the baseline model outperforms the marked-up model 5 times out of 8 for both the depressed and non-depressed classes. This confirms our previous findings from section §5.4, showing that the baseline architecture already attends to clinically annotated sentences, thus reducing the impact of the marked-up strategy. Figure 5.5 compares both baseline and marked-up models, with plots showing similar behaviors of attending to the annotated sentences although with different amplitude. In particular, the marked-up model tends to pay high attention to the middle of the transcripts thus failing to highlight important information from other regions. This is not the case for the baseline model, which has more evenly distributed attention values, while still being consistent with psychiatrist annotations.



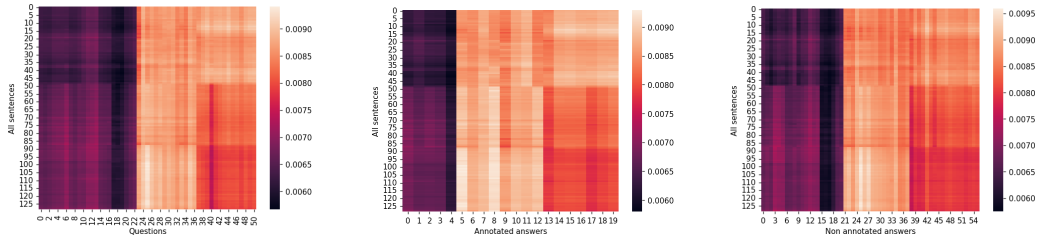
(a) Patient id: 340 (PHQ-8=1). Average attention scores (Q, A, N) = (3.536e-05, 3.528e-05, 3.535e-05)



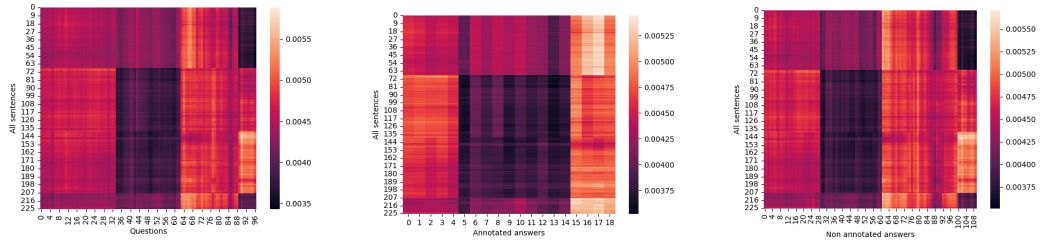
(b) Patient id: 336 (PHQ-8=7). Average attention scores (Q, A, N) = (2.221e-05, 2.320e-05, 2.215e-05)



(c) Patient id: 410 (PHQ-8=12). Average attention scores (Q, A, N) = (1.995e-05, 1.967e-05, 1.991e-05)



(d) Patient id: 362 (PHQ-8=20). Average attention scores (Q, A, N) = (4.085e-05, 4.302e-05, 4.187e-05)



(e) Patient id: 311 (PHQ-8=21). Average attention scores (Q, A, N) = (2.383e-05, 2.320e-05, 2.385e-05)

Figure 5.7: Heatmaps of the sentence-level attention scores for three different examples calculated on Baseline model.

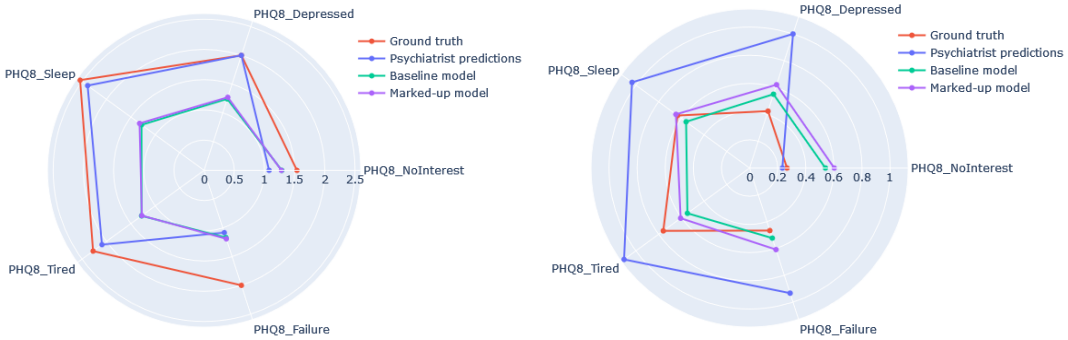
Symptoms	**Psychiatrist Pred.**		**Baseline model**		**Marked-up model**	
	Depr.	Non-Depr.	Depr.	Non-Depr.	Depr.	Non-Depr.
Loss of interest	0.615	0.366	**0.611**	**0.431**	0.699	0.485
Feeling of depression	0.571	0.696	**0.884**	**0.443**	0.939	0.465
Sleeping habits	0.615	0.533	0.761	**0.691**	**0.651**	0.808
Tiredness	0.727	0.689	**0.797**	0.711	0.812	**0.666**
Feeling of failure	1.083	0.800	0.820	**0.543**	**0.786**	0.573
Lack of concentration	-	-	**1.332**	0.521	1.361	**0.475**
Loss of appetite	-	-	**0.932**	0.745	1.037	**0.628**
Lack of movement	-	-	1.008	**0.105**	**0.964**	0.125

