AFFECT AND PERSONALITY AWARE ANALYSIS OF SPEECH CONTENT FOR AUTOMATIC ESTIMATION OF DEPRESSION SEVERITY

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

By

Kaan Gönç

September 2023

Affect and Personality Aware Analysis of Speech Content for Automatic Estimation of Depression Severity
By Kaan Gönç
September 2023

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Hamdi Dibeklioğlu(Advisor)

Shervin Rahimzadeh Arashloo

Pınar Duygulu Şahin

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

AFFECT AND PERSONALITY AWARE ANALYSIS OF SPEECH CONTENT FOR AUTOMATIC ESTIMATION OF DEPRESSION SEVERITY

Kaan Gönç
M.S. in Computer Engineering
Advisor: Hamdi Dibeklioğlu
September 2023

The detection of depression has gained a significant amount of scientific attention for its potential in early diagnosis and intervention. In light of this, we propose a novel approach that places exclusive emphasis on textual features for depression severity estimation. The proposed method seamlessly integrates affect (emotion and sentiment), and personality features as distinct yet interconnected modalities within a transformer-based architecture. Our key contribution lies in a masked multimodal joint cross-attention fusion, which adeptly combines the information gleaned from these different text modalities. This fusion approach empowers the model not only to discern subtle contextual cues within textual data but also to comprehend intricate interdependencies between the modalities. A comprehensive experimental evaluation is undertaken to meticulously assess the individual components comprising the proposed architecture, as well as extraneous ones that are not inherent to it. The evaluation additionally includes the assessments conducted in a unimodal setting where the impact of each modality is examined individually. The findings derived from these experiments substantiate the self-contained efficacy of our architecture. Furthermore, we explore the significance of individual sentences within speech content, offering valuable insights into the contribution of specific textual cues and we perform a segmented evaluation of the proposed method for different ranges of depression severity. Finally, we compare our method with existing state-of-the-art studies utilizing different combinations of auditory, visual, and textual features. The final results demonstrate that our method achieves promising results in depression severity estimation, outperforming the other methods.

Keywords: depression severity estimation, deep learning, natural language processing, multimodal fusion.

ÖZET

DEPRESYON ŞİDDETİNİN OTOMATİK TAHMİNİ İÇİN KONUŞMA İÇERİĞİNİN DUYGULANIMA VE KİŞİLİĞE BAĞLI ANALİZİ

Kaan Gönç
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Hamdi Dibeklioğlu
Eylül 2023

Depresyon tespiti, erken teşhis ve müdahale potansiyeli dolayısıyla bilimsel açıdan önemli ölçüde ilgi çekmektedir. Bu sebeple, bu tezde depresyon şiddetinin tahmini için yalnızca metin özniteliklerine bağlı kalan yeni bir yaklaşım önerilmektedir. Önerilen bu yaklaşım, dönüştürücü tabanlı bir yapı içerisinde duygulanım (duygu ve his) ve kişilik özniteliklerini farklı ancak birbirine bağlı kipler hâlinde entegre etmektedir. Bu tezin ana katkısı, farklı metin kiplerinden elde edilen bilgileri birleştirmeyi sağlayan maskeli ve çok kipli ortak çapraz dikkat füzyon yaklaşımıdır. Bu füzyon yaklaşımı, modelin sadece metin verileri içindeki gizli bağlamsal ipuçlarını ayırt etmesine değil, aynı zamanda modaliteler arasındaki karmaşık bağımlılıkları da anlamasına olanak tanımaktadır. Önerilen mimaride var olan bileşenler ile var olmayan bileşenler ayrıntılı olarak incelenmek üzere kapsamlı deneysel değerlendirmelere tabi tutulmaktadır. Bu değerlendirmeler, her kipin ayrı ayrı incelendiği tek kipli bir ortamda gerçekleştirilen deneyleri de içerir. Değerlendirmelerden elde edilen bulgular, önerilen mimarinin kendi kendine yeterli etkinliğini doğrulamaktadır. Bunlara ek olarak, bu tezde konuşma içeriği içindeki cümlelerin önemini inceleyerek belirli metin ipuçlarının katkısına dair değerli bilgiler sunulmaktadır. Aynı zamanda, önerilen yöntemin farklı depresyon şiddeti aralıkları için değerlendirmeleri yer almaktadır. Son olarak, önerilen yöntem farklı işitsel, görsel ve metinsel özellik kombinasyonları kullananan mevcut en ileri düzey çalışmalar ile karşılaştırılmaktadır. Sonuçlar, önerilen yöntemin depresyon şiddeti tahmininde umut verici sonuçlar elde ettiğini ve diğer yöntemleri geride bıraktığını göstermektedir.

Anahtar sözcükler: depresyon şiddeti tahmini, derin öğrenme, doğal dil işleme, çok kipli füzyon.

Acknowledgement

I would like to express my heartfelt gratitude to my supervisor, Asst. Prof. Dr. Hamdi Dibeklioğlu, for his unwavering support and guidance throughout my academic journey. His expertise in academic research and writing has been instrumental in shaping this thesis. Mr. Dibeklioğlu's incredible patience and consistent encouragement have not only facilitated my academic growth but have also been a source of invaluable mental support. Moreover, he has become an inspiration and role model to me, demonstrating the highest standards of scholarship and dedication. I am truly fortunate to have had the privilege of working with such a dedicated and knowledgeable mentor, and I am deeply appreciative of his contributions to my academic development.

I would like to extend my heartfelt appreciation to my loving family, who have been my unwavering pillars of support and inspiration throughout my academic journey. My sister, Aslı, has been a constant source of encouragement, sharing in both my joys and challenges, and her belief in me has been a driving force behind my success. My mother, Nazlı, and my father, Uğur, have nurtured my growth with unwavering dedication and love, instilling in me the values of diligence and perseverance. Their sacrifices and endless encouragement have been the foundation upon which I have built my academic achievements. Their success throughout their lives has been a tremendous source of inspiration for me. Witnessing their determination and achievements in their respective fields has motivated me to push the boundaries of my own potential and strive for excellence. Together, they have provided the abiding belief and inspiration that have empowered me to reach this point in my academic pursuits. Furthermore, Olaf, our cherished canine companion, holds a special place in our family, and his unwavering loyalty and companionship have provided solace and joy during the demanding times of research and writing.

I would like to thank my dear friends and colleagues, Onat and Baturay, who have played an indispensable role in my academic journey. Their friendship and camaraderie have been a constant source of inspiration and support. As both friends and colleagues, they have not only shared in the joys and challenges of academic life but have also motivated me with their relentless pursuit of success. Their boundless enthusiasm and the way they consistently push themselves to achieve their goals have been a beacon of motivation for me, spurring me onward during the most challenging times. I am immensely fortunate to have friends like Onat and Baturay, and their friendship has been a vital component of my academic success.

I want to express my deepest appreciation to my special friends, İdil, Esin, Hande, and Eda who have been unwavering sources of mental support, motivation, and understanding throughout my academic journey. Their friendship has been a priceless gift, and their constant encouragement has been instrumental in keeping me focused and driven in my studies. Whenever I faced academic challenges, they were always there with a helping hand and comforting words, reminding me that I was never alone in this journey. İdil, Esin, Hande, and Eda have not only been friends but also pillars of strength, lifting me up during times of self-doubt and pushing me to reach for my goals. Their friendship has been an essential part of my academic success, and I am truly grateful for their unwavering presence in my life.

I would like to thank Turkish Academy of Sciences (TÜBA) for their partial support in our research through the GEBIP award program. Lastly, I want to express my gratitude to The Scientific and Technological Research Council of Türkiye (TÜBİTAK) for their financial support under grant number 122E134.

Contents

1 Introduction **1**

2 Related Work **4**

3 Method **10**

3.1 Auxiliary Networks 11

3.1.1 Abstract Network 11

3.1.2 Emotion & Sentiment Networks 11

3.1.3 Personality Network 12

3.2 Dimension Standardization 14

3.3 Addition of Regression Token 15

3.4 Transformer Block 16

3.5 Masked Multimodal Joint Cross-Attention Fusion (MMJCA-Fusion) 18

3.6 Feed-Forward Network & Regression 20

4 Experimental Setup 21

4.1 Datasets 22

4.1.1 Dataset Used for the Assessment of the Proposed Depression Severity Estimation Network 22

4.1.2 Datasets Used for the Fine-tuned Auxiliary Networks . . . 24

4.2 Training Configuration 26

4.3 Hyperparameter Tuning 26

4.4 Model Selection 27

4.5 Evaluation Metrics 27

5 Experiments & Results 29

5.1 Evaluations in the Unimodal Setting 30

5.1.1 Assessment of Auxiliary Networks Individually 30

5.1.2 Assessment of Temporal Modeling in the Unimodal Setting 32

5.2 Evaluations in the Multimodal Setting 39

5.2.1 Assessment of Various Combinations of Auxiliary Networks 39

5.2.2 Assessment of Temporal Modeling in the Multimodal Setting 40

5.2.3 Assessment of Regression Token Approach 43

5.2.4 Assessment of Fusion Approach 44

5.3 Segmented Evaluation of The Method Across Different Ranges of True Values 46

5.4 Assessment of Attention Weights of the MMJCA-Fusion Module on the Test Set 49

5.4.1 Discussion on the High True PHQ-8 Score with Accurate Prediction Scenario 53

5.4.2 Discussion on the High True PHQ-8 Score with Inaccurate Prediction Scenario 53

5.4.3 Discussion on the Low True PHQ-8 Score with Accurate Prediction Scenario 54

5.4.4 Discussion on the Low True PHQ-8 Score with Inaccurate Prediction Scenario 55

5.5 Comparison to Other Methods 56

6 Conclusion **58**

List of Figures

3.1 High-level architecture of the method 11

3.2 Masked Multimodal Joint Cross-Attention Fusion 19

4.1 Data distributions across different intervals of PHQ-8 Score. . . . 23

5.1 High-level architecture of the method modified for the unimodal setting 30

5.2 High-level architecture of the method modified for the unimodal setting including temporal modeling 32

5.3 High-level architecture of the method modified with the addition of temporal modeling 41

5.4 RMSE and MAE values across different groups in the validation set, each encompassing a specific range of PHQ-8 scores. 47

5.5 RMSE and MAE values across different groups in the test set, each encompassing a specific range of PHQ-8 scores. 48

List of Tables

4.1 Hyperparameters 27

5.1 Results for individual auxiliary networks 31

5.2 Results of temporal modeling in the unimodal setting 35

5.3 Results of bidirectional recurrent layers in the unimodal setting . 37

5.4 Results of excluding the transformer block in the unimodal setting 38

5.5 Results for different combinations of auxiliary networks 40

5.6 Results of different temporal modeling approaches 42

5.7 Results of different pooling methods 43

5.8 Results of different fusion approaches 45

5.9 Average attention weights for the high true PHQ-8 score with accurate prediction scenario 51

5.10 Average attention weights for the high true PHQ-8 score with inaccurate prediction scenario 51

5.11 Average attention weights for the low true PHQ-8 score with accurate prediction scenario 52

5.12 Average attention weights for the low true PHQ-8 score with inaccurate prediction scenario 52

5.13 Results of comparison to other methods 57

Chapter 1

Introduction

Depression is a common mental disorder affecting millions worldwide. It can cause persistent feelings of sadness, hopelessness, and loss of interest in daily activities [1]. Depression can also impair physical health, social functioning, and quality of life. Therefore, it is important to detect and treat depression early and effectively. However, detecting depression can be challenging, as many people may not seek professional help or disclose their symptoms due to stigma, lack of awareness, or other barriers. Moreover, people do not give the same response or express the same emotions in the face of situations they experience. Due to this, experts have to make subjective evaluations specific to the patient throughout the treatment. Therefore, developing auxiliary tools that can make objective and generalizable determinations is preferred for obtaining accurate results. Following this, we propose a novel multimodal approach to automatically predict the Personal Health Questionnaire (PHQ-8) score [2] from textual transcripts of clinical interviews.

Recent advances in the literature enabled the development of automatic methods for depression detection based on auditory and visual cues [3–7]. These methods analyze the speech and facial expressions of the patients to infer their emotional state and level of depression. However, auditory and visual cues may not be reliable for depression detection, as they can be affected by noise, lighting, or masking in unregulated or unpredictable environments. Therefore, using only the text information without relying on any auditory or visual features can be sufficient for depression detection, as it can capture the cognitive and linguistic aspects of depression [8]. Moreover, text information is less likely to raise privacy concerns than auditory and visual information, which may contain sensitive personal or biometric data that could be exploited by malicious actors. Therefore, the audio or video records of the clinical interviews are not commonly shared and distributed but it is more likely to get access to transcripts of the interviews. Considering these reasons, the proposed method totally relies on text information extracted from the transcripts of clinical interviews.

As Large Language Models (LLMs) have emerged as pioneering models in the domain of Natural Language Processing (NLP), LLMs, such as BERT [9], GPT [10], RoBERTa [11], and MPNet [12], are state-of-the-art models capable of understanding, inferring, and generating human-like text. Leveraging the power of these advanced LLMs, we employ them to create text representations for each line existing in the transcripts. The representations extracted from pretrained LLMs capture the contextual information and semantic meaning of the input text at various levels of abstraction. While these abstract text representations offer valuable linguistic features, relying solely on them can be limiting because there are various factors that reflect an individual's mental state and indicate the presence or severity of depression. In this study, we utilize external sources of information to acquire knowledge regarding the subject's personality traits, emotions, and sentiments based on the transcript data. Personality traits are relatively stable patterns of thinking, feeling, and behaving that characterize an individual. Emotions are short-term affective states that arise in response to specific events or stimuli. Sentiments are opinionated tones, typically categorized as positive, negative, or neutral. These three aspects of natural language data can provide valuable insights into one's psychological well-being and mood disorders. For example, some studies have found that certain personality traits, such as neuroticism or extraversion, are associated with a higher or lower risk of depression [13]. Other studies have shown that the recognition and understanding of emotions play a crucial role in the accurate detection and assessment of depression, highlighting the significance of emotional factors in comprehensive diagnostic procedures. [14]. Furthermore, sentiments, such as positive or negative polarity, can indicate one's satisfaction or dissatisfaction with life aspects and the severity of their mental disorder [15]. In addition to the abstract representations, our approach extracts text representations for affect and personality features from LLMs that have been fine-tuned for the corresponding classification tasks.

The main contributions of this thesis are listed below:

- We propose a novel architecture for the automatic depression severity estimation task based on multiple text modalities.
- To our best knowledge, we are the first to utilize the emotion, sentiment, and personality features in a single approach for the automatic depression severity estimation task.
- We design a new multimodal joint cross-attention technique to fuse multiple text modalities.
- We fine-tune the pretrained RoBERTa model for the utterance-based personality traits classification task and leverage it as an auxiliary model to our main approach.
- We conduct comprehensive experimental analyses for the proposed method and provide elaborative discussions on the results.