Table 5.5: MAE calculated against patient's self-assessment scores by symptoms over the DAIC-WOZ test set. Results are averaged over 5 runs for the automated models. Psychiatrist prediction evidences the difference between the patients' assessments and the psychiatrists' ones.

Symptoms	Depr.		Non-Depr.	
	Over	Under	Over	Under
Psychiatrist Prediction				
Loss of Interest	1	5	3	6
Feeling of depression	3	3	16	2
Sleeping habits	3	3	10	2
Tiredness	2	3	12	5
Feeling of failure	1	8	13	5
Baseline Model				
Loss of Interest	4	9	24	5
Feeling of depression	2	12	24	9
Sleeping habits	1	12	19	10
Tiredness	1	10	14	14
Feeling of failure	1	11	20	9
Marked-up model				
Loss of Interest	4	9	27	3
Feeling of depression	3	11	26	7
Sleeping habits	1	12	19	11
Tiredness	1	10	15	14
Feeling of failure	2	10	23	7

Table 5.6: Number of over- and under-evaluated transcripts in the test set for the baseline model, the marked-up model, and the psychiatrists' scorings.

In order to put prediction results into perspective further, we calculate the Mean Absolute Error (MAE) between the psychiatrist's PHQ-8 scores and patients' self-assessments. Results in table 5.5, calculated by taking self-assessment as ground truth, show that *Psychiatrist Predictions* outperform automated models for most of the symptoms (feeling of failure and loss of interest being exceptions). Further analysis of psychiatrist scoring confirms findings from the medical domain [27], show-



(a) Depressed class (b) Non-Depressed class

Figure 5.8: Radar plots showing symptom-wise average scores for the different automated models, the patient self-assessments, and the psychiatrists' ratings over the test set of the DAIC-WOZ. Note that only 5 symptoms are illustrated, which refer to the ones that psychiatrists could reliably annotate.

ing that clinicians tend to under-evaluate the PHQ-8 scores for the depressed class while over-evaluating those for the non-depressed class. Intriguingly, we observe the same behavior for the automated models as illustrated in table 5.6. The figures indicate that both the *Baseline Model* and the *Marked-up Model* demonstrate behavior similar to that of psychiatrists. This observation further bolsters our claim regarding shared psychological tendencies between our proposed model and psychiatrists. As expected, the number of transcripts misdiagnosed by the automated models far exceeds those misdiagnosed by psychiatrists. This is due to the fact that models generate floating point predictions whereas psychiatrists' predictions are based on a Likert scale ranging from 0 to 3.

To further analyze the behavior of over and under-evaluation, we plot the symptom-wise average scores for the different automated models, the patient self-assessments, and the psychiatrists' ratings in figure 5.8. The illustrations show a high correlation between the results from the two automated models. Both the *Baseline Model* and *Marked-up Model* generate the same average scores for the depressed class while for the non-depressed class, the values are very close. This confirms that introduction of annotations into the learning process through markup strategy does not provide significant performance gain. These plots also support the claims of over and under-evaluation of PHQ-8 scores, and showcase a similar pattern as seen in table 5.6.

5.6 Conclusion

In this part of the thesis, we examine automated depression estimation through the prism of psychiatric expertise and compare the behavior of automated models against clinical annotators. The aim of this work has been twofold: (1) generate clinical annotations for the DAIC-WOZ dataset to encourage multi-disciplinary research in the field, and (2) analyze and compare neural network behavior against that of psychiatrists to validate their reliability as predictive models within the clinical setting.

The limited availability of medical experts has a significant impact on multi-disciplinary research in the field of automated depression estimation, with most researchers treating this as a purely computational task. Through this research we provide clinical annotations of the DAIC-WOZ dataset, encouraging researchers to incorporate domain knowledge within their research. Annotations within our endeavor were carried out with the help of trained medical professionals with expertise in mental health assessment, ensuring their reliability and correctness. Our choice of annotators sets us apart from previous annotation attempts [4] where crowd-workers without medical knowledge were employed as annotators. Our confidence in these annotations is based on the academic background and extensive training of our annotators in the field of mental health assessment. Initial analysis of our annotations highlighted discrepancies between the information contained within the transcripts and the self-assessment PHQ-8 based ground-truth associated with the problem. Our study shows the lack of symptom-level information within interview transcripts, highlighting the fact that for neural network models trained on the DAIC-WOZ dataset, the ground-truth (PHQ-8 scores) is defined outside the context of the model input (interview transcripts). This leads to unreasonable learning goals for models trained only on transcribed interview data.

Our studies show a strong correlation between psychiatrist annotations and the decision-making process of the proposed neural network architecture. Analysis of the sentence level attention scores from the neural network shows that *Baseline model* learns to analyze the interview transcripts in ways similar to a psychiatrist despite the lack of any medical knowledge in the input. A global trend of increased attention values around annotated sentences was observed indicating that our model bases its prediction on the same information sought after by the medical professionals (figure 5.5). However, despite their importance, basing the final prediction solely on psychiatrist annotations can be counter-productive (table 5.3). This indicates the importance of retaining non-annotated sentences within model input, especially to

maintain the coherence of the discourse.

Our annotations further quantify the discrepancy between patients' mental health assessments based on clinical evaluations and self-assessment scores (table 5.5). Moreover, we also establish a strong correlation between the psychological tendencies of the medical professionals and those of our neural network model in the context of predicting the final PHQ-8 scores. Taking patient's self-assessment as the ground-truth, both the neural network and medical professionals share an inclination to underestimate the scores for people suffering from depression, while overestimating the scores for patients without depression (figure 5.6). This further supports the reliability of our model and validates their role as predictive models for clinicians in psychiatry. Finally, the proposed architectures are compared against recent research initiatives, with *Baseline model* providing new state-of-the-art results over the DAIC-WOZ test set (table 5.2).

Chapter 6

Conclusion and Future Work

6.1 Conclusions

Depression is a serious mental disorder affecting millions of people worldwide, encoring huge social and financial losses. The severity and widespread impact of depression has prompted significant research initiatives in the field of automated depression estimation in an attempt to alleviate the pressure on the healthcare systems and extend the reach of such services. Within this context, researchers have exploited different neural network architectures and models to define systems for automated depression estimation. This dissertation discussed the need for automated depression estimation and presented recent initiatives in the field. Despite the extensive literature, some research directions have been overlooked in the context of ADD research based on clinical interviews. Three major research questions were raised within this thesis that are missing from the current literature on automated depression detection. This dissertation emphasizes the importance of answering these research gaps and showcases the advantages of exploring these research directions both in terms of improving the predictive performance of the models and providing explainability and useful insights for medical professionals. The remainder of this chapter concludes the individual research discussed in the preceding chapters and provides an overlook of the possible future research directions.

The first important question raised within this thesis concerns the relevance of discourse structure while processing dyadic patient-therapist interviews. The argument is based on the fact that the dyadic nature of the conversation implies an inherent structure within the discourse. Chapter 3 presents the models and experiments defined in the context of answering this question. Firstly, the relevance of

therapist's questions is established in the context of DAIC-WOZ dataset by verifying the presence of relevant information in the questions asked by the therapist during the interview. This raises further questions regarding the ideal method for combining the two input streams, therapist's questions and patient's answers, with the sequential combination proving to be sub-optimal even compared to using the patient's input alone (chapter 3, table 3.1). Multi-view architectures are defined that divide the input into two views, the patient view and the therapist view, and process them both independently and co-dependently as a possible way of incorporating the discourse structure into the learning process of the models. The proposed architecture not only allows the model to handle the discourse structure and symmetry but also reduces the number of noisy interactions encountered during the training process for more efficient learning, resulting in a significantly higher predictive performance. Furthermore, ablation studies were conducted in order to understand the importance of different interactions within the discourse with *MV-Inter-Att. (Mean)* configuration out-performing the corresponding sequential baseline configuration (*Hierarchical Baseline (Patient+Therapist)*) by 13.84% on macro F1 evaluation metric. Furthermore, *MV-Inter-Att. (Mean)* also outperformed all other multi-view based configurations validating the need to account for both interactions within the two views and the interactions between the corresponding questions and answers. Finally, experiments were conducted with two different text encoding methodologies, hierarchical text encoding, and Sentence Transformer based text encoding, to further strengthen the need for incorporating the discourse structure in the learning process, and the validity of Multi-view architectures.