Chapter 2

Related Work

Textual features have been integrated into studies focused on the automatic depression detection task, manifesting in various methodologies. Similar to our proposed approach, several studies introduce techniques centered exclusively on text modality for this purpose. For instance, Mallol-Ragolta et al. [16] devise an architectural framework in which they input GloVe embeddings [17] of clinical transcripts into a hierarchical attention network. This network hierarchically weighs the textual components to predict the binary PHQ-8 label, indicative of depression presence or absence. In a parallel vein, Xezonaki et al. [18] formulate a hierarchical attention network aimed at PHQ-8 label prediction, utilizing transcripts. Additionally, they incorporate *external knowledge conditioning* into their methodology, encompassing facets like emotional tone, sentiment, and psycholinguistic attributes. While akin to our methodology, they leverage manually crafted features, whereas we derive them from latent representations of fine-tuned LLMs. Dinkel et al. [19] architect an ensemble comprising cascaded bidirectional gated recurrent unit (GRU) [20] layers, followed by classification and regression layers. They supply this architecture with embeddings at word and sentence levels, harnessed via Word2Vec [21], fastText [22], ELMo [23], and BERT models. Rutowski et al. [24] adapt the AWD-LSTM architecture [25] for the depression detection task, adroitly fine-tuning the core model for the target task. This fine-tuning methodology derives inspiration from the principles of ULMFiT [26]. Li et al. [27] conceptualize an architecture in which utterances within transcripts are encoded through concurrent bidirectional LSTM layers. Specifically, these bidirectional LSTM layers are nourished with BERT embeddings of said utterances. This encoding is pursued by a subsequent phase involving a mutual self-attention mechanism and a fusion operation.

As the global prominence of social media continues to surge, there arises a corresponding increase in the accessibility of personal information. The copious volumes of textual user-generated content facilitate the extraction of individual characteristics through linguistic attributes. Concurrently, social media profiles afford the monitoring of user actions, thereby furnishing supplementary insights into user mood and personality. Studies such as [28, 29] propose multimodal frameworks that amalgamate these textual attributes with user conduct indicators for the task of depression detection. In these studies, a common approach involves the utilization of hierarchical attention networks to encode user-generated posts, coupled with the incorporation of behavioral cues encompassing social network, emotional, and topic-related features. Notably, similar to our methodology, they harness ancillary information for analysis. Nevertheless, a point of departure lies in their reliance on hand-crafted features, while our approach entails the derivation of latent representations facilitated by LLMs.

Various studies involve amalgamating textual attributes with auditory and visual counterparts, with the overarching objective of imbuing the analytical paradigms with heightened informational content. Many instances of social media allow posting either textual or visual content, it enables the analysis of textual and visual features combined for depression detection from social media. Shet et al. [30] propose a method that involves a cross-domain framework. This method aims to enable depression detection in an online setting via social media for more countries with different cultural settings. Their cross-domain framework transfers the relevant information across heterogeneous domains. While doing it, they incorporate hand-crafted emotion-based textual features along with color-related visual attributes. Similar to [28, 29], they also involve behavioral cues in their input features. Lin et al. [31]. also enable online detection of depression followed by an offline training phase. They combine textual features and visual features, both collected from social media, using a low-rank multimodal fusion technique proposed by [32]. They extract the textual features from BERT and the visual features from a CNN-based binary classifier for the training procedure. As a very different approach, Gui et al. [33] adopt the reinforcement learning (RL) paradigm in their study in order to automatically select related indicator texts and images from the past posts of the users. They use a GRU layer and the pretrained VGG-Net to extract the textual and visual features, respectively. Later, a multilayer perceptron is responsible for depression classification using the features selected by the multi-agent RL model that they introduce.

Given the ease of transforming speech recordings into text through automatic speech recognition tools, and the inherent alignment between sequential textual and auditory data, the concurrent utilization of auditory and textual features has emerged as a highly favored approach. Therefore, many recent studies amalgamate textual features with auditory counterparts, with the overarching objective of imbuing the analytical paradigms with heightened informational content. In the work by Lam et al. [34], a hybrid approach is adopted, where data is initially transformed into mel-frequency spectrograms. Subsequently, textual attributes undergo encoding through transformer-based encoders [35], while spectrogram features are subjected to convolutional neural network (CNN) encoding. The ensuing step involves the concatenation of the latent representations emanating from these two distinct modalities. Meanwhile, Ghadiri et al. [36] employ a multi-pronged strategy, incorporating auditory data through low-level attributes like mel-frequency cepstral coefficients, spectrograms, and mel-frequency spectrograms. Further inclusions encompass openSMILE features [37] and graph-based descriptors. For textual input, a pre-trained BERT model is leveraged for transcript encoding. Zhang et al. [38] dissect auditory and textual features discretely for the depression detection task. In the textual realm, Doc2Vec embeddings [39] are harnessed and channeled into an Adaboost classifier. Simultaneously, various audio-text paradigms accommodate affect and personality features. In this context, Fan et al. [40] extract nuanced linguistic features through NLTK [41], complemented by sentiment facets extracted from a fine-tuned BERT model rooted in the Stanford Sentiment Treebank dataset [42]. The emphasis here diverges from our approach, where latent representations are culled from the terminal stratum of the fine-tuned network. In tandem, auditory components undergo meticulous feature engineering, followed by encoding using a multi-scale temporal dilated CNN architecture that they architect. Concatenation with textual features finalizes this fusion. Moreover, Van Steijn et al. [43] furnish an assemblage of linguistic attributes, comprising representations from a Sentence BERT network [44] and the Linguistic Inquiry and Word Count (LIWC) 2015 [45] features. The latter's demonstrated correlation with personality traits [46] bolsters its inclusion. Additionally, sentiment traits are discerned through Flair's sentiment analysis library [47], culminating in a comprehensive ensemble of attributes for analysis.

There also has been a notable surge in research endeavors dedicated to amalgamating textual, auditory, and visual modalities, with the objective of addressing the depression detection task from a multifaceted perspective encompassing diverse aspects. Pampouchidou et al. [48] present a comprehensive framework that seamlessly merges high-level and low-level features extracted from audio, video, and text data to enhance depression assessment accuracy. The proposed method involves the extraction of high-level features such as mel-frequency cepstral coefficients (MFCCs) from audio, facial expressions from video, and textual sentiment analysis from text. These features are then fused to provide a holistic view of an individual's mental state. Later, a decision tree method is used for the classification problem. Williamson et al. [49] introduce a pioneering approach that leverages vocal, facial, and semantic cues, amalgamating information from these modalities to offer a more holistic view of emotional well-being. The study employs feature extraction techniques such as pitch analysis for vocal cues, facial expression recognition for facial cues, and sentiment analysis using GloVe embeddings for semantic cues. These features are then combined and fed to a Gaussian staircase model to enhance depression detection accuracy. Sun et al. [50] focus on selecting and combining the most relevant textual, auditory, and visual features. Specifically, the method extracts relevant audio features such as MFCCs, spectrograms, and possibly prosodic cues from speech recordings. Visual features are also extracted, which may include facial expressions, body language, and other visual cues obtained from video data. These extracted auditory and visual features are then carefully selected to identify the most informative ones for depression assessment. Later, they are integrated with selected text-based features, which can include linguistic patterns, sentiment analysis, and semantic information from text data. A random forest regression model is then trained on this combined feature set to predict and assess depression levels. These studies rely on hand-crafted features and conventional machine learning algorithms. Over time, the field has transitioned towards employing more sophisticated deep learning techniques. Ray et al. [51] employ a multi-level attention network to jointly process text, audio, and video data, enhancing depression prediction through attentive feature extraction. The method involves extracting features like text embeddings from BERT models, auditory features through audio signal processing, and visual features using techniques like facial expression recognition. The multi-level attention network then dynamically weighs and fuses these features for enhanced prediction. Makiuchi et al. [52] exploit deep learning techniques that are harnessed to fuse representations from textual, auditory, and visual data, showcasing the power of neural networks in extracting meaningful features across modalities. Text data undergoes encoding by CNN layers interpreting the BERT embeddings, while auditory and visual data are processed through gated CNNs. These representations are then combined with simple concatenation. Furthermore, Zheng et al. [53] highlight the significance of modeling inter-modal relationships through a graph attention model to improve depression detection accuracy. The method involves the extraction of features from different modalities, such as textual, auditory, and visual data, and constructs a graph-based representation to capture the relationships between these features. Wei et al. [54] explore sub-attentional fusion to estimate depression across various modalities, enhancing the robustness of depression assessment. Their approach involves the extraction of sub-attentional features from textual, auditory, and visual data, with a focus on capturing subtle cues that may indicate depression. To achieve that, they employ convolutional bidirectional LSTM as their backbone architecture. These sub-attentional features are then combined using an attentional fusion approach that is inspired by the work of Dai et al. [55]. This fusion approach relies on the attention mechanism similar to our multimodal fusion approach but it is designed specifically for the feature fusion of three-dimensional input data. Lastly, Saggu et al. [56]

introduce DepressNet, a hierarchical attention mechanism that adeptly combines insights from multiple modalities, capturing both global and local patterns within the data, thereby elevating the accuracy of depression detection. The method extracts features from textual, auditory, and visual data and uses hierarchical attention mechanisms to weigh and combine these features effectively.

As mentioned, various ways of text embedding and encoding techniques are employed by existing studies. For the purpose of transcript data embedding, we leverage auxiliary networks, which have been fine-tuned across multiple downstream tasks. In congruence with [34], our text encoding is executed through the utilization of transformer encoders. While certain studies advocate the integration of varied cues to fortify the text modality, the absence of advanced fusion techniques is evident. Within the realm of processing multimodal textual features, we introduce a novel joint cross-attention fusion module into our architectural framework.

Chapter 3

Method

The primary objective of the proposed method is to accurately predict the PHQ-8 score of a given clinical interview transcript that consists of a sequence of sentences. It comprises several interconnected components. Initially, each sentence within a transcript undergoes a sequence of auxiliary networks, each designed to capture distinct representation types. Once we obtain the sentence embeddings, which we refer to as sentence tokens, for each modality, we standardize the feature dimensions to ensure equitable treatment of modalities during encoding. Subsequently, we inject a new token, called regression token, at the outset of the hidden representation matrix for each modality. These modified hidden representations then pass through a transformer block, consisting of specialized cascaded transformer encoders tailored to each modality. Later, the encoded representations of each modality are combined through a novel fusion approach: Masked Multimodal Joint Cross-Attention Fusion (MMJCA-Fusion). The resulting fused output is pooled by extracting the hidden representations associated with the [REG] token. Finally, the pooled output is fed through a feed-forward network and a regression layer. The output of the regression layer yields a prediction of the PHQ-8 score. The high-level architecture of the proposed method is visualized in Figure 3.1. In the subsequent sections of this chapter, we provide a comprehensive and detailed exposition of each component of the proposed methodology.

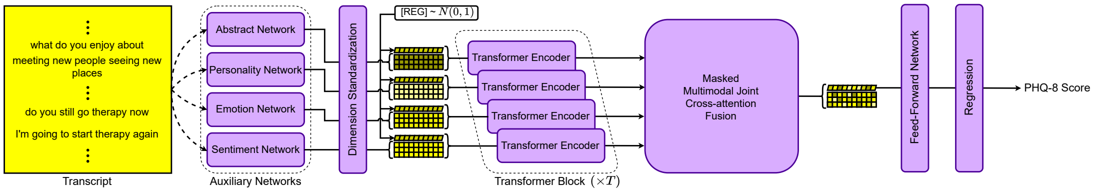
what do you enjoy about meeting new people seeing new places
do you still go therapy now
I'm going to start therapy again
Transcript
Abstract Network
Personality Network
Emotion Network
Sentiment Network
Auxiliary Networks
Dimension Standardization
[REG] ~ $\mathcal{N}(0,1)$
Transformer Encoder
Transformer Encoder
Transformer Encoder
Transformer Encoder
Transformer Block $(\times T)$
Masked Multimodal Joint Cross-attention Fusion
Feed-Forward Network
Regression
PHQ-8 Score

Figure 3.1: High-level architecture of the method. The dashed arrows indicate that there is no back-propagation in those connections.

3.1 Auxiliary Networks

We exploit four different types of representations, each enriching the model with a different aspect. Three of these representation types are for personality, emotion, and sentiment. The other one is for the abstract representations that provide the contextual information and the semantic meaning of the input sentence. We extract these different types of representations from several auxiliary pretrained LLMs.

3.1.1 Abstract Network

For the abstract representations, we employ *all-mpnet-base-v2* introduced by Reiemrs and Gurevych [44], that is the pretrained MPNet model optimized for sentence embedding.

3.1.2 Emotion & Sentiment Networks

For the affect, including emotion and sentiment, representations, we employ LLMs that are fine-tuned for the corresponding classification tasks. While choosing and adopting the fine-tuned LLMs, we ensure fairness between the representations extracted from separate models using the same fine-tuned architecture for each model. It implies that they have been pretrained and fine-tuned using similar configurations and assumptions. This mitigates potential biases that could arise if the representations were learned from models with different architectures. For the emotion representations, we employ the model introduced by Barbieri et al. [57], which has been fine-tuned for the emotion classification task. For the sentiment representations, we employ the SiEBERT model [58], which has been fine-tuned for the sentiment classification task. Both models have been fine-tuned on pretrained RoBERTa architecture.

3.1.3 Personality Network

As we endeavor to employ equitable architectures for the extraction of affect and personality representations, it is also equally imperative to acquire a corresponding framework for the extraction of personality trait representations. However, transformer-assisted text-based studies in the literature do not offer such architectures satisfying the expectations. First of all, our method requires an auxiliary model that is trained for utterance-based personality trait classification since our method proposes to extract personality representations for each sentence existing in the transcripts individually and independently. However, most of the existing studies [59–61] approach the personality traits classification task in the person-based setting. It means they make a single prediction for multiple utterances (posts, documents, etc.) that belong to the same person rather than making a prediction for each utterance itself. Some other studies [62, 63] involve psycholinguistic features alongside text information. Although Li et al. [64] introduce a considerably expedient work, they do not offer any available dataset or ready-to-use model. In light of this, we train our own personality trait detection network that we utilize to extract the corresponding representations.

We instantiate the personality trait detection network through the utilization of the pretrained RoBERTa model followed by a multi-label classification layer. The multi-label classification layer is connected to the hidden units associated with the BOS (Beginning of Sentence) token in the last layer of the RoBETRTa model. We have the multi-label classification layer since there are multiple binary personality trait categories, each indicating a different characteristic of the personality. We train the whole network, including fine-tuning all the layers of the pre-trained RoBERTAa model, using multi-label binary cross-entropy loss (L_{MLBCE}) along with a regularization factor (L_R). The total loss L_{total} can be formulated as follows:

$$L_{total} = L_{MLBCE} + \lambda L_R \tag{3.1}$$

where λ is the regularization parameter. For the calculation of L_{MLBCE}, we compute the loss value ($l_{c,b}$) for the c-th class and b-th sample in the batch. This computation is as follows:

$$\begin{aligned} l_{c,b} = &- w_c y_{c,b} \cdot \log(\sigma(\hat{y}_{c,b})) \\ &- (1 - y_{c,b}) \cdot \log(1 - \sigma(\hat{y}_{c,b})) \end{aligned} \tag{3.2}$$

where $y_{c,b} \in \{0, 1\}$ is the target value and $\hat{y}_{c,b}$ is the output of the last linear layer for the c-th class and b-th sample in the batch. Also, $w_c = \frac{N_{c,positive}}{N_{c,negative}}$ is the weight of the positive answer for the c-th class, where $N_{c,positive}$ is the total number of positive samples and $N_{c,negative}$ is the total number of negative samples. We add these weights to avoid possible biases that could occur due to class imbalance. Then, the loss vector for the c-th class becomes $L_c = [l_{c,1}, ..., l_{c,B}]$, where B is the batch size. On top of this, the loss matrix becomes $L = [L_1; ...; L_C]$, where C is the number of classes. Afterward, we compute the final scalar L_{MLBCE} value for each batch by calculating the mean of the L matrix as:

$$L_{MLBCE} = \text{mean}(L) \tag{3.3}$$

In addition, we compute L_R as:

$$L_R = \text{Variance}(\{||w_1||_2, ..., ||w_C||_2\}) \tag{3.4}$$

where w_c is the weight vector associated with the c-th output neuron. The reason we use the regularization factor is we can assume that the four categories of personality types cannot be totally independent of each other. Therefore, we avoid the output weights drifting apart from each other using this factor.

To extract the affect and personality representations from the auxiliary fine-tuned RoBERTa models, we derive the hidden representations of the BOS token from the last layer of the models. By deriving the hidden representations of the BOS token from the last layer, we obtain a condensed representation of the input sentence that captures the learned features and context relevant to the corresponding traits. We further exploit these representations for the depression severity estimation task, treating them as distinct modalities.

Once we derive the representations from the auxiliary networks for each sentence existing in a transcript, we obtain a sequence of embedded sentences. We refer to these as sentence tokens. To standardize the number of sentence tokens in each transcript, we apply [PAD] tokens to equalize the number of sentence tokens in each transcript. As a result, each transcript is represented by a unique matrix with the shape of $S_{max} \times D_m$. Here, S_{max} denotes the sequence length of the transcript with the maximum number of samples among the transcripts that exist in the training set. D_m refers to the hidden dimension of the m-th auxiliary network.

3.2 Dimension Standardization

Since the text representations are extracted from a separate auxiliary network, the feature dimension for each modality may differ. To ensure fair competition between the modalities during encoding, we standardize the feature dimension to a specific value, D. In cases where the feature dimension was initially not equal to D, we pass it through linear transformation, GELU activation [65], and layer normalization [66] layers to standardize it.

GELU is an activation function commonly used in deep neural networks. Given a sequential input tensor x of shape (N, S, D), where N is the batch size, S is the sequence length ($S = S_{max}$ through the dimension standardization process), and D is the feature dimension, the GELU activation function is applied element-wise as follows:

$$\mathrm{GELU}(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \cdot x \tag{3.5}$$

Here, x is the input tensor element and $\mathrm{erf}(\cdot)$ is the error function, which is a standard mathematical function used to calculate the error probability in statistics and the cumulative distribution function of the standard normal distribution in probability theory. This function maps each element of the input tensor x to its corresponding GELU-activated value.

Layer Normalization is a technique used to normalize the activations of a layer within a neural network. In order to apply Layer Normalization to a sequential input tensor x of shape (N, S, D), we first compute the mean μ and standard deviation σ for each example and sequence independently across the feature dimension:

$$\mu_{js} = \frac{1}{D}\sum_{d=1}^{D} x_{jds} \tag{3.6}$$

$$\sigma_{js} = \sqrt{\frac{1}{D}\sum_{d=1}^{D}(x_{jds} - \mu_{js})^2} \tag{3.7}$$

Here, j indexes the examples in the batch, ranging from 1 to N. s indexes the sequences in the input, ranging from 1 to S. d indexes the features in the input tensor, ranging from 1 to D.

Then, Layer Normalization standardizes the input tensor x for each example and sequence independently as follows:

$$\mathrm{LN}(x)_{jds} = \frac{x_{jds} - \mu_{js}}{\sigma_{js}} \tag{3.8}$$

where $\mathrm{LN}(x)$ is the normalized output, μ_{js} is the mean computed across the feature dimension, and σ_{js} is the standard deviation computed across the feature dimension.

3.3 Addition of Regression Token

Inspired by the [CLS] token technique introduced in BERT, we include an additional token in our method. We refer to this token as the regression token ([REG] token). The primary aim of the [REG] token is to facilitate the regression task in the context of multi-modal sequential data processing. We prepend a [REG] token to the beginning of each modality's sequential data. This token is initialized with values drawn from the Normal distribution $N(\mu = 0, \sigma = 1)$. As a result of this addition, the shape of the hidden representation matrices becomes $S \times D$, where $S = S_{max} + 1$.

By incorporating the [REG] token into the input sequence, we aim to provide the model with dedicated representations that encapsulate essential information about the regression task. During the fusion of multiple modalities, we ensure that the information from the [REG] token is seamlessly integrated into the combined output. After the fusion process is complete, we employ a pooling strategy where we extract the representations associated with the [REG] token. These [REG] token representations serve as a critical bridge between the modalities and the regression task, enabling our model to focus on relevant information. These dedicated representations ensure that our model can effectively leverage the information encapsulated by the [REG] token to make informed predictions, effectively addressing the regression problem within the multimodal context.

3.4 Transformer Block

In the transformer block, there exist T number of cascaded transformer encoders [35] for each modality. Inside a single transformer encoder, we first feed $X_{m,t}$, which is the input for the t-th transformer encoder of the m-th modality, to the masked multi-head self-attention layer. This layer is composed of the concatenation of H number of heads which are obtained in a parallel manner. The computation of the heads incorporates the self-attention mechanism. The self-attention mechanism is a variant of scaled dot product attention where the query, key, and value matrices are derived from the linear projection of the same input matrix. The query, key, and value matrices of each head utilize unique linear projections of $X_{m,t}$. For the h-th head, the query ($Q_{h,m,t}$), key ($K_{h,m,t}$), and value ($V_{h,m,t}$) matrices are calculated as follows:

$$Q_{h,m,t} = X_{m,t} W_{query,h,m,t} \tag{3.9}$$

$$K_{h,m,t} = X_{m,t} W_{key,h,m,t} \tag{3.10}$$

$$V_{h,m,t} = X_{m,t} W_{value,h,m,t} \tag{3.11}$$

where $W_{query,h,m,t} \in \mathbf{R}^{D \times D_h}$, $W_{key,h,m,t} \in \mathbf{R}^{D \times D_h}$, $W_{value,h,m,t} \in \mathbf{R}^{D \times D_h}$ are the weight matrices. Here, D_h is the same for all heads and is calculated as $D_h = \frac{D}{H}$. Accordingly, we calculate the attention scores of the h-th head as follows:

$$A_{h,m,t} = \mathrm{softmax}(\frac{Q_{h,m,t} K^T_{h,m,t}}{\sqrt{D_h}}) \tag{3.12}$$

The attention scores represent the weights assigned to sentence tokens in the input sequence, reflecting their importance for the context. However, the [PAD] tokens that we insert into the sequence do not contribute any information to the context. Therefore, it is necessary to exclude them from the attention map. In light of this, we mask out the attention scores associated with the [PAD] tokens by setting their values to $-\infty$. On top of this, we calculate the result of the h-th head by multiplying the attention scores with the value matrix as follows:

$$head_{h,m,t} = A_{h,m,t} V_{h,m,t} \tag{3.13}$$

Further, we concatenate the heads and linearly project them to obtain the output Om, t of the masked multi-head self-attention layer.

$$O_{m,t} = [head_{1,m,t}, \ldots, head_{H,m,t}] W_{output,m,t} \tag{3.14}$$

where $W_{output,m,t} \in \mathbf{R}^{D \times D}$ is the weight matrix. Later, we feed this output matrix to a residual connection and a Layer Normalization layer.

$$X'_{m,t} = \mathrm{LN}(O_{m,t} + X_{m,t}) \tag{3.15}$$

where LN stands for Layer Normalization. For the rest, we employ a feed-forward network followed by the repetition of Eq. 3.15.

$$X_{m,t+1} = \mathrm{LN}(\mathrm{FFN}_{m,t}(X'_{m,t}) + X'_{m,t}) \tag{3.16}$$

Here, $\text{FFN}_{m,t}$ stands for a two-layer feed-forward network. The first layer projects the hidden representations from D dimensional space to $4D$ dimensional space and the second layer projects back to D dimensional space. Both layers utilize the GELU activation function for non-linearity.

For each modality m, we obtain $X_{m,T+1}$ as the result of the last (T-th) transformer encoder. This is also the output of the transformer block and is denoted as $\hat{X}_m$, which is equal to $X_{m,T+1}$.

3.5 Masked Multimodal Joint Cross-Attention Fusion (MMJCA-Fusion)

As different modalities convey diverse information related to their own context, it is crucial to effectively capture their complementary relationship. To merge these modalities, we utilize a cross-attention-based fusion approach, which encodes inter-modal information while preserving the intra-modal dependencies. To achieve that, we rely on the cross-attention between the individual modalities and the joint representation, which is the concatenation of the modalities over the feature dimension. Previous works such as [67] and [68] also propose to incorporate joint representations for their cross-attention fusion mechanisms. Their attention mechanism operates on the feature or modality levels. They aim to capture the dependencies between visual and audial representations across feature or modality dimensions. In our method, the positions of sentence tokens correspond to each other across the modalities. In light of this, we compute the attention maps between the hidden representations of the sentence tokens along separate modalities to capture the dependencies between the sentence tokens located in the same position. This enables sequence information modeling for each transcript along multiple modalities. The architecture of the Masked Multimodal Joint Cross-Attention Fusion (MMJCA-Fusion) module is visualized in Figure 3.2.

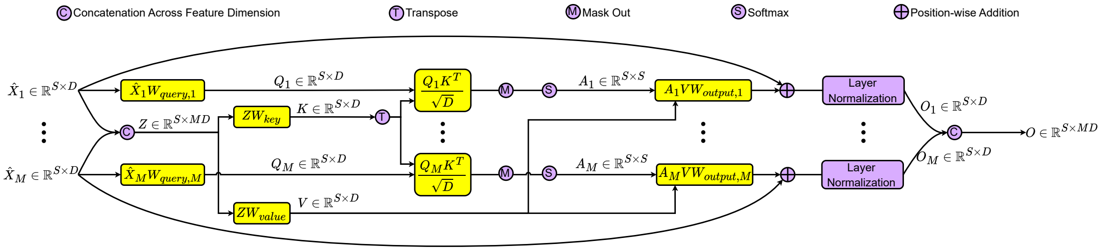
C Concatenation Across Feature Dimension
T Transpose
M Mask Out
S Softmax
Position-wise Addition
$\hat{X}_1 \in \mathbb{R}^{S\times D}$
$\hat{X}_M \in \mathbb{R}^{S\times D}$
$\hat{X}_1 W_{query,1}$
$\hat{X}_M W_{query,M}$
$Z \in \mathbb{R}^{S\times MD}$
ZW_{key}
ZW_{value}
$Q_1 \in \mathbb{R}^{S\times D}$
$K \in \mathbb{R}^{S\times D}$
$Q_M \in \mathbb{R}^{S\times D}$
$V \in \mathbb{R}^{S\times D}$
$\frac{Q_1 K^T}{\sqrt{D}}$
$\frac{Q_M K^T}{\sqrt{D}}$
$A_1 \in \mathbb{R}^{S\times S}$
$A_M \in \mathbb{R}^{S\times S}$
$A_1 V W_{output,1}$
$A_M V W_{output,M}$
Layer Normalization
Layer Normalization
$O_1 \in \mathbb{R}^{S\times D}$
$O_M \in \mathbb{R}^{S\times D}$
$O \in \mathbb{R}^{S\times MD}$

Figure 3.2: Masked Multimodal Joint Cross-Attention Fusion

First, we linearly project the joint representation matrix, $Z \in \mathbb{R}^{S\times Md_{model}}$, two times separately using the transformation matrices $W_{key} \in \mathbb{R}^{MD\times D}$ and $W_{value} \in \mathbb{R}^{MD\times D}$. Here, S denotes the sequence length. Basically, the key and value matrices for the cross-attention operation become:

$$K = ZW_{key} \tag{3.17}$$

$$V = ZW_{value} \tag{3.18}$$

To obtain the query matrix, we also linearly project the encoded representation matrix $\hat{X}_m \in \mathbb{R}^{S\times D}$ for the m-th modality where $i \in \{1, ..., M\}$ and $M > 1$ is the number of modalities. Hereby, the query matrix becomes:

$$Q_i = \hat{X}_m W_{query,m} \tag{3.19}$$

where $W_{query,m} \in \mathbb{R}^{D\times D}$ is the transformation matrix. Further, we calculate the attention scores, similar to Eq. 3.12.

$$A_m = \text{softmax}(\frac{Q_m K^T}{\sqrt{D}}) \tag{3.20}$$

Similar to the approach that we employ in the transformer block, we mask out the attention scores associated with the [PAD] tokens by setting their values to ∞. On top of it, we obtain the result of each attention operation by multiplying the attention scores with the value matrix followed by a linear projection.

$$O'_m = A_m V W_{output,m} \tag{3.21}$$

where $W_{output,m} \in \mathbb{R}^{D\times D}$ is the transformation matrix for the linear projection. Afterward, we employ residual connection and layer normalization on top of the attention operation as follows:

$$O_m = \text{LN}(O'_m + \hat{X}_m) \tag{3.22}$$

If we denote $o_{reg,i}$ as the vector that corresponds to the [REG] token index in O_i, the final output, $O = [o_{reg,1}, o_{reg,2}, ..., o_{reg,M}]$ is the concatenation of each $o_{reg,i}$ across the feature dimension.

3.6 Feed-Forward Network & Regression

At the final stage of our model architecture, we employ a feed-forward network followed by a regression layer. The feed-forward network consists of a series of neural network layers. It extracts higher-level information from O by progressively transforming it into lower-dimensional spaces. Considering the number of layers is denoted by L, the set, $\Sigma = \{\Sigma_1, \dots \Sigma_L\}$ indicates the number of neurons that exist in the neural network layers. For instance, the l-th layer contains Σ_l neurons. In addition, each layer utilizes GELU activation function for undergoing non-linear transformation.

The output of the last neural network layer is connected to the regression layer. Through a linear regression operation, this layer predicts a scalar continuous value that corresponds to the PHQ-8 score.

For the optimization of the weights in the architecture, we employ we employ the Concordance Correlation Coefficient (CCC) as the loss function, which is effective in measuring the agreement between two variables and highly adaptive to the regression tasks. CCC is calculated as follows:

$$CCC = \frac{2\rho\sigma_{\hat{y}}\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2} \tag{3.23}$$

where $\sigma_{\hat{y}}$ and σ_y are the standard deviations and $\mu_{\hat{y}}$ and μ_y are the means of the predictions and the target values, respectively. Also, ρ is Pearson's correlation coefficient between the predictions and the true values. Regarding this, we modify CCC to obtain the loss function as follows:

$$L_{CCC} = 1 - CCC \tag{3.24}$$

Chapter 4

Experimental Setup

We build the experimental setup with the aim of analyzing the effects of components that are included in or excluded from the model architecture. These components involve employed auxiliary networks, the adaptation of temporal modeling, the existence of the transformer block, the type of the pooling method, and the fusion approach. We define a single experiment as the process of obtaining the best model state utilizing a determined set of model components. Each experiment consists of training procedures where we tune the hyperparameter values. During each of these procedures, the model parameters are trained on only the predetermined training set. The model does not encounter any samples from the predetermined validation or test sets during the learning phase. The validation set is used to evaluate the model performance during the training. The test set is used to measure the test performance of the final model that is obtained after each experiment. In this section, we explain how we configure the training procedures, tune the hyperparameters, and select the best model state for each experiment. Then, we define the evaluation metric that we use to perform the assessments.

4.1 Datasets

In this section, we provide a comprehensive overview of the datasets employed in our study to evaluate the proposed depression severity estimation network, and the datasets that have been used for fine-tuning the sentiment, emotion, and personality networks, leveraged as integral components.