The second question relates to the role played by the input representation in the learning ability and interpretability of the neural network models. The research in this direction is outlined in chapter 4, which suggests utilizing graph-based representations of the input transcripts. This stems from the necessity of integrating discourse structure into the learning process, and employs Sentence Similarity Graphs (SSG) representations to embed the said structure directly into the input representation. Furthermore, it is argued that graphs are not only a more suitable data structure for representing the non-linear interactions inherent in conversations, but different graph definitions can highlight different relevant aspects of the same input, which might not be possible with sequential encodings. Sentence Similarity Graph and Keyword Correlation Graphs [19] (KCG) structures are explored within this context, respectively encoding local transcript level information and global corpus level knowledge into individual transcript representations to study the impact of different

input representations on the model's performance. Experimental results presented in chapter 4 validate the advantages of using graph-based input representations for improving the predictive abilities of the models, with SSG based input representations out-performing the corresponding sequential and KCG based representations (table 4.2). Furthermore, the multi-view concept was also applied to the graph structures in order to highlight the difference in perspective between the patient and the therapist. Experimental results verify the model-agnostic nature of the multi-view concept with multi-view based configurations out-performing the corresponding baselines across all input representations considered. This further shows that view based division of the input, proposed in the multi-view architecture, not only encodes discourse structure but also highlights the difference in perspective between the patient and the therapist. This is more evident in the case of multi-view based KCG representations where the individual topic models perfectly encode the difference in perspective with the therapist's topic models representing the relevant aspects of the patient's life discussed during the interview, while the patient's emotions and feelings towards these aspects are encoded in the patient topic model. It is shown that visualization of these graph structures can provide further insights into the mental health of the patient. Within this context, KCG visualizations are used to show that therapist behavior also changes in response to the patient's mental status, with visualization of therapist sentence similarity graphs providing a visual representation of therapist behavior during the interview. It is further shown how the key terms within the transcript change with respect to the mental health status of the patient, with terms like "depressed" and "therapy" only appearing in the graphs of depressed patients. Finally, the model performance was also compared against recent state-of-the-art results with the graph based models out-performing all recent initiatives and giving new state-of-the-art performance for the binary classification task within the ADD domain.

The final question pertains to the involvement of medical professionals within the learning process of neural network architectures and their trustworthiness as predictive models within the healthcare setting. Chapter 5 emphasizes the importance of incorporating domain expertise into the learning process of the neural network models, and provides clinical annotations of the DAIC-WOZ dataset in an attempt to encourage multi-disciplinary research into the ADD field. Firstly, clinical annotation of the DAIC-WOZ dataset was carried out to not only highlight medically relevant information within the individual transcripts but also to obtain a medical diagnosis of the participants involved. The clinical annotations open the possibility

to incorporate medical knowledge into the neural network's learning process and account for the lack of domain expertise, not only within the research discussed in this dissertation but clinical interview based automated depression estimation in general. These annotations are also used as context to analogize the predictive tendencies of neural network models and mental health professionals. Initial analysis of the annotations confirm that medical professionals tend to focus more on patient's responses rather than their questions (which they likely process sub-consciously) with almost all annotations appearing in patient responses, while the AI models consider both patient and therapist inputs in their predictive process. Furthermore, they also point towards a lack of information within the transcripts for reliable evaluation of all PHQ-8 indicators with most transcripts lacking information on at least 3 out of 8 symptoms (refer figure 5.1).

Sentence level attention analysis of the proposed *Baseline Model* (model trained without clinical annotations) reveals a strong correlation between psychiatric annotations and the decision-making process of the neural network architecture. Notably, there is an observed trend of increased attention scores corresponding to an increased number of annotations in the region (examples in figure 5.5). Although network attention is distributed over a wider region in contrast to clinical annotations that highlight specific sentences, the general trend supports the argument that both automated models and medical professionals attend to the same information in the transcript, indicating similarities in their decision-making process.

Psychiatric evaluations were also compared against the patient's self-assessment scores, confirming the well-documented discrepancy between medical evaluation and self-assessment of mental health. Assuming self-assessment to be the ground truth, which is the standard practice in the field, symptom level predictions of the *Baseline Model* (defined in chapter 5) follow the same pattern as medical evaluations. In general, the *Baseline model* over-evaluated the non-depressed individuals while under-evaluating the scores for depressed patients, a well-documented behavior of medical professionals which is also observed in the clinical annotations collected as part of this work. Although this work does not provide any explanation of the decision-making process of the proposed neural network models, the overall analysis shows shared psychological tendencies between medical professionals, and the neural network, supporting their use as reliable and trustworthy predictive tools within the healthcare setting. The models proposed in chapter 5 were also compared against relevant initiatives within recent literature with the proposed *Baseline model* providing new state-of-the-art results over the test set of DAIC-WOZ dataset.

6.2 Future Work

Although the models and architectures proposed in this dissertation meet the current expectations and provide improvements in predictive performance over the recent initiatives in the field, they represent initial steps towards the proposed research directions with a great deal of unexplored possibilities. Research in two major directions is currently underway with plans to explore other possibilities in the near future.

The first set of experiments are focused on the multi-view based KCG representations of input transcripts. Efforts are being made to exploit linguistic knowledge to further fine-tune the KCG structure for a better representation of individual perspectives of the two interlocutors, the patient and the therapist, within clinical interviews. The current research revolves around studying the impact of Part Of Speech (POS) tags on the training of topic models within the KCG definition. The aim is to learn more customized view representations, eventually leading to more complex graph definitions allowing a more detailed study of interactions between the individual perspectives, and providing valuable insights and clues not only in the context of the model's predictive process but also the patient's mental health. Working in this direction, experiments are being run that combine SSG and KCG structures to define a multi-level graph representation incorporating both local and corpus level knowledge. Furthermore, we plan to expand the multi-view concept and define the two views as different representations of the input transcript in place of agent-based splitting of the patient-therapist interview. We plan to use SSG and KCG structures to experiment with this updated definition of multi-view architectures.

The second research direction currently being explored is the incorporation of medical knowledge into the learning process of proposed neural network models. Although the initial experiments in this direction did not evidence promising results (table 5.5), different ways of incorporating external knowledge into the model are being explored. In particular, the current experiments are based on the findings of Deshpande et al. [23] and explore the use of clinical annotations within the guided attention mechanism for improved results.

Acronyms

ADD Automated Depression Detection.

BERT Bidirectional Encoder Representations from Transformers [26].

DAIC-WOZ Distress Analysis Interview Corpus - Wizard of Oz [38].

ELMo Embeddings from Language Models [65].

GAT Graph Attention Network [87].

GCN Graph Convolution Network [47].

GloVe Global Vectors [64].

GNN Graph Neural Network.

GPC General Psychotherapy Corpus.

GRU Gated Recurrent Unit.

HCAG Hierarchical Context-Aware Model [59].

KCG Keyword Correlation Graphs [19].

LLM Large Language Models.

MDD Major Depressive Disorder.

MLP Multi Layer Perceptron.

NLP Natural Language Processing.

NMF Non-negative Matrix Factorization [33].

PHQ-8 Patient Health Questionnaire-8.

PTSD Post-Traumatic Stress Disorder.

QA-pair Question-Answer Pair.

RNN Recurrent Neural Network.

SGNN Schema-Based Graph Neural Network.

SSG Sentence Similarity Graphs.

Bibliography

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Usman Ahmed, Jerry Chun-Wei Lin, and Gautam Srivastava. Graph attention network for text classification and detection of mental disorder. *Association of Computing Machinery (ACM) Transactions on the Web*, 2023.

[3] Laura Andrade, Jorge J Caraveo-Anduaga, Patricia Berglund, Rob V Bijl, Ron De Graaf, Wilma Vollebergh, Eva Dragomirecka, Robert Kohn, Martin Keller, Ronald C Kessler, et al. The epidemiology of major depressive episodes: results from the international consortium of psychiatric epidemiology (icpe) surveys. 2003.

[4] Alina Arseniev-Koehler, Sharon Mozgai, and Stefan Scherer. What type of happiness are you looking for? - a closer look at detecting mental health from language. In *5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (CLPSYCH) associated to 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018.

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. In *arXiv preprint arXiv:1607.06450*, 2016.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR*, 2015.

[7] Andrew Bailey and Mark D. Plumbley. Gender bias in depression detection

using audio features. In *29th European Signal Processing Conference (EUSIPCO)*, 2021.

[8] Jasmijn Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. Graph convolutional encoders for syntax-aware neural machine translation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1957–1967, 2017.

[9] Aaron T Beck, Robert A Steer, Gregory K Brown, et al. Beck depression inventory. *Psychological Corporation San Antonio, TX*, 1996.

[10] Daniel Beck, Gholamreza Haffari, and Trevor Cohn. Graph-to-sequence learning using gated graph neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 273–283, 2018.

[11] Jared D Bernard, Jenna L Baddeley, Benjamin F Rodriguez, and Philip A Burke. Depression, language, and affect: an examination of the influence of baseline depression and affect induction on language. *Journal of Language and Social Psychology*, 35(3):317–326, 2016.