4.1.1 Dataset Used for the Assessment of the Proposed Depression Severity Estimation Network

For the assessment of the proposed depression severity estimation network, we use the E-DAIC dataset [69], which is an extension of the DAIC-WoZ dataset [70], provided by the AVEC'19 Detecting Depression with AI Subchallenge [71]. The data consists of semi-clinical video interviews, including video features, audio recordings, and automatic transcriptions generated by Google's Automatic Speech Recognition (ASR) tool. The dataset is divided into fixed sets for training, development, and testing, comprising 163, 56, and 56 interviews, respectively. The interviews were conducted in a Wizard-of-Oz (WoZ) scenario by two humans controlling a virtual agent (Ellie) or by a fully automated AI. The training and development sets contain a mix of WoZ and AI settings, while the test set only includes the AI setting. All interviewees filled out the eight-item Patient Health Questionnaire (PHQ-8), providing scores for each of the eight symptoms and their total depression score. The total depression score, ranging from 0 to 24, is the sum of the eight-item scores. The distribution of the number of subjects over PHQ-8 scores for each set is detailed in Figure 4.1.

The quality and interpretability of the dataset are crucial for machine learning tasks. However, the datasets come with challenges and limitations most of the time so the proposed methods should be capable of accomplishing the challenges and overcoming the limitations. The dataset that we use also introduces some, which we detail as follows:

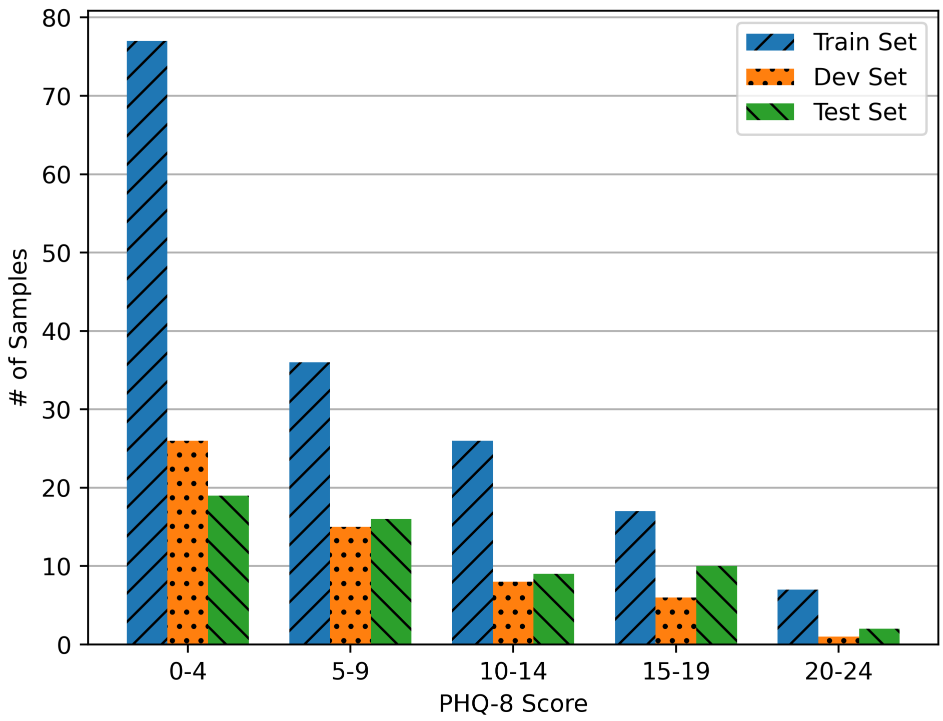

Figure 4.1: Data distributions across different intervals of PHQ-8 Score.

- The acoustic of environments and the quality of recordings differ for separate interviews. These cause noisy variances among auditory features and inaccuracies in automatic generations of transcripts.

- There are no human interventions or manual corrections throughout the preparation of transcripts. Thus, we encounter mistakes in the text, caused by the failures that occurred during the automatic speech recognition. The textual feature extraction is affected by such mistakes significantly.

- Aside from the lack of manual correction, there is no manually tagged speaker information on transcripts. So, it is not known which parts of the transcripts belong to the interviewee and which parts belong to Ellie or the AI agent. This information can be derived from audial features but it is a certain bottleneck for textual features.

- Even though we use the extended version of DAIC-WoZ dataset, the size of the dataset is considerably small. Modern technologies in the machine learning field require reasonably large datasets to train and validate the models they include. Due to consistently growing architectures, it gets harder to avoid the overfitting issue. Therefore, the size of the dataset is a significant challenge for adjusting the complexity of our proposed methods.

- As can be observed from Figure 4.1, the dataset is plagued by significant class imbalance. This leads to deflections in the prediction of PHQ-8 scores. It impels the model to predict scores from a certain scope of the range and reduces the performance during the validation and testing phases.

4.1.2 Datasets Used for the Fine-tuned Auxiliary Networks

In order to fine-tune the RoBERTa model for the sentiment classification task, Hartmann et al. [58] exploit 15 different datasets, introduced by the studies [72–79] and offered publicly on Kaggle[1] and Yelp[2]. The combination of these datasets encompasses various domains, including tweets, movie reviews from IMDb[3] and Rotten Tomatoes[4], product and kitchen appliance reviews from Amazon[5], and restaurant reviews from Yelp[6]. It consists of 1,253,000 samples and is labeled with fine-grained sentiment scores, ranging from very negative to very positive. The authors only consider the binary version of the dataset, where the sentences are classified as either positive or negative.

In order to fine-tune the RoBERTa model for the emotion recognition task, Barbieri et al. [57] exploit the Affect in Tweets dataset, introduced by Mohammad et al. [80]. The dataset consists of 174,356 tweets annotated with multi-labeled emotion classes that are listed as follows:

[1] https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews

[2] https://www.yelp.com/dataset

[3] https://www.imdb.com/

[4] https://www.rottentomatoes.com/

[5] https://www.amazon.com/

[6] https://www.yelp.com/

- anger, including annoyance, rage
- anticipation, including interest, vigilance
- disgust, including disinterest, dislike, loathing
- fear, including apprehension, anxiety, terror
- joy, including serenity, ecstasy
- love, including affection
- optimism, including hopefulness, confidence
- pessimism, including cynicism, no confidence
- sadness, including pensiveness, grief
- surprise, including distraction, amazement
- trust, including acceptance, liking, admiration
- neutral or no emotion

Barbieri et al. re-purpose this multi-label dataset into a multi-class classification problem, keeping only the tweets labeled with a single emotion. In order to tackle the scarcity of the number of tweets with single labels, they select the tweets annotated with the most common four emotions: anger, joy, sadness, and optimism. This ends up with a total number of 5,052 samples for the fine-tuning process.

We fine-tune the personality network, using the Kaggle MBTI dataset[7]. Notably, this dataset has been leveraged in contemporary research endeavors for the personality trait classification task [60, 63, 81]. The dataset is sourced from PersonalityCafe[8], a platform where individuals share their personality types and engage in discussions on topics such as health, behavior, and personal growth.

[7] https://www.kaggle.com/datasets/datasnaek/mbti-type/

[8] https://www.personalitycafe.com/

The dataset comprises 8675 users, with each user contributing approximately 45-50 posts. The data is labeled based on the Myers-Briggs Type Indicator (MBTI) taxonomy [82], which divides personality types into four categories: Introversion / Extroversion, Sensing / Intuition, Thinking / Feeling, and Perception / Judging.

4.2 Training Configuration

Throughout the training process, the model parameters (weights and biases) are initialized and then aimed at achieving optimization subject to iterative updates. Exclusively, we freeze the fine-tuned parameters of the auxiliary networks since they do not involve back-propagation. We use Xavier's method [83] for the initialization of the parameters and we use Adam optimizer [84] with weight decay of 10^{-5} and an epsilon value of 10^{-8} to update them. We designate the learning rate (lr) as a hyperparameter, dictating the magnitude of parameter updates during the optimization process. Furthermore, we apply the dropout technique [85] on all model parameters as a regularization technique. We also designate the dropout probability (p) as a hyperparameter.

4.3 Hyperparameter Tuning

For each experiment, we conduct an extensive hyperparameter tuning process. The selection of hyperparameters is based on the random search algorithm. During the random search, multiple training procedures are executed with randomly selected configurations of hyperparameter values. The configuration that has achieved the highest validation score indicates the best model for that experiment. The model selection process is explained elaborately in Section 4.4. We ensure the amount of training procedures is the same for each experiment. The considered values for each hyperparameter are listed in Table 4.1.

Table 4.1: Considered values for each hyperparameter.

Hyperparameter	**Considered Values**
# of Transformer Encoders (K)	{1, 2, 3}
# of Heads in Self-Attention Layers (H)	{1, 2, 4, 8}
# of Neurons in The Layers of FFN (Σ)	{{1024, 256, 16}, {1024, 64}, {256, 64}, {64, 16}, {256}}
Learning Rate (lr)	$[10^{-6}, 10^{-4}]$
Dropout Probability (p)	{0, 0.1, 0.2, 0.3}
Batch Size	{8, 16, 32, 64}

4.4 Model Selection

Each training procedure runs for a maximum value of 250 epochs. We also apply the early-stoppage criterion with a patient parameter of 10. We evaluate the validation score after each epoch is completed. In case the validation score does not improve for 10 epochs, the training procedure terminates. Applying the early-stoppage criterion reduces both the risk of overfitting and the training time to be consumed. As the training procedure ends, we select the model state that has achieved the highest validation score as the candidate model. At the end of each experiment, we obtain the best model similarly by selecting among the candidate models.

4.5 Evaluation Metrics

To assess the performance of the methods that are utilized in the experiments, we use three distinct evaluation metrics: Concordance Correlation Coefficient (CCC), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

The CCC measures the agreement between predicted values and ground truth labels, considering both the mean and variance of the data, as formulated in Eq. 3.23. A higher CCC score indicates better agreement between predictions and true values, with 1 indicating perfect agreement and 0 representing no agreement.

RMSE is a common regression metric that quantifies the average deviation between predicted and target values, giving more weight to larger errors due to the squaring of differences. It calculates the square root of the mean of squared differences, providing a measure of the model's accuracy in predicting continuous values. Lower RMSE values indicate better model performance. Considering the number of test samples as N, the prediction value as $\hat{y}$, and the true value as y, RMSE is formulated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{4.1}$$

MAE is another regression metric measuring the average deviation between predicted and target values. It provides a similar evaluation as RMSE but is less sensitive to outliers since MAE treats all errors equally, without giving more weight to larger errors. Like RMSE, lower MAE values indicate better model accuracy. MAE is formulated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{4.2}$$

To monitor the progress of the model during the training procedures and to compare the empirical results during the assessments, we consistently utilize the CCC metric. The reasons are (1) the CCC metric effectively assesses how well the model predictions match the ground truth values and (2) the AVEC'19 Subchallenge declares it as the main metric of the task. During the analyses demonstrated in Sections 5.3 and 5.5, where we take account of only our final model, we also involve the RMSE and MAE metrics.

Chapter 5

Experiments & Results

In this section, we expound upon the executed experiments and present their outcomes, accompanied by meticulous analyses. Initially, our focus revolves around conducting a series of experiments aimed at assessing the significance of the components comprising our model architecture. These experimental procedures are primarily segregated into two distinct sections: encompassing the evaluations of unimodality and multimodality. For each of these sections, we compare the validation result of each experiment with the validation result achieved by the best model state we have obtained at the end of all experiments throughout the analyses of the empirical outcomes. Moreover, we ensure to keep other components the same as controlling variables while experimenting for one. three distinct subgroups, encompassing the assessments of various combinations of auxiliary networks, the existence of temporal modeling, and the fusion approach. Throughout the analyses of the empirical outcomes, we compare the validation result of each experiment with the validation result achieved by the best model state we have obtained at the end of all experiments.

Further, we enhance our analyses by investigating the performance of the best model over the test set. First, we assess the attention weights extracted from the MMJCA-Fusion module while inferring two samples selected from the test set. Second, we execute a more elaborate evaluation by applying the test metrics to small portions of the test set that have been created with respect to the ground truth labels. Lastly, we compare the final results on the test set with other state-of-the-art studies in order to ensure the credibility of our findings.

5.1 Evaluations in the Unimodal Setting

The objective of this assessment is to explore the impact of auxiliary networks individually on the performance of our depression severity estimation model. To achieve this, we conduct a series of experiments in the unimodal setting.

5.1.1 Assessment of Auxiliary Networks Individually

First, we examine the model for each modality independently. To achieve this, the MMJCA-Fusion module is omitted from the architecture, and the network is instantiated separately for each modality. This assessment provides insights into the individual performance of each modality when they are treated in isolation. The high-level architecture of the method modified for the unimodal setting is visualized in Figure 5.1 and the individual performance of each auxiliary network is represented in Table 5.1.

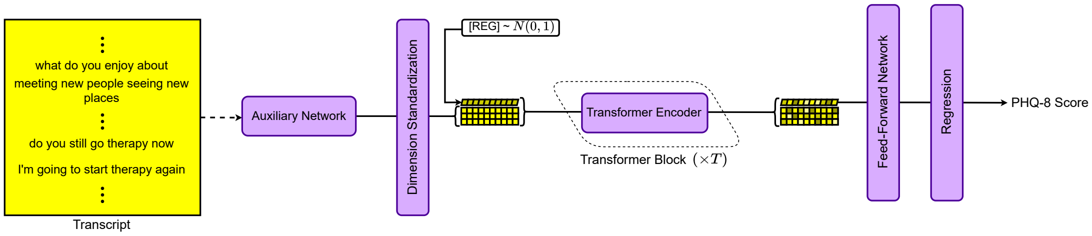
⋮
what do you enjoy about meeting new people seeing new places
⋮
do you still go therapy now
I'm going to start therapy again
⋮
Transcript
Auxiliary Network
Dimension Standardization
[REG] ~ $N(0,1)$
Transformer Encoder
Transformer Block $(\times T)$
Feed-Forward Network
Regression
PHQ-8 Score

Figure 5.1: High-level architecture of the method modified for the unimodal setting. The dashed arrow indicates that there is no back-propagation in that connection.

Table 5.1: Development and test CCC scores when the auxiliary networks are utilized individually. The bold values indicate the best scores.

Auxiliary Network				Dev CCC	Test CCC
Personality	Emotion	Sentiment	Abstract		
✓				0.636	0.448
	✓			**0.645**	**0.640**
		✓		0.456	0.457
			✓	0.640	0.618

The findings presented in Table 5.1 demonstrate that each auxiliary network contributes valuable information to depression severity estimation. However, notable differences are observed, particularly with the sentiment network, which lags significantly behind the other networks in terms of the CCC results. This discrepancy can be attributed to the multifaceted nature of depression, which encompasses various emotional, cognitive, and behavioral aspects. Simple positive or negative sentiment analysis is insufficient to grasp the complexity of this mental health condition. Instead, a comprehensive understanding requires incorporating personality traits and emotions, as they offer a more nuanced and enriched perspective on an individual's mental state.

Additionally, it is intriguing to note that the utilization of the abstract network achieves a CCC score comparable to that of the personality traits and emotion networks. This suggests that our proposed method effectively accomplishes the depression severity estimation task even without employing affect or personality representations. Instead, it successfully captures the intricate dependencies between abstract representations derived from the transcripts and the PHQ-8 score, implying that the abstract network is capable of extracting meaningful patterns and features related to depression.

5.1.2 Assessment of Temporal Modeling in the Unimodal Setting

Further, we delve into investigating the influence of introducing temporal modeling on individual modalities in the unimodal setting. To achieve this, a temporal modeling module that is composed of a variant of recurrent neural networks is inserted subsequent to the transformer block of the network. This enables the network to capture temporal dependencies and sequential patterns within each modality. Similar to Section 5.1.1, the MMJCA-Fusion module is excluded to maintain the unimodal nature of this comparison.

Moreover, we add a pooling layer after the temporal modeling module to obtain the comprehended representation of the input sequence. We do not incorporate the [REG] token approach since the hidden state associated with the [REG] token after passing through a recurrent neural network layer would not capture the same type of high-level information as the [REG] token's original representations. [REG] token embedding is designed to encapsulate the entire sequence's information, while that hidden state represents the sequence in the context of the recurrent neural network's own internal processing. The high-level architecture of the method modified for the unimodal setting including temporal modeling is visualized in Figure 5.2

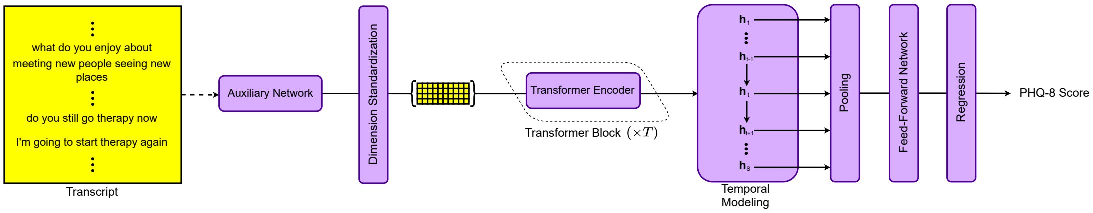

Figure 5.2: High-level architecture of the method modified for the unimodal setting including temporal modeling. The dashed arrow indicates that there is no back-propagation in that connection. h_t represents the hidden state for the t-th time step in the temporal modeling module.

We explore variations in the type of recurrent layer, the number of cascaded recurrent layers, and the pooling method, seeking to identify the optimal configuration that yields superior performance. The evaluation process involves employing different combinations of these components and comparing their results against one another. The results are presented in Table 5.2

Two prominent recurrent layer types, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), are considered for capturing temporal dependencies within the unimodal data. These layers exhibit different gating mechanisms and memory retention characteristics. The LSTM cell is characterized by its separate memory cell and output gate, while the GRU combines these gates into a unified update gate and reset gate. Mathematically, LSTM is defined as follows:

Mathematically, the LSTM unit consists of several key components that govern its operation at each time step t. Given an input sequence x_t at time t, an LSTM unit computes the following transformations:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{5.1}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{5.2}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{5.3}$$

$$g_t = \tanh(W_g x_t + U_g h_{t-1} + b_g) \tag{5.4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{5.5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{5.6}$$

$$\tag{5.7}$$

Here, f_t, i_t, and o_t represent the forget, input, and output gates respectively, controlling the flow of information in the LSTM. g_t is the candidate value to be added to the memory cell, c_t, while h_t is the LSTM's hidden state at time t. W and U represent the weight matrices, and b represents the bias terms for the respective gates. The symbol $\odot$ denotes element-wise multiplication, σ represents the sigmoid activation function, and tanh stands for the hyperbolic tangent activation. In words, the forget gate decides which information from the previous cell state c_{t-1} to discard, the input gate determines new information to be added to the cell state, and the output gate controls the information to be exposed in the hidden state h_t. The candidate value g_t is computed based on the current input and the previous hidden state, which, after gating, contributes to the updated cell state c_t.

Given an input sequence x_t at time t similar to the defined LSTM unit, a GRU unit is defined as follows:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \tag{5.8}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \tag{5.9}$$

$$\tilde{h}_t = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t] + b_h) \tag{5.10}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{5.11}$$

Here, h_{t-1} is the hidden state from the previous time step, z_t is the update gate that determines how much of the previous state should be retained, and r_t is the reset gate that controls the information from the current input that should be added to the new memory content. Similar to the LSTM unit, W represents the weight matrices, and b represents the bias terms for the respective gates. In this formulation, the update gate allows the network to decide whether to blend the previous hidden state with the new memory content, enabling the model to remember or forget information as needed. The reset gate controls how much of the previous state information to reset based on the current input, enhancing the model's adaptability to different patterns in the data.

The number of cascaded recurrent layers is another critical factor to consider. It determines the depth of the temporal modeling hierarchy within the network. We experiment with employing one and two layers of the chosen recurrent type. This allows us to gauge whether deeper temporal modeling yields improved performance compared to a single-layer approach.

Pooling methods are employed to aggregate the temporal information extracted by the recurrent layers into a fixed-length representation. In this assessment, we focus on two pooling methods: Last-Pooling and Max Pooling. In the last-pooling method, the output of the last time step from the recurrent layer is selected as the aggregated representation. This captures the final temporal state of the sequence. Max pooling involves selecting the maximum value along each dimension of the temporal sequence. This method aims to capture the most salient features present at different time steps.

Table 5.2: Development and test CCC Scores across different configurations: auxiliary network, recurrent layer type, number of recurrent Layers, and pooling method. The bold values indicate the best scores.

Auxiliary Network	**Recurrent Layer Type**	**# of Recurrent Layers**	**Pooling Method**	**Dev CCC**	**Test CCC**
Abstract	LSTM	1	Last	0.595	0.616
			Max	0.541	0.639
		2	Last	0.575	0.664
			Max	0.540	0.623
	GRU	1	Last	0.627	0.654
			Max	0.575	0.527
		2	Last	0.638	0.635
			Max	0.567	0.604
Emotion	LSTM	1	Last	0.589	0.607
			Max	0.631	0.575
		2	Last	0.577	0.626
			Max	0.611	0.650
	GRU	1	Last	0.596	0.620
			Max	**0.650**	**0.675**
		2	Son	0.601	0.648
			Max	0.610	0.611
Sentiment	LSTM	1	Last	0.440	0.34
			Max	0.467	0.416
		2	Last	0.413	0.364
			Max	0.431	0.392
	GRU	1	Last	0.441	0.415
			Max	0.466	0.411
		2	Last	0.436	0.358
			Max	0.414	0.388
Personality	LSTM	1	Last	0.625	0.621
			Max	0.634	0.619
		2	Last	0.626	0.552
			Max	0.623	0.512
	GRU	1	Last	0.616	0.553
			Max	0.641	0.556
		2	Last	0.624	0.555
			Max	0.636	0.520

After conducting an exhaustive assessment of the various configurations, it is ascertained that the outcomes align consistently with the data presented in Table 5.1 when accounting for the interplay among the auxiliary networks.” The results also show that employing a single layer of GRU with max pooling with the emotion network achieves the highest CCC score. This indicates that the GRU’s gating mechanism combined with max pooling is particularly effective in capturing relevant temporal patterns. The single-layer architecture suggests that for the dataset and task under consideration, additional layers did not significantly contribute to improved performance.

To further scrutinize the impact of different temporal modeling components, we conduct an additional assessment, focusing on the integration of bidirectional recurrent layers. As employing LSTM and two cascaded layers does not consistently improve the performance, we aim to understand the influence of bidirectional information flow on the unimodal temporal modeling task while maintaining the recurrent layer type as GRU and a single cascaded layer to mitigate computational complexity. This exploration allows us to investigate whether bidirectional modeling could enhance the network’s ability to capture temporal dependencies effectively. The results are presented in Table 5.3

Bidirectional recurrent layers enable the network to consider both past and future contexts when processing each time step, potentially capturing a more comprehensive representation of temporal patterns. In our assessment, we utilized the bidirectional variant of GRU. Mathematically, bidirectional GRU can be defined as follows:

$$\overrightarrow{z}_t = \sigma(W_{\overrightarrow{z}} x_t + U_{\overrightarrow{z}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{z}}) \tag{5.12}$$

$$\overleftarrow{z}_t = \sigma(W_{\overleftarrow{z}} x_t + U_{\overleftarrow{z}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{z}}) \tag{5.13}$$

$$\tilde{h}_t = \tanh(W_h x_t + U_{\overrightarrow{h}}(\overrightarrow{r}_t \odot \overrightarrow{h}_{t-1}) + U_{\overleftarrow{h}}(\overleftarrow{r}_t \odot \overleftarrow{h}_{t+1}) + b_h) \tag{5.14}$$

$$\overrightarrow{h}_t = (1 - \overrightarrow{z}_t) \odot \overrightarrow{h}_{t-1} + \overrightarrow{z}_t \odot \tilde{h}_t \tag{5.15}$$

$$\overleftarrow{h}_t = (1 - \overleftarrow{z}_t) \odot \overleftarrow{h}_{t+1} + \overleftarrow{z}_t \odot \tilde{h}_t \tag{5.16}$$

$$h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t] \tag{5.17}$$

Here, $\overrightarrow{z}_t$ and $\overleftarrow{z}_t$ represent the update gates for the forward and backward directions respectively. The weight matrices $W_{\overrightarrow{z}}$, $U_{\overrightarrow{z}}$, $W_{\overleftarrow{z}}$, and $U_{\overleftarrow{z}}$ correspond to input-to-gate and hidden-to-gate connections for the forward and backward directions. Additionally, $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ denote the hidden states for the forward and backward directions, while $\overrightarrow{r}_t$ and $\overleftarrow{r}_t$ stand for the reset gates for these directions. The weight matrices W_h, $U_{\overrightarrow{h}}$, and $U_{\overleftarrow{h}}$ capture the interactions within the bidirectional hidden states. The bias terms $b_{\overrightarrow{z}}$ and $b_{\overleftarrow{z}}$ correspond to the update gates, and b_h is the bias term for the candidate hidden state.

Table 5.3: Development and test CCC scores of unidirectional and bidirectional recurrent layers in the Unimodal Temporal Modeling baseline, with fixed GRU recurrent layer type and a single recurrent layer, while exploring various auxiliary network and pooling method combinations. The bold values indicate the best scores.

Auxiliary Network	**Pooling Method**	**Recurrent Layer Direction Type**	**Dev CCC**	**Test CCC**
Abstract	Last	Unidirectional	0.627	0.654
		Bidirectional	0.602	0.404
	Max	Unidirectional	0.575	0.527
		Bidirectional	0.590	0.420
Emotion	Last	Unidirectional	0.596	0.620
		Bidirectional	0.625	0.550
	Max	Unidirectional	**0.650**	**0.675**
		Bidirectional	0.626	0.604
Sentiment	Last	Unidirectional	0.441	0.415
		Bidirectional	0.462	0.355
	Max	Unidirectional	0.466	0.411
		Bidirectional	0.470	0.447
Personality	Last	Unidirectional	0.616	0.553
		Bidirectional	0.479	0.398
	Max	Unidirectional	0.641	0.556
		Bidirectional	0.503	0.453

Contrary to our expectations, the integration of bidirectional GRU does not consistently lead to improved performance across all configurations. While bidirectional modeling exhibits potential in certain instances, it does not yield consistently higher CCC scores compared to the unidirectional GRU configurations. This observation is intriguing, as the bidirectional mechanism intuitively offers access to a broader temporal context. The results indicate that bidirectional modeling might not inherently contribute to better performance. This could be attributed to the deficiency of computational complexity and the nature of the temporal patterns they encapsulate.

In our ongoing endeavor to comprehensively explore the impact of temporal modeling components, we conduct yet another crucial assessment. This time, our focus shifts to the existence of a transformer block, specifically targeting text representations as a precursor to temporal modeling. The primary objective is to understand the influence of the transformer block while maintaining consistency by utilizing a single-layer unidirectional GRU configuration. Since the results in Table 5.3 indicate that employing bidirectional GRU is not consistently superior to the unidirectional GRU, we persevere with the unidirectional approach to avoid unnecessary computational complexity. The results are presented in Table 5.4.

Table 5.4: Development and test CCC scores comparing the inclusion and exclusion of the Transformer block in the Unimodal Temporal Modeling baseline, with a fixed GRU recurrent layer type, a single recurrent layer, unidirectional recurrent layer direction, and max-pooling, while exploring various auxiliary network configurations. The bold values indicate the best scores.

Auxiliary Network	Transformer Block	Dev CCC	Test CCC
Abstract	✓	0.575	0.527
	✗	0.565	0.543
Emotion	✓	**0.650**	**0.675**
	✗	0.519	0.440
Sentiment	✓	0.466	0.411
	✗	0.390	0.362
Personality	✓	0.641	0.556
	✗	0.335	0.234

The outcomes of this assessment yield compelling insights into the significance of incorporating the transformer block. The existence of the transformer block consistently leads to a significant improvement in the CCC scores across all modalities. This marked improvement suggests that the transformer's attention mechanisms excel not only in capturing interdependencies within the input data but also in distilling crucial temporal features. The observed performance boost aligns with the transformer's inherent strength in capturing long-range dependencies and contextual nuances. By incorporating the transformer block, the model effectively enhances the text representations before temporal modeling, enabling the subsequent layers to operate on more informative and contextually enriched inputs. This corroborates the effectiveness of the attention mechanisms in capturing intricate temporal dynamics present in the data.

5.2 Evaluations in the Multimodal Setting

In this section, we present the assessments conducted for our main multimodal framework. We divide this section into four distinct subgroups, encompassing the assessments of various combinations of auxiliary networks, the existence of temporal modeling, the influence of [REG] token, and the fusion approach.

5.2.1 Assessment of Various Combinations of Auxiliary Networks

We evaluate the impact of combining different auxiliary networks on depression severity estimation performance. We experiment with various combinations of the auxiliary networks, and the results are presented in Table 5.5. While executing these experiments, we stick to our main architecture, including the MMJCA-Fusion module.

The results presented in Table 5.5 demonstrate a coherent alignment with the findings in Table 5.1. Specifically, combinations involving the abstract and emotion networks exhibit superior performance. Despite the sentiment network's

Table 5.5: Development and test CCC scores when different combinations of auxiliary networks are utilized. The bold values indicate the best scores.

Auxiliary Networks				**Dev CCC**	**Test CCC**
Personality	**Emotion**	**Senitment**	**Abstract**		
		✓	✓	0.632	0.640
	✓		✓	0.679	0.677
	✓	✓		0.644	0.633
	✓	✓	✓	0.660	0.681
✓			✓	0.641	0662
✓		✓		0.586	0.612
✓	✓			0.627	0.638
✓		✓	✓	0.638	0.666
✓	✓		✓	0.675	0.652
✓	✓	✓		0.648	0.665
✓	✓	✓	✓	**0.690**	**0.748**

tendency to diminish performance in certain instances, the amalgamation of multiple auxiliary networks consistently enhances depression severity estimation compared to the utilization of individual auxiliary networks in general. Notably, the fusion of all auxiliary networks yields the most favorable performance outcome, indicating that the integration of diverse representations substantially augments the model's capacity to discern intricate patterns associated with depression. This implies that the combined utilization of complementary information sources significantly enriches the model's grasp of the multifaceted aspects inherent in depression severity estimation.

5.2.2 Assessment of Temporal Modeling in the Multimodal Setting

In this section, we investigate the impact of temporal modeling on the performance of the depression severity estimation network. The primary objective is to determine whether the incorporation of temporal information through different modeling techniques enhances or hinders the multimodal framework's ability to predict PHQ-8 scores from transcripts. Two distinct approaches for temporal modeling are explored: (1) utilizing a single unidirectional GRU layer with max-pooling and last-pooling, and (2) adding learnable positional embeddings.

In the first approach, we delve into investigating the influence of introducing temporal modeling in our multimodal framework. The introduced temporal modeling mechanism is applied after the MMJCA-Fusion module, ensuring that both modalities' fused information is effectively utilized. We integrate the temporal modeling module subsequent to the MMJCA-Fusion module, aiming to capture the sequential information from the fused representations. The underlying premise is that incorporating a separate temporal modeling module for each modality before the fusion would lead to a substantial increase in model complexity. Akin to Section 5.1.2, we insert a pooling layer after the temporal modeling module. The high-level architecture of the method modified with the addition of temporal modeling is visualized in Figure 5.3.