[12] HUANG Binxuan and KM CARLEY. Syntax-aware aspect level sentiment classification with graph attention networks. In *In Proceedings of Empirical Methods in Natural Language Processing (EMNLP) and the 9th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 5469–5477, 2019.

[13] Keivan Borna and Reza Ghanbari. Hierarchical lstm network for text classification. *SN Applied Sciences*, 1(9):1–4, 2019.

[14] Lila Boualili, José G. Moreno, and Mohand Boughanem. Markedbert: Integrating traditional IR cues in pre-trained language models for passage retrieval. In *43rd International ACM SIGIR conference on research and development in Information Retrieval (SIGIR)*, pages 1977–1980, 2020.

[15] Scott R Braithwaite, Christophe Giraud-Carrier, Josh West, Michael D Barnes, and Carl Lee Hanson. Validating machine learning algorithms for twitter data against established measures of suicidality. *Journal of Medical Internet Research (JMIR) mental health*, 3(2):e4822, 2016.

[16] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

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

[18] Stevie Chancellor and Munmun De Choudhury. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43, 2020.

[19] Billy Chiu, Sunil Kumar Sahu, Derek Thomas, Neha Sengupta, and Mohammady Mahdy. Autoencoding keyword correlation graph for document clustering. In *Proceedings of the 58th annual meeting of the association for computational linguistics (ACL)*, pages 3974–3981, 2020.

[20] Zhijun Dai, Heng Zhou, Qingfang Ba, Yang Zhou, Lifeng Wang, and Guochen Li. Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis. *Journal of Affective Disorders*, 295:1040–1048, 2021.

[21] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual Association of Computational Machinery (ACM) web science Conference*, pages 47–56, 2013.

[22] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the international Association for the Advancement of Artificial Intelligence (AAAI) conference on web and social media*, pages 128–137, 2013.

[23] Ameet Deshpande and Karthik Narasimhan. Guiding attention for self-supervised learning with transformers. In *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4676–4686. Association for Computational Linguistics, 2020.

[24] Roger Detels and Chorh Chuan Tan. The scope and concerns of public health. *Oxford Textbook of Global Public Health*, pages 3–18, 2009.

[25] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *International Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, pages 1061–1068, 2014.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[27] Marc Domken, Jan Scott, and Peter Kelly. What factors predict discrepancies between self and observer ratings of depression? *Journal of Affective Disorders*, 1994.

[28] Vijay Prakash Dwivedi, Chaitanya K Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. *Journal of Machine Learning Research*, 24(43):1–48, 2023.

[29] Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208, 2018.

[30] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.

[31] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[32] Patrick Fernandes, Miltiadis Allamanis, and Marc Brockschmidt. Structured neural summarization. In *International Conference on Learning Representations (ICLR)*, 2019.

[33] Cédric Févotte and Jérôme Idier. Algorithms for Nonnegative Matrix Factorization with the β-Divergence. *Neural Computation*, 23(9):2421–2456, 2011.

[34] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *Workshop on Representation Learning on Graphs and Manifolds associated with International Conference on Learning Representation (ICLR)*, 2019.

[35] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

[36] Centers for Disease Control and Prevention (CDC). Behavioural risk factor surveillance system survey data. In *Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention*, 2008,2011,2012.

[37] George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta. Characterisation of mental health conditions in social media using informed deep learning. *Scientific reports*, 7(1):45141, 2017.

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

[39] Tao Gui, Yicheng Zou, Qi Zhang, Minlong Peng, Jinlan Fu, Zhongyu Wei, and Xuan-Jing Huang. A lexicon-based graph neural network for chinese ner. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1040–1050, 2019.

[40] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 2017.

[41] Léo Hemamou, Ghazi Felhi, Vincent Vandenbussche, Jean-Claude Martin, and Chloé Clavel. Hirenet: A hierarchical attention model for the automatic analysis of asynchronous video job interviews. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 573–581, 2019.

[42] Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.

[43] Simin Hong, Anthony Cohn, and David Crossland Hogg. Using graph representation learning with schema encoders to measure the severity of depressive symptoms. In *International Conference on Learning Representations*, 2022.

[44] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556. Association for Computational Linguistics, 2019.

[45] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. MentalBERT: Publicly available pretrained language models for mental healthcare. In *13th Language Resources and Evaluation Conference (LREC)*, pages 7184–7190, 2022.

[46] Mohsinul Kabir, Tasnim Ahmed, Md Bakhtiar Hasan, Md Tahmid Rahman Laskar, Tarun Kumar Joarder, Hasan Mahmud, and Kamrul Hasan. Deptweet: A typology for social media texts to detect depression severities. *Computers in Human Behavior*, 139:107503, 2023.