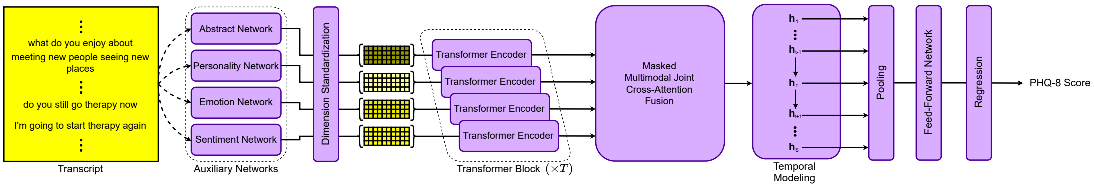

Figure 5.3: High-level architecture of the method modified with the addition of temporal modeling. The dashed arrows indicate that there is no back-propagation in those connections. h_t represents the hidden state for the t-th time step in the temporal modeling module.

Following the indications of the results obtained in Section 5.1.2, we use unidirectional GRU inside the temporal modeling module. Akin to the experiments conducted in Section 5.1.2, we employ a single layer and two cascaded layers of unidirectional GRU. Further, we employ two pooling methods: max-pooling and last-pooling to extract relevant information from the GRU output. The former captures the most salient features across the temporal dimension, while the latter focuses on the final time step's representation, assuming it carries the most critical information.

Since the multimodal framework has significantly more complex architecture compared to the unimodal framework, we introduce the second approach. This alternative approach represents a lighter temporal modeling method that necessitates notably lower computational resources in contrast to the first approach. In this approach, we augment the baseline model by incorporating learnable positional embeddings. These embeddings aim to provide the model with additional positional information, which might help in discerning the order of the sentences within the transcript during prediction. For each modality m, the positional embeddings matrix $P_m \in \mathbf{R}^{S \times D}$ are initialized from $N(\mu = 0, \sigma = 1)$. We inject them into the network prior to the transformer block. With the integration of these embeddings, the input of the transformer block for the modality m transforms into $X_{m,t=1} \bigoplus P_m$, where $\bigoplus$ denotes the position-wise addition.

We present the results of our experiments in Table 5.6, comparing the performance of each temporal modeling approach against the baseline model.

Table 5.6: Development and test CCC scores of different temporal modeling approaches. The bold values indicate the best scores.

Temporal Modeling	**Pooling Method**	**Dev CCC**	**Test CCC**
GRU (single layer)	Last	0.616	0.621
	Max	0.670	0.669
GRU (two cascaded layers)	Last	0.623	0.670
	Max	0.625	0.700
Positional Embedding	[REG] Token	0.649	0.574
None (Our Best)	[REG] Token	**0.690**	**0.748**

The results in Table 5.6 indicate that not employing any temporal modeling outperforms any other temporal modeling approach. The potential reasons are (1) data characteristics and (2) trade-offs between dataset limitations and model complexity. First of all, the transcripts may not have strong temporal dependencies or sequential patterns that can be effectively captured by temporal modeling techniques. In cases where the information relevant to depression severity estimation is mostly contained within individual sentences or short segments, temporal modeling might introduce noise and unnecessary complexity, leading to suboptimal performance. Secondly, the addition of temporal modeling can significantly increase the model's capacity, making it more prone to overfitting, especially considering the fact that our dataset is limited. In contrast, the model without temporal modeling is simpler and less susceptible to overfitting, leading to better generalization and overall performance. Furthermore, the transcripts contain various linguistic noises, hesitations, or repetitions. Temporal modeling methods inadvertently emphasize these noisy elements, leading to a negative impact on performance. In short, these outcomes can be summarized as the incorporation of temporal modeling acquiesces to the dataset limitations. Drawing from these potential reasons and the significant difference between the results, we deem it superfluous to advance to temporal modeling experiments within the context of the multimodal framework, given that the incorporation of temporal modeling techniques conspicuously diminishes performance.

5.2.3 Assessment of Regression Token Approach

In this section, we explore and analyze the impact of different pooling methods in the network. The original model utilizes the [REG] token approach for pooling contextual information from the fused representations. We conducted experiments to compare the performance of this [REG] token approach with two alternative pooling methods: max-pooling and mean-pooling. The experiment results are represented in Table 5.7.

Table 5.7: Development and test CCC scores of different pooling methods. The bold values indicate the best scores.

Pooling Method	Dev CCC	Test CCC
Mean	0.636	0.711
Max	0.646	0.693
[REG] Token	**0.690**	**0.748**

The results in Table 5.7 indicate that the original [REG] token approach achieves the highest performance among the three pooling methods. The model is able to effectively capture and summarize the contextual information from the interview transcripts, leading to superior predictions of PHQ-8 scores. The max-pooling approach, which selects the maximum value from each dimension across the token representations, exhibits significantly lower performance compared to the [REG] token approach. While max-pooling is a simple and efficient method, it seems to be not ideal for this specific task, as it tends to focus on the most salient features while potentially discarding relevant context. Similarly, the mean-pooling approach, which calculates the average of the token representations, demonstrates lower performance compared to the [REG] token approach. Mean-pooling may not adequately capture the nuanced patterns and interactions present in the transcripts, resulting in less accurate predictions.

In summary, the superiority of the [REG] token approach over max-pooling and mean-pooling is attributed to its ability to leverage the entire context of the transcripts for prediction. The [REG] token carries aggregated information from the transformer's attention mechanism, allowing it to encapsulate the most pertinent information for the task at hand. In contrast, max-pooling and mean-pooling may fail to preserve important contexts, leading to suboptimal performance.

5.2.4 Assessment of Fusion Approach

We evaluate the performance of the proposed MMJC-Fusion module in conjunction with the transformer block by comparing its results with two existing fusion approaches from other studies, as well as the basic concatenation method. The goal is to demonstrate the effectiveness and superiority of the proposed approach. In addition, we experiment with each fusion approach with and without the transformer block in order to assess its impact effectively.

For comparison, we include two existing fusion approaches relying on joint cross-attention, each proposed for emotion recognition tasks involving different modalities:

- Praveen et al. [67] propose a fusion approach designed for emotion recognition tasks that involve fusing auditory and visual modalities. It operates on the feature dimension, attending to specific features extracted from each modality to capture their complementary information for emotion prediction.

- Zhang et al. [68] propose a fusion approach designed for emotion recognition tasks that involve fusing textual, auditory, and visual modalities. Unlike our MMJC-Fusion approach that attends to tokens, it operates on the modality dimension, allowing the network to focus on the most informative modality during the emotion prediction task.

We present the results of the experiments in Table 5.8.

Table 5.8: Development and test CCC scores of different fusion approaches. The bold values indicate the best scores.

Fusion Approach	Transformer Block	Dev CCC	Test CCC
Concatenation	✗	0.000	0.000
	✓	0.650	0.654
Praveen et al. [67]	✗	0.376	0.349
	✓	0.641	0.680
Zhang et al. [68]	✗	0.000	0.000
	✓	0.650	0.676
MMJCA-Fusion	✗	0.570	0.669
	✓	**0.690**	**0.748**

As shown in the results, the combination of the transformer block and the MMJC-Fusion module outperforms all other fusion approaches with or without the Transformer. The superior performance of this combined approach is attributed to its ability to effectively capture long-range intra- and inter-level dependencies, contextual relationships, and fine-grained token-level information from auxiliary networks.

The fusion approaches proposed by Praveen et al. and Zhang et al. were originally developed for emotion recognition tasks that involve different modalities. However, in the context of depression severity estimation from transcripts, the MMJC-Fusion approach, which operates along the token dimension, demonstrates better performance compared to attention along the feature dimension (FD-Attention) and the modality dimension (MD-Attention).

The joint cross-attention mechanism along the token dimension in our MMJC-Fusion approach enables the depression severity estimation network to attend to specific sentences and phrases within the clinical interview transcripts. Therefore, the network gains a deeper understanding of the text data and can capture the nuanced linguistic patterns and contextual cues indicative of depression symptoms. This token-level attention allows the network to effectively integrate information from different personality traits, emotion, and sentiment representations, leading to superior predictions of PHQ-8 scores. In contrast, the joint cross-attention along the feature dimension and the modality dimension do not fully exploit the fine-grained information present in the text data. Praveen et al. emphasize specific features extracted from each modality, which might miss the context and relationships between tokens. Similarly, Zhang et al. focus on modalities, potentially overlooking the importance of specific sentences or phrases that carry critical information about a patient's mental state.

Furthermore, the positive impact of masking in the MMJC-Fusion approach reinforces the effectiveness of the token-level attention and the careful handling of text data during the fusion process. By considering only meaningful tokens and filtering out padding tokens, the depression severity estimation network can fully utilize the valuable information present in the clinical interview transcripts, resulting in a more accurate and interpretable prediction of PHQ-8 scores. In contrast, the fusion approaches without this masking process might inadvertently allocate attention to padding tokens, which may hinder the model's ability to focus on the critical content of the text. This results in suboptimal predictions and lower scores.

5.3 Segmented Evaluation of The Method Across Different Ranges of True Values

In this section, we present a detailed analysis of our proposed method's performance across different ranges of true PHQ-8 values. The purpose of this analysis is to examine whether the predictive accuracy of our method remains consistent across the entire spectrum of depression severity, despite the class imbalance present in the dataset. To achieve this, we partition the validation and test sets into distinct groups based on the true PHQ-8 scores. Each group encompasses a specific range of PHQ-8 values, allowing us to investigate how our method's prediction errors are distributed within these partitions.

For each evaluation on both the validation and test sets, we segment the samples into distinct groups according to the following PHQ-8 score ranges: [0,4], [5,9], [10,14], [15,19], and [20,24]. We compute the RMSE and MAE for each group, providing us with insights into the accuracy of our method's predictions within different ranges of depression severity. By analyzing the trends in RMSE and MAE across these groups, we aim to gain a comprehensive understanding of our method's behavior across the entire depression severity spectrum.

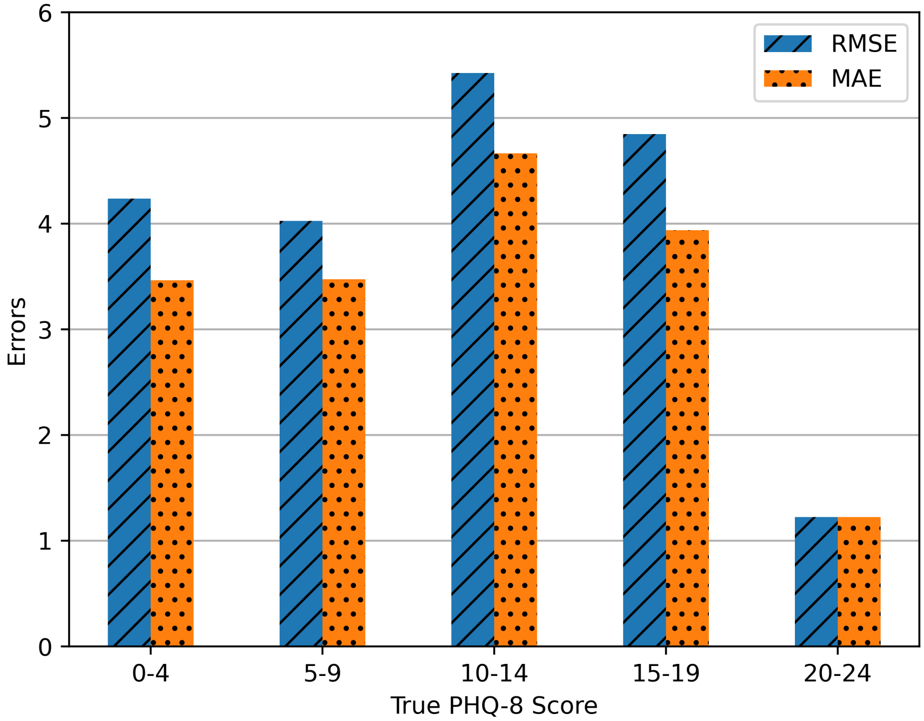

Figure 5.4: RMSE and MAE values across different groups in the validation set, each encompassing a specific range of PHQ-8 scores.

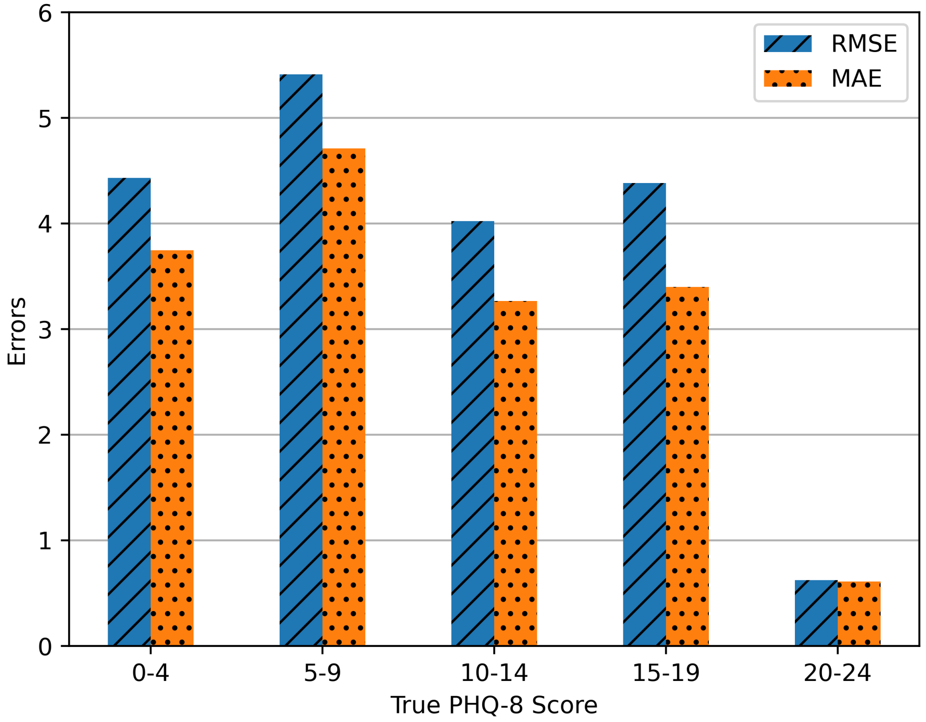

Figure 5.5: RMSE and MAE values across different groups in the test set, each encompassing a specific range of PHQ-8 scores.

The results of our segmented evaluation, shown in Figures 5.4 and 5.5, shed light on the robustness of our proposed method. Across both the validation and test sets, we observe consistent patterns in the distribution of prediction errors. Notably, the RMSE and MAE values demonstrate remarkably similar trends for the groups encompassing the [0,4], [5,9], [10,14], and [15,19] PHQ-8 score ranges. This finding suggests that our method's accuracy remains stable across a wide range of depression severity, indicating its resilience against potential bias arising from class imbalance of the AVEC'19 dataset.

Furthermore, a noteworthy observation emerges from the evaluation of the [20,24] PHQ-8 score range. In both the validation and test sets, this group exhibits substantially lower RMSE and MAE values compared to the other groups. This phenomenon indicates that our method excels in predicting cases of higher depression severity, showcasing its potential clinical relevance in accurately identifying individuals with a higher tendency towards depression.

Our method's ability to maintain consistent performance across varying levels of depression severity is a testament to its generalizability and reliability. The fact that it performs well even in the presence of imbalanced class distribution demonstrates its capacity to capture the intricate relationships between input features and depression severity, without succumbing to undue influence from the prevalence of lower PHQ-8 scores.