[47] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[48] Jun Kong, Jin Wang, and Xuejie Zhang. Hierarchical bert with an adaptive fine-tuning strategy for document classification. *Knowledge-Based Systems*, 238:107872, 2022.

[49] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613, 2001.

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

[51] Clinton Lau, Xiaodan Zhu, and Wai-Yip Chan. Automatic depression severity assessment with deep learning using parameter-efficient tuning. *Frontiers in Psychiatry*, 14:1160291, 2023.

[52] David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

[53] Ying Luo and Hai Zhao. Bipartite flat-graph network for nested named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6408–6418. Association for Computational Linguistics (ACL), 2020.

[54] Jason B Luoma, Catherine E Martin, and Jane L Pearson. Contact with mental health and primary care providers before suicide: a review of the evidence. *American Journal of Psychiatry*, 159(6):909–916, 2002.

[55] Adria Mallol-Ragolta, Ziping Zhao, Lukas Stappen, Nicholas Cummins, and Björn W. Schuller. A hierarchical attention network-based approach for depression detection from transcribed clinical interviews. In *Interspeech (INTERSPEECH)*, pages 221–225, 2019.

[56] Kirill Milintsevich, Kairit Sirts, and Gaël Dias. Towards automatic text-based estimation of depression through symptom prediction. *Brain Informatics*, 10(1):1–14, 2023.

[57] Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. Towards transparent and explainable attention models. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4206–4216, July 2020.

[58] Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*, 2010.

[59] Meng Niu, Kai Chen, Qingcai Chen, and Lufeng Yang. Hcag: A hierarchical context-aware graph attention model for depression detection. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4235–4239. IEEE, 2021.

[60] Xiaolei Niu and Yuexian Hou. Hierarchical attention blstm for modeling sentences and documents. In *24th International Conference on Neural Information Processing (ICONIP)*, pages 167–177. Springer, 2017.

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

[62] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[63] Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. Semantic graphs for generating deep questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1463–1475, 2020.

[64] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[65] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.

[66] Inna Pirina and Çağrı Çöltekin. Identifying depression on Reddit: The effect of training data. In *Proceedings of 3rd Social Media Mining for Health Applications Workshop & Shared Task SMM4H associated with Empirical Methods in Natural Language Processing (EMNLP)*, pages 9–12. Association for Computational Linguistics, 2018.

[67] Meng Qu, Tianyu Gao, Louis-Pascal Xhonneux, and Jian Tang. Few-shot relation extraction via bayesian meta-learning on relation graphs. In *International conference on machine learning*, pages 7867–7876. PMLR, 2020.

[68] Syed Arbaaz Qureshi, Gaël Dias, Mohammed Hasanuzzaman, and Sriparna Saha. Improving depression level estimation by concurrently learning emotion intensity. *IEEE Computational Intelligence Magazine*, 15(3):47–59, 2020.

[69] Syed Arbaaz Qureshi, Sriparna Saha, Mohammed Hasanuzzaman, and Gaël Dias. Multitask representation learning for multimodal estimation of depression level. *IEEE Intelligent Systems*, 34(5):45–52, 2019.

[70] Lenore Sawyer Radloff. The ces-d scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1:385–401, 1977.

[71] Eva-Maria Rathner, Julia Djamali, Yannik Terhorst, Björn W. Schuller, Nicholas Cummins, Gudrun Salamon, Christina Hunger-Schoppe, and Harald Baumeister. How did you like 2017? detection of language markers of depression and narcissism in personal narratives. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association (ISCA)*, pages 3388–3392, 2018.

[72] Eva-Maria Rathner, Julia Djamali, Yannik Terhorst, Björn W. Schuller, Nicholas Cummins, Gudrun Salamon, Christina Hunger-Schoppe, and Harald Baumeister. How did you like 2017? detection of language markers of depression and narcissism in personal narratives. In B. Yegnanarayana, editor, *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association*, pages 3388–3392. ISCA, 2018.

[73] Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg. Multi-level attention network using text, audio and video for depression prediction. In *9th International on Audio/Visual Emotion Challenge and Workshop (AVEC)*, page 81–88, 2019.

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

[75] Benjamin J Ricard, Lisa A Marsch, Benjamin Crosier, and Saeed Hassanpour. Exploring the utility of community-generated social media content for detecting depression: an analytical study on instagram. *Journal of medical Internet research*, 20(12):e11817, 2018.

[76] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133, 2004.

[77] Devendra Singh Sachan, Lingfei Wu, Mrinmaya Sachan, and William Hamilton. Stronger transformers for neural multi-hop question generation. *arXiv preprint arXiv:2010.11374*, 2020.

[78] Farig Sadeque, Dongfang Xu, and Steven Bethard. Measuring the latency of depression detection in social media. In *Proceedings of the Eleventh ACM*

International Conference on Web Search and Data Mining, page 495–503. Association for Computing Machinery, 2018.

[79] Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. Inter-sentence relation extraction with document-level graph convolutional neural network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4309–4316. Association for Computational Linguistics, 2019.

[80] Elvis Saravia, Chun-Hao Chang, Renaud Jollet De Lorenzo, and Yi-Shin Chen. Midas: Mental illness detection and analysis via social media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1418–1421. IEEE, 2016.

[81] Judy Hanwen Shen and Frank Rudzicz. Detecting anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 58–65. Association for Computational Linguistics, 2017.

[82] Cheolmin Shin, Seung-Hoon Lee, Kyu-Man Han, Ho-Kyoung Yoon, and Changsu Han. Comparison of the usefulness of the phq-8 and phq-9 for screening for major depressive disorder: Analysis of psychiatric outpatient data. *Psychiatry Investigation*, 16(4):300–305, 2019.

[83] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *57th Conference of the Association for Computational Linguistics (ACL)*, pages 2895–2905, 2019.

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

[85] Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432, 2015.

[86] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[87] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

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

[89] Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. Detecting and characterizing eating-disorder communities on social media. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, page 91–100. Association for Computing Machinery, 2017.

[90] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, 2019. Association for Computational Linguistics.

[91] Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. A multilevel predictive model for detecting social network users with depression. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 130–135. IEEE, 2018.

[92] Danai Xezonaki, Georgios Paraskevopoulos, Alexandros Potamianos, and Shrikanth Narayanan. Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews. In *Interspeech (INTERSPEECH)*, pages 4556–4560, 2020.

[93] Yujing Xia, Lin Liu, Tao Dong, Juan Chen, Yu Cheng, and Lin Tang. A depression detection model based on multimodal graph neural network. *Multimedia Tools and Applications*, pages 1–17, 2024.

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

[95] Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. In *28th International Conference on Computational Linguistics (COLING)*, pages 696–709, 2020.

[96] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. Towards interpretable mental health analysis with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[97] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

[98] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377, 2019.

[99] Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. Graph-based neural multi-document summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 452–462. Association for Computational Linguistics, 2017.

[100] Andrew Yates, Arman Cohan, and Nazli Goharian. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

[101] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019.

[102] Jianming Zheng, Yupu Guo, Chong Feng, and Honghui Chen. A hierarchical neural-network-based document representation approach for text classification. *Mathematical Problems in Engineering*, 2018, 2018.

Automated Depression Level Estimation: A Study on Discourse Structure, Input Representation and Clinical Reliability

Keywords: Automated depression estimation, Discourse structure, Graph-based input representation, Multi-view architectures

Résumé: La recherche discutée dans cette thèse vise à répondre à trois questions majeures dans le domaine de l'estimation automatisée de la dépression ; (1) le rôle de la structure du discours dans la compréhension de la santé mentale, (2) la pertinence de la représentation de l'entrée, et (3) l'importance de la connaissance médicale dans l'analyse automatisée de la santé mentale. Ceci constitue la base de ma recherche sur les architectures multi-vues pour encoder la structure du discours des entretiens dyadiques patient-thérapeute. Les représentations graphiques des transcriptions sont également explorées pour modéliser les interactions non linéaires au sein des conversations dyadiques et la génération d'idées. Nous intégrons en outre le concept de vues multiples dans ces structures graphiques, l'établissant comme une méthodologie agnostique du modèle. Nous examinons également le comportement de notre modèle dans le cadre des annotations cliniques afin d'établir des parallèles avec leurs tendances prédictives, améliorant ainsi la fiabilité des modèles de réseaux neuronaux en tant qu'outils prédictifs dans les systèmes de soins de santé.

Abstract: The research discussed within this thesis aims to answer three major questions in the domain of automated depression estimation; (1) the role of discourse structure in mental health understanding, (2) the relevance of input representation, and (3) the importance of medical knowledge in automated mental health analysis. This forms the basis for my research on multi-view architectures for encoding the discourse structure of dyadic patient-therapist interviews. Graph-based transcript representations are also explored for modeling non-linear interactions within dyadic conversations along with insight generation. We further incorporate the multi-view concept within these graph structures, establishing it as a model-agnostic methodology. We also examine our model's behavior within the framework of clinical annotations to draw parallels with their predictive tendencies, thus enhancing the reliability of neural network models as predictive tools in healthcare systems.