5.4 Assessment of Attention Weights of the MMJCA-Fusion Module on the Test Set

In the pursuit of enhancing the accuracy and interpretability of the proposed method, we delve into an essential aspect of our proposed architecture – the MMJCA-Fusion module. This section focuses on a comprehensive assessment of the attention weights generated by this module when applied to transcripts from the clinical interview test set. Understanding the distribution and significance of attention across sentences within a transcript is pivotal in comprehending the model's decision-making process.

Our novel architecture aims to harness multiple text modalities, incorporating emotional, sentiment, and personality trait features extracted from auxiliary fine-tuned networks for each sentence in the transcripts. The MMJCA-Fusion module, a critical component of our framework, facilitates the fusion of these diverse modalities while assigning attention weights to each modality during feed-forward propagation. This enables the model to dynamically adapt to the most informative aspects of the input data. Throughout this evaluation process, we extract the attention weights associated with the [REG] token for each sentence in a transcript. Since the [REG] token encapsulates the captured information for each sentence, these attention weights reflect the importance or significance of the corresponding sentences in the context of the entire transcript. To obtain a scalar attention weight score for each sentence, we calculate the mean of these attention weights across all modalities. This scalar score encapsulates the relative importance of each sentence within the transcript, as determined by the model's attention mechanism.

In this section, we embark on an exploratory journey into the attention weights assigned by the MMJCA-Fusion module under various scenarios. Four distinct scenarios will be analyzed, each shedding light on the model's behavior in different situations:

- **High True PHQ-8 Score with Accurate Prediction:** This scenario explores a specific transcript with a high true PHQ-8 score, where our model accurately predicts the corresponding depression severity level. By examining the sentences that garner the highest attention, we aim to decipher what characteristics the model identifies as indicative of depression in these cases.
- **High True PHQ-8 Score with Inaccurate Prediction:** Here, we investigate a case where the true PHQ-8 score is high, but the model's prediction deviates from accuracy. The analysis of attention weights in such an instance can unveil potential challenges or shortcomings in our architecture when handling severe depression cases.
- **Low True PHQ-8 Score with Accurate Prediction:** In this scenario, we focus on a specific transcript with a low true PHQ-8 score, which is correctly identified by our model. Understanding which sentences receive heightened attention during prediction can provide insights into the model's ability to recognize non-depressive attributes within the text.
- **Low True PHQ-8 Score with Inaccurate Prediction:** Lastly, we delve into a case where a specific transcript exhibits a low true PHQ-8 score, but our model's prediction is inaccurate. Analyzing the sentences prioritized by the model's attention can uncover challenges in distinguishing low-level depressive symptoms from other factors.

By scrutinizing the attention weights attributed to individual sentences in these scenarios, we aim to enhance our understanding of the model's decision-making process, gain insights into the salient textual cues it leverages, and identify areas for potential improvement. This analysis not only contributes to the interpretability of our automatic depression severity estimation system but also offers valuable feedback for refining its performance and robustness. The results of the scenarios are presented in Tables 5.9, 5.10, 5.11, and 5.12.

Table 5.9: The top ten sentences that have achieved the highest average attention weıghts for the high true PHQ-8 score with accurate prediction scenario(True PHQ-8 Score: 20, Predicted PHQ-8 Score: 20.49).

Sentence	**Average Attention Weight** ($\times 10^2$)
how easy is it for you to get a good night sleep not very	1.804
because I knew I was going to kill myself if I didn't	1.776
and it wasn't easy at all especially during that time	1.687
over at a friend's house I really wanted to leave felt it was rude to leave so early so I stuck it out for 6 hours and I just was probably the worst guest	1.586
depressed mostly	1.578
not very good	1.569
no I'm even more irritable I have even more of a flash camper it makes depression worse anxiety worse	1.551
I want to take a boat load of medications now treatment for PTSD it's also for depression	1.435
due to PTSD	1.342
I'm not a therapist	1.323

Table 5.10: The top ten sentences that have achieved the highest average attention weıghts for the high true PHQ-8 score with inaccurate prediction scenario (True PHQ-8 Score: 19, Predicted PHQ-8 Score: 9.74).

Sentence	**Average Attention Weight** ($\times 10^2$)
yes I'm a little bit more upbeat	2.688
I'm very good at controlling my temper	2.598
I went to a ladies luncheon and I enjoyed myself	2.516
I went to the movies and saw Monsters University	2.490
how would your best friend describe	2.429
my home is filled with a lot of negative energy and I don't have any friends to rely on	2.422
the shopping the museums	2.372
I recently got involved in a book club	2.369
I went to Walt Disney World for my 21st birthday	2.366
I read a book	2.276

Table 5.11: The top ten sentences that have achieved the highest average attention weights for the low true PHQ-8 score with accurate prediction scenario (True PHQ-8 Score: 0, Predicted PHQ-8 Score: 1.61).

Sentence	**Average Attention Weight** ($\times 10^2$)
I think it's a very useful tool and many people have been successful in receiving therapy	1.668
how easy is it for you to get a good night's sleep very easy I usually retire about 11:30 12:00 at night and sleep through for 6 to 7 hours	1.482
you feel therapy is useful I think it is for some people yes probably for most	1.342
and it's usually as a result of being ill if I have the flu or cold or whatever then I'm sluggish the next day or for two days but I'm usually able to bounce back	1.312
the 18th of July and then they'll be out on the Great Lawn on August 3rd so it'll be the August 3rd event and crate they're bringing a trailer ring in BBQ for everybody and it's going to be a wonderful time we have anything will send it in or you're more than happy and I okay at night I don't think I do but I met you sent me an email and I noticed you have the military would have to understand it from West Point okay and they would I mean that we're going to have generals were going to be on during World War two people they might love to come oh yeah	1.284
I got a feeling good I've been very fortunate since having open heart surgery I've been relatively healthy and I'm very involved in the community and I enjoy every every aspect of that sometimes it's very tiring but it's a good tired	1.269
many of the veterans with whom I live in work I have had extensive therapy and it's proven to be very successful	1.227
visiting SEC liqueur in that my Mart in Paris on my first day in France in Europe	1.206
I always having a full plate	1.203
we didn't have any virtual people but it was nice	1.201

Table 5.12: The top ten sentences that have achieved the highest average attention weights for the low true PHQ-8 score with accurate prediction scenario (True PHQ-8 Score: 0, Predicted PHQ-8 Score: 9.12).

Sentence	**Average Attention Weight** ($\times 10^2$)
I hardly ever not sleep or get sleep	3.399
not too easy	2.740
Come Easy	2.580
I'm happy to be alive	2.313
[REG]	2.215
I need a kiss to my head together	2.210
tell me about the hardest.	2.198
try to help her out	2.121
it's Friday or is it	2.088
it's simple knowing is an A+ affect my life	2.070

5.4.1 Discussion on the High True PHQ-8 Score with Accurate Prediction Scenario

In this scenario, the model accurately predicts a high PHQ-8 score, indicating severe depressive symptoms. The sentences that received the highest attention weights predominantly revolve around themes associated with depression:

- Sentences like ”how easy is it for you to get a good night’s sleep not very” and ”and it wasn’t easy at all especially during that time” highlight sleep disturbances, a common symptom of depression [86, 87].
- ”because I knew I was going to kill myself if I didn’t” is a particularly alarming statement, signifying a high risk of self-harm or suicide, which aligns with a high PHQ-8 score [88, 89].
- ”depressed mostly,” ”not very good,” and ”it makes depression worse anxiety worse” directly mention feelings of depression and anxiety, reinforcing the severity of the condition.
- The mention of ”treatment for PTSD” and ”due to PTSD” indicates the presence of comorbid conditions, which can contribute to a higher PHQ-8 score.

The attention weights in this scenario reflect the model’s ability to appropriately identify and prioritize sentences indicative of severe depression, leading to an accurate prediction.

5.4.2 Discussion on the High True PHQ-8 Score with Inaccurate Prediction Scenario

In this case, despite the high true PHQ-8 score, the model’s prediction is inaccurate. The sentences receiving the highest attention weights seem to focus on positive or neutral aspects of the individual’s life:

- Sentences such as ”I’m very good at controlling my temper,” ”I went to a ladies luncheon and I enjoyed myself,” and ”I went to the movies and saw Monsters University” convey positive experiences and emotional stability, which may have influenced the model’s prediction.

- Mentions of engaging in activities like ”the shopping,” ”the museums,” and ”a book club” indicate an active and socially connected lifestyle, potentially leading the model to underestimate the depression severity as the plenitude of the social activities indicates a lower risk of depression [90, 91].

- ”I recently got involved in a book club” and ”I read a book” highlight engagement in intellectually stimulating activities, suggesting a positive mental state.

The attention weights in this scenario suggest that the model might have been overly influenced by the presence of positive or neutral cues in the text, leading to an inaccurate prediction despite the high true PHQ-8 score.

5.4.3 Discussion on the Low True PHQ-8 Score with Accurate Prediction Scenario

In this situation, the model accurately predicts a low PHQ-8 score, reflecting the absence or mild nature of depressive symptoms. The sentences with the highest attention weights emphasize positive attributes and well-being:

- Sentences like ”I think it’s a very useful tool,” ”how easy is it for you to get a good night’s sleep very easy,” and ”I’ve been very fortunate since having open heart surgery” underscore a positive outlook on life, good sleep quality, and overall well-being.

- Mentions of engaging in activities, attending events, and enjoying social interactions (”they’re bringing a trailer ring in BBQ,” ”I’m very involved in the community”) indicate an active and socially connected lifestyle.

- The statement "many of the veterans with whom I live in work I have had extensive therapy and it's proven to be very successful" suggests a support network and successful therapeutic interventions, contributing to the low PHQ-8 score.

The attention weights in this scenario align with the model's accurate prediction, highlighting the absence of significant depressive cues and the presence of positive indicators.

5.4.4 Discussion on the Low True PHQ-8 Score with Inaccurate Prediction Scenario

In this scenario, the model's prediction is notably inaccurate, given the low true PHQ-8 score. The sentences with the highest attention weights do not strongly indicate depressive symptoms. However, it's essential to highlight why this inaccurate prediction occurred:

- Sentences like "I hardly ever not sleep or get sleep" and "not too easy" do mention sleep difficulties and possible emotional distress, but they do not strongly suggest severe depression. The model might have placed undue emphasis on these mild cues, leading to an inaccurate prediction.
- It is essential to note that the model's prediction, though elevated compared to the low true PHQ-8 score, does not reach the upper range of possible PHQ-8 scores. Phrases like "I'm happy to be alive" and "it's simple knowing is an A+ affect my life" convey a positive attitude and optimism. These positive expressions may have influenced the model's decision positively, but the overall prediction remains within a moderate range.
- The mention of "try to help her out" and "it's Friday or is it" does not provide clear evidence of depression. However, the model might have mistakenly interpreted these sentences as neutral statements, which could have contributed to the model's erroneous high prediction.

In this case, the model's attention weights seem to have been swayed by sentences with mild emotional cues, potentially leading to the inaccurate prediction. This highlights the challenge of distinguishing between mild depressive symptoms and non-depressive cues, a crucial area for future model improvement.

5.5 Comparison to Other Methods

In this section, we present a comprehensive comparison of the performance of the proposed method to existing state-of-the-art methods. Akin to ours, these methods address the depression severity estimation task on the AVEC'19 dataset. This comparison allows us to place our results in the broader context of the existing research landscape, providing insights into the strengths and limitations of our approach. By benchmarking our performance against other cutting-edge methods, we can ascertain the competitiveness of our model and its potential to outperform or align with the best-performing techniques in the field. The practice of comparing our outcomes with state-of-the-art studies promotes transparency and encourages rigorous evaluation, thereby enhancing the reliability of our research. It showcases the significance of independent validation and strengthens the credibility of our contributions to the scientific community.

Table 5.13 presents a meticulous comparison between our text-based depression severity estimation network and prominent state-of-the-art methodologies. A distinguishing factor is our exclusive reliance on the text modality, while other methods combine multiple modalities such as textual, auditory, and visual inputs. The results decisively highlight the supremacy of our approach despite the scarcity of the utilized modalities.

Table 5.13: Results of comparison to other methods. Bold values indicate the best results for the corresponding evaluation metric. The methods are sorted by year.

Method	**Year**	**Modalities**			**Dev**			**Test**		
		Textual	**Audial**	**Visual**	**CCC**	**RMSE**	**MAE**	**CCC**	**RMSE**	**MAE**
Zhang et al. [38]	2019	✓	✓		-	-	-	-	6.78	5.77
Ray et al. [51]	2019	✓	✓	✓	-	4.37	-	0.670	4.73	4.02
Makiuchi et al. [52]	2019	✓	✓	✓	0.696	3.86	-	0.403	6.11	-
Fan et al. [40]	2019	✓	✓		0.466	5.07	4.06	0.430	5.91	4.39
Sun et al. [6]	2021		✓	✓	**0.733**	**3.78**	-	-	-	-
Rutowski et al. [24]	2021	✓			-	-	-	-	5.51	4.20
Van Steijn et al. [43]	2022	✓	✓		0.61	5.10	-	0.62	6.06	-
Saggu et al. [56]	2022	✓	✓	✓	0.662	4.32	-	0.457	5.36	-
Ours	2023	✓			0.690	4.32	**3.62**	**0.748**	**4.37**	**3.46**

Chapter 6

Conclusion

In this thesis, we have presented a comprehensive approach to automatic depression severity estimation through the analysis of clinical interview transcripts. Our contributions have advanced the field by proposing a novel architecture that effectively utilizes multiple text modalities, including emotion, sentiment, and personality. Through our research, we have demonstrated the effectiveness of these features in a single, integrated framework for the automatic depression severity estimation task. To derive motion, sentiment, and personality features, we have incorporated various auxiliary networks that are LLMs fine-tuned for the corresponding classification tasks. Our pioneering use of these multiple features within a novel transformer-based approach represents a significant step forward in the development of accurate and comprehensive depression severity estimation systems. By leveraging these diverse aspects of textual information, our model captures a more holistic representation of an individual's mental state, enhancing the predictive capabilities and robustness of the system.

Furthermore, we have introduced a new multimodal joint cross-attention fusion technique (MMJCA-Fusion) that effectively combines information from different text modalities. This technique not only improves the interpretability of our model but also ensures that relevant features are adequately integrated, leading to enhanced predictive performance.

An essential aspect of our work involves the fine-tuning of the pretrained RoBERTa model, as one of the auxiliary networks, for the classification of personality traits in utterances. This auxiliary network not only contributes to the accuracy of our main approach but also underscores the potential for leveraging pre-existing models to augment specific tasks within a larger framework.

We have assessed the proposed method with unimodal and multimodal settings. During the evaluations with unimodal settings, we have examined the impact of each auxiliary network on PHQ-8 score estimation. To convert the original architecture to an unimodal one, we have excluded the MMJCA-Fusion module. The first assessment with the unimodal setting has been executed by employing each auxiliary network individually. The results have indicated that each network has contributed valuable information to depression severity estimation, whereas the sentiment network has achieved significantly lower results than others. Later, we have assessed the impact of adding temporal modeling into the unimodal network. During this assessment, we have conducted experiments across different configurations of auxiliary networks, recurrent layer types, the number of recurrent layers, and pooling methods. We have considered using two types of recurrent layers: LSTM and GRU, using a single layer and two cascaded recurrent layers, and last-pooling and max-pooling methods. One observation that has been obtained from the results is that using temporal modeling has slightly improved the results in the unimodal setting. So, we have proceeded with the assessments of temporal modeling. Another observation has been that the configurations, including LSTM and two cascaded recurrent layers, have not consistently improved the model performance. To avoid an unnecessary increase in model complexity, we have taken account of using a single GRU layer in the proceeding assessments of temporal modeling in the unimodal setting. Afterward, we have explored the influence of employing bidirectional recurrent layers instead of unidirectional ones. For each auxiliary network, we have conducted experiments for utilizing a single bidirectional GRU layer with both max-pooling and last-pooling methods. Contrary to our expectations, the integration of bidirectional GRU has not consistently led to improved performance across the configurations. In this manner, we have stuck with utilizing a single unidirectional GRU layer in the subsequent evaluation, that is the assessment of the existence of the transformer block. In this assessment, we have both included the transformer block in and excluded it from the network for each auxiliary network. The findings have demonstrated the importance of the transformer block in our architecture as including it has consistently led to a significant improvement in the results.

For the multimodal evaluation, we have first conducted experiments utilizing all combinations of the auxiliary networks. The results have indicated that combining multiple auxiliary networks has consistently improved the PHQ-8 estimation performance and the combination of all the auxiliary networks has yielded the best results. Similar to the unimodal evaluation, we have included temporal modeling in the multimodal setting for the subsequent assessment. In this assessment, we have utilized a single GRU layer and two cascaded layers GRU layers with both last-pooling and max-pooling. In addition, we have experimented with a positional embedding approach. The findings have shown that the utilization of any temporal modeling has not managed to improve the performance. Further, we have conducted assessments for different pooling methods to explore the impact of the [REG] token, and for different fusion approaches from other studies to explore the impact of the MMJCA-Fusion approach. The findings have demonstrated the original proposed architecture has outperformed other configurations.

Moreover, we have performed a segmented evaluation of the proposed method across different ranges of true PHQ-8 scores on both the validation set and the test set. The findings have demonstrated that our method's accuracy remains stable across a wide range of depression severity, indicating its resilience against potential bias arising from class imbalance of the AVEC'19 dataset. We have also observed from the results that the [20,24] PHQ-8 score range has exhibited significantly lower error values compared to the other ranges. Subsequently, we have explored the attention weights generated by the MMJCA-Fusion module. We have examined the average attention weights assigned for the sentences that exist in selected samples from the test set. The outcomes have provided valuable insights into the correlations between the observed textual cues and the depression severity. The outcomes have also matched existing psychological studies. Finally, we have pursued a comprehensive comparison of the performance of the proposed method to other methods that also address the depression severity estimation task on the AVEC'19 dataset. A notable factor in this comparison is that other studies exploit the different combinations of text, audio, and vision modalities. The results have highlighted the supremacy of our method.

The remarkable attainment of our text-based depression severity estimation network deserves profound recognition. By exclusively focusing on textual data, our approach circumvents the intricacies involved in integrating and processing multimodality. This underscores the latent potential residing within linguistic constructs present within the textual content for discerning and precise depression severity estimation. Furthermore, the singularity of text modality integration conveys pragmatic implications for real-world implementation. The streamlined utilization of text data not only mitigates resource and computational demands but also amplifies the method's practicality and seamless integration into existing mental health assessment frameworks.

Bibliography

[1] W. H. Organization, *Depression and Other Common Mental Disorders: Global Health Estimates*. World Health Organization, 2017.

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

[3] S. H. Dumpala, S. Rodriguez, S. Rempel, R. Uher, and S. Oore, "Significance of speaker embeddings and temporal context for depression detection," *CoRR*, vol. abs/2107.13969, 2021.

[4] Z. Huang, J. Epps, D. Joachim, B. Stasak, J. R. Williamson, and T. F. Quatieri, "Domain adaptation for enhancing speech-based depression detection in natural environmental conditions using dilated cnns," *Interspeech 2020*, 2020.

[5] W. C. de Melo, E. Granger, and M. B. López, "Mdn: A deep maximization-differentiation network for spatio-temporal depression detection," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, p. 578–590, 2023.

[6] H. Sun, J. Liu, S. Chai, Z. Qiu, L. Lin, X. Huang, and Y. Chen, "Multi-modal adaptive fusion transformer network for the estimation of depression level," *Sensors*, vol. 21, no. 14, p. 4764, 2021.

[7] Z. Dai, Q. Li, Y. Shang, and X. Wang, “Depression detection based on facial expression, audio and gait,” *2023 IEEE 6th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC)*, 2023.

[8] J. M. Havigerová, J. Haviger, D. Kučera, and P. Hoffmannová, “Text-based detection of the risk of depression,” *Frontiers in Psychology*, vol. 10, 2019.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[10] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pretraining,” *URL: https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/language-unsupervised/language_understanding_paper. pdf*, 2018.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized BERT pretraining approach,” *CoRR*, vol. abs/1907.11692, 2019.

[12] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, “Mpnet: Masked and permuted pre-training for language understanding,” in *NeurIPS 2020*, ACM, September 2020.

[13] D. Klein, R. Kotov, and S. Bufferd, “Personality and depression: Explanatory models and review of the evidence,” *Annual review of clinical psychology*, vol. 7, pp. 269–95, 04 2010.

[14] A. Compare, C. Zarbo, E. Shonin, W. V. Gordon, and C. Marconi, “Emotional regulation and depression: A potential mediator between heart and mind.,” *Cardiovascular Psychiatry and Neurology*, vol. 2014, pp. 324374–10, 2014.

[15] N. V. Babu and E. Kanaga, "Sentiment analysis in social media data for depression detection using artificial intelligence: A review," *SN Computer Science*, vol. 3, 01 2022.

[16] A. Mallol-Ragolta, Z. Zhao, L. Stappen, N. Cummins, and B. W. Schuller, "A hierarchical attention network-based approach for depression detection from transcribed clinical interviews," *Interspeech 2019*, 2019.

[17] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

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

[19] H. Dinkel, M. Wu, and K. Yu, "Text-based depression detection on sparse data," 2020.

[20] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL* (A. Moschitti, B. Pang, and W. Daelemans, eds.), pp. 1724–1734, ACL, 2014.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2013.

[22] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[23] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 2227–2237, 2018.

[24] T. Rutowski, E. Shriberg, A. Harati, Y. Lu, R. Oliveira, and P. Chlebek, "Cross-demographic portability of deep nlp-based depression models," *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021.

[25] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing LSTM language models," in *International Conference on Learning Representations*, 2018.

[26] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *ACL*, Association for Computational Linguistics, 2018.

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

[28] H. Zogan, I. Razzak, S. Jameel, and G. Xu, "Depressionnet: Learning multi-modalities with user post summarization for depression detection on social media," *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

[29] H. Zogan, I. Razzak, X. Wang, S. Jameel, and G. Xu, "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media," *World Wide Web*, vol. 25, no. 1, p. 281–304, 2022.

[30] T. Shen, J. Jia, G. Shen, F. Feng, X. He, H. Luan, J. Tang, T. Tiropanis, T.-S. Chua, and W. Hall, "Cross-domain depression detection via harvesting social media," *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018.

[31] C. Lin, P. Hu, H. Su, S. Li, J. Mei, J. Zhou, and H. Leung, "Sensemood: Depression detection on social media," *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 2020.

[32] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Bagher Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

[33] T. Gui, L. Zhu, Q. Zhang, M. Peng, X. Zhou, K. Ding, and Z. Chen, "Cooperative multimodal approach to depression detection in twitter," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, p. 110–117, 2019.

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

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[36] N. Ghadiri, R. Samani, and F. Shahrokh, "Integration of text and graph-based features for detecting mental health disorders from voice," 2022.

[37] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, (New York, NY, USA), p. 1459–1462, Association for Computing Machinery, 2010.

[38] L. Zhang, J. Driscol, X. Chen, and R. Hosseini Ghomi, "Evaluating acoustic and linguistic features of detecting depression sub-challenge dataset," *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019.

[39] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML’14, p. II–1188–II–1196, JMLR.org, 2014.

[40] W. Fan, Z. He, X. Xing, B. Cai, and W. Lu, “Multi-modality depression detection via multi-scale temporal dilated cnns,” *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019.

[41] E. Loper and S. Bird, “Nltk: The natural language toolkit,” in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP ’02, (USA), p. 63–70, Association for Computational Linguistics, 2002.

[42] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, “Recursive deep models for semantic compositionality over a sentiment treebank,” in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, (Seattle, Washington, USA), pp. 1631–1642, Association for Computational Linguistics, Oct. 2013.

[43] F. Van Steijn, G. Sogancioglu, and H. Kaya, “Text-based interpretable depression severity modeling via symptom predictions,” *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*, 2022.

[44] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *EMNLP/IJCNLP (1)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), pp. 3980–3990, Association for Computational Linguistics, 2019.

[45] J. Pennebaker, R. Boyd, K. Jordan, and K. Blackburn, *The development and psychometric properties of LIWC2015*. University of Texas at Austin, 2015.

[46] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, “Using linguistic cues for the automatic recognition of personality in conversation and text,” *J. Artif. Int. Res.*, vol. 30, p. 457–500, nov 2007.

[47] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "FLAIR: An easy-to-use framework for state-of-the-art NLP," in *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59, 2019.

[48] A. Pampouchidou, O. Simantiraki, A. Fazlollahi, M. Pediaditis, D. Manousos, A. Roniotis, G. Giannakakis, F. Meriaudeau, P. Simos, K. Marias, and et al., "Depression assessment by fusing high and low level features from audio, video, and text," *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016.

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

[50] B. Sun, Y. Zhang, J. He, L. Yu, Q. Xu, D. Li, and Z. Wang, "A random forest regression method with selected-text feature for depression assessment," *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017.

[51] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level attention network using text, audio and video for depression prediction," *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019.

[52] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019.

[53] W. Zheng, L. Yan, C. Gou, and F.-Y. Wang, "Graph attention model embedded with multi-modal knowledge for depression detection," *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020.

[54] P.-C. Wei, K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen, “Multi-modal depression estimation based onnbsp;sub-attentional fusion,” *Lecture Notes in Computer Science*, p. 623–639, 2023.

[55] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, “Attentional feature fusion,” *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[56] G. S. Saggu, K. Gupta, and K. V. Arya, “Depressnet: A multimodal hierarchical attention mechanism approach fordepression detection,” *International Journal of Engineering Sciences*, vol. 15, no. 1, 2022.

[57] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves, “Tweeteval: Unified benchmark and comparative evaluation for tweet classification,” *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

[58] J. Hartmann, M. Heitmann, C. Siebert, and C. Schamp, “More than a feeling: Accuracy and application of sentiment analysis,” *International Journal of Research in Marketing*, vol. 40, no. 1, p. 75–87, 2023.

[59] X. Sun, B. Liu, J. Cao, J. Luo, and X. Shen, “Who am i? personality detection based on deep learning for texts,” *2018 IEEE International Conference on Communications (ICC)*, 2018.

[60] F. Yang, X. Quan, Y. Yang, and J. Yu, “Multi-document transformer for personality detection,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, p. 14221–14229, 2021.

[61] V. Lynn, N. Balasubramanian, and H. A. Schwartz, “Hierarchical modeling for user personality prediction: The role of message-level attention,” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[62] E. Kerz, Y. Qiao, S. Zanwar, and D. Wiechmann, “Pushing on personality detection from verbal behavior: A transformer meets text contours of

psycholinguistic features," *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment amp;amp; Social Media Analysis*, 2022.

[63] T. Yang, F. Yang, H. Ouyang, and X. Quan, "Psycholinguistic tripartite graph network for personality detection," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021.

[64] B. Li, Y. Weng, Q. Song, F. Ma, B. Sun, and S. Li, "Prompt-based pretrained model for personality and interpersonal reactivity prediction," *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment amp;amp; Social Media Analysis*, 2022.

[65] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv: Learning*, 2016.

[66] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[67] R. G. Praveen, W. C. de Melo, N. Ullah, H. Aslam, O. Zeeshan, T. Denorme, M. Pedersoli, A. L. Koerich, S. Bacon, P. Cardinal, and et al., "A joint cross-attention model for audio-visual fusion in dimensional emotion recognition," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

[68] S. Zhang, R. An, Y. Ding, and C. Guan, "Continuous emotion recognition using visual-audio-linguistic information: A technical report for abaw3," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

[69] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L.-P. Morency, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 International*

Conference on Autonomous Agents and Multi-Agent Systems, AAMAS '14, (Richland, SC), p. 1061–1068, International Foundation for Autonomous Agents and Multiagent Systems, 2014.

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

[71] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, and et al., "Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition," *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019.

[72] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, (USA), p. 142–150, Association for Computational Linguistics, 2011.

[73] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, (USA), p. 115–124, Association for Computational Linguistics, 2005.

[74] P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, and T. Wilson, "SemEval-2013 task 2: Sentiment analysis in Twitter," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, (Atlanta, Georgia, USA), pp. 312–320, Association for Computational Linguistics, June 2013.

[75] S. D., K. L., and C. E., "Tweetgeist: Can the twitter timeline reveal the structure of broadcast events?," *Horizon, In CSCW 2010*, 2010.

[76] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, (Seattle, Washington, USA), pp. 1631–1642, Association for Computational Linguistics, Oct. 2013.

[77] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, (Prague, Czech Republic), pp. 440–447, Association for Computational Linguistics, June 2007.

[78] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: Understanding rating dimensions with review text," in *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, (New York, NY, USA), p. 165–172, Association for Computing Machinery, 2013.

[79] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 task 4: Sentiment analysis in Twitter," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, (Vancouver, Canada), pp. 502–518, Association for Computational Linguistics, Aug. 2017.

[80] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 task 1: Affect in tweets," in *Proceedings of the 12th International Workshop on Semantic Evaluation*, (New Orleans, Louisiana), pp. 1–17, Association for Computational Linguistics, June 2018.

[81] Y. Zhu, L. Hu, X. Ge, W. Peng, and B. Wu, "Contrastive graph transformer network for personality detection," *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022.

[82] I. B. Myers and K. C. Briggs, "The myers-briggs type indicator," *Educational and Psychological Measurement*, vol. 3, no. 3, pp. 229–237, 1943.

[83] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International*

Conference on Artificial Intelligence and Statistics (Y. W. Teh and M. Titterington, eds.), vol. 9 of *Proceedings of Machine Learning Research*, (Chia Laguna Resort, Sardinia, Italy), pp. 249–256, PMLR, 13–15 May 2010.

[84] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *International Conference on Learning Representations (ICLR)*, (San Diega, CA, USA), 2015.

[85] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[86] H. Fang, S. Tu, J. Sheng, and A. Shao, “Depression in sleep disturbance: a review on a bidirectional relationship, mechanisms and treatment,” *Journal of Cellular and Molecular Medicine*, vol. 23, no. 4, pp. 2324–2332, 2019.

[87] P. L. Franzen and D. J. Buysse, “Sleep disturbances and depression: risk relationships for subsequent depression and therapeutic implications,” *Dialogues in Clinical Neuroscience*, 2022.

[88] Y. Takahashi, “Depression and suicide,” *Japan Medical Association Journal*, vol. 44, no. 8, pp. 359–363, 2001.

[89] H. J. Jeon, “Depression and suicide,” *Journal of the Korean Medical Association*, vol. 54, no. 4, pp. 370–375, 2011.

[90] L. M. De Wit, M. Fokkema, A. van Straten, F. Lamers, P. Cuijpers, and B. W. Penninx, “Depressive and anxiety disorders and the association with obesity, physical, and social activities,” *Depression and Anxiety*, vol. 27, no. 11, pp. 1057–1065, 2010.

[91] K. Holtfreter, M. D. Reisig, and J. J. Turanovic, “Depression and infrequent participation in social activities among older adults: the moderating role of high-quality familial ties,” *Aging & Mental Health*, vol. 21, no. 4, pp. 379–388, 2017.