MODELING SPEECH TRANSCRIPTIONS FOR AUTOMATIC ASSESSMENT 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
Ergün Batuhan Kaynak
September 2022

Modeling Speech Transcriptions for Automatic Assessment of Depression Severity
By Ergün Batuhan Kaynak
September 2022

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)

Selim Aksoy

Albert Ali Salah

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

MODELING SPEECH TRANSCRIPTIONS FOR AUTOMATIC ASSESSMENT OF DEPRESSION SEVERITY

Ergün Batuhan Kaynak
M.S. in Computer Engineering
Advisor: Hamdi Dibeklioğlu
September 2022

It is true that everyone has bad days from time to time. Unfortunately, for people suffering from depression, every day is a constant battle for motivation to do even the simplest of things, all the while dealing with hopelessness, physical and emotional fatigue, and sadness. Considering the ever-increasing number of people suffering from this disease, the necessity for automated depression severity assessment systems is profound. These systems can be used in treatment procedures, and the findings provided from learned models can help us better understand the dynamics of depression.

To help in the solution to this illness, we propose a modular deep learning pipeline that uses speech transcripts as input for depression severity prediction. Through our pipeline, we investigate the role of popular deep learning architectures in creating representations for depression assessment. To extend the depression assessment literature on text modality, we provide a thorough analysis of sentence statistics and their effects on model training. We also present an investigation regarding the use of sentiment information for depression assessment.

Evaluation of the proposed architectures is performed on the publicly available Extended Distress Analysis Interview Corpus dataset (E-DAIC). Through the results and discussions, we show that informative representations for depression assessment can be obtained without exploiting the temporal dynamics between sentences. Our proposed non-temporal model outperforms the state of the art by %8.8 in terms of Concordance Correlation Coefficient (CCC). In light of our findings on trained models and data statistics, we discuss how recurrent structures can have a bias toward certain sequence lengths during training and that shorter

sentences can be more informative during inference. Our experimental results suggest that relying on semantic information rather than sentiment information, contrary to previous literature, may be more reliable for depression assessment.

Keywords: depression severity assessment, speech transcription analysis, text analysis, deep learning.

ÖZET

DEPRESYON ŞİDDETİ DEĞERLENDİRMESİ İÇİN KONUŞMA ÇEVRİYAZILARININ MODELLENMESİ

Ergün Batuhan Kaynak
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Hamdi Dibeklioğlu
Eylül 2022

Herkesin zaman zaman kötü günleri olduğu doğrudur. Ne yazık ki, bu günler depresyondan muzdarip insanlar çok daha fazladır, ve neredeyse her gün en basit şeyleri bile yapabilmek için bir motivasyon savaşı verirler. Bu sırada sürekli olarak umutsuzluk, fiziksel ve duygusal yorgunluk ve üzüntü ile uğraşırlar. Bu hastalıktan muzdarip insan sayısının giderek arttığı göz önüne alındığında, otomatik depresyon şiddeti değerlendirme sistemlerine duyulan ihtiyacın önemi anlaşılmaktadır.

Bu hastalığın çözümüne yardımcı olmak adına, depresyon şiddeti tahmini için girdi olarak konuşma dökümlerini kullanan modüler bir derin öğrenme ardışık düzeneği öneriyoruz. Ardışık düzeneğimiz aracılığıyla, popüler derin öğrenme mimarilerinin depresyon değerlendirmesi için temsiller oluşturmadaki rolünü araştırıyoruz. Metin kipine ilişkin depresyon değerlendirme literatürünü genişletmek amacıyla cümle istatistiklerinin kapsamlı bir analizini ve bunların model eğitimi üzerindeki etkilerini sunuyoruz. Ayrıca, depresyon değerlendirmesi için duygu bilgilerinin kullanımına ilişkin bir araştırmaya da yer veriyoruz.

Önerilen mimarilerin değerlendirilmesi, halka açık Genişletilmiş Tehlike Analizi Mülakat Derlemi veri setinde (E-DAIC) gerçekleştirilmiştir. Sonuçlar ve tartışmalar aracılığıyla, cümleler arasındaki zamansal dinamiklerden yararlanmadan depresyon değerlendirmesi için bilgilendirici temsillerin elde edilebileceğini gösteriyoruz. Önerdiğimiz bu zamana bağlı olmayan model, Uyum Korelasyon Katsayısı (CCC) açısından güncel en iyi teknolojiden %8.8 daha iyi performans göstermektedir. Eğitilmiş modeller ve veri istatistiklerine ilişkin bulgularımız ışığında, tekrarlayan yapıların eğitim sırasında belirli dizi uzunluklarına karşı nasıl bir önyargıya sahip olabileceğini ve çıkarım aşamasında daha kısa cümlelerin

daha bilgilendirici olabileceğini tartışıyoruz. Deneysel sonuçlarımız, önceki literatürün aksine, duygu bilgisinden ziyade anlamsal bilgiye güvenmenin depresyon değerlendirmesi için daha güvenilir olabileceğini düşündürmektedir.

Anahtar sözcükler: Depresyon şiddeti değerlendirmesi, konuşma içeri analizi, metin analizi, derin öğrenme.

Acknowledgement

I would like to express my eternal gratitude to my advisor Asst. Prof. Dr. Hamdi Dibeklioğlu. Without his great expertise, unending academic and moral support throughout my studies, and most importantly, everlasting patience, I would not be where I am today.

I am also grateful for my friends and family, who listened to me talk about my thesis these past two years. Even though they possibly did not understand most of what I said, their unceasing support and trust in me is the foundation on which this thesis is built.

The View From Halfway Down

The weak breeze whispers nothing
the water screams sublime.
His feet shift, teeter-totter
deep breaths, stand back, it's time.

Toes untouch the overpass
soon he's water-bound.
Eyes locked shut but peek to see
the view from halfway down.

A little wind, a summer sun
a river rich and regal.
A flood of fond endorphins
brings a calm that knows no equal.

You're flying now, you see things
much more clear than from the ground.
It's all okay, or it would be
were you not now halfway down.

Thrash to break from gravity
what now could slow the drop?
All I'd give for toes to touch
the safety back at top.

But this is it, the deed is done
silence drowns the sound.
Before I leaped I should've seen
the view from halfway down.

I really should've thought about
the view from halfway down.
I wish I could've known about
the view from halfway down—

Alison Tafel

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Depression is a mental disorder that negatively affects the feelings, behaviours, and thoughts of individuals. Overwhelming feelings caused by depression can hinder the individual by leading to disinterest in daily activities and reduced concentration. It can even manifest itself as physical pain. Diagnosis of depression is very important as individuals, in the worst case, can be driven to suicide without proper treatment. Depression has many challenges, both regarding its diagnosis and treatment. Mental health issues are mistakenly not taken as seriously as physical health issues, and most people can show reluctance to accept they are suffering from an illness and seek professional help. This is exacerbated in the case of depression since depressed individuals generally do not have the motivation to perform simple daily tasks, let alone seek treatment. The difficulty of understanding the human psyche is also a primary concern. This can cause misdiagnosis of the severity of depression, as the symptoms can vary depending on individual differences of the patient.

To control this uncertainty, standardized tests are proposed. A popular test is Hamilton Depression Rating Scale (HDRS) [1]. This test contains point scales in many depression cues, such as sleep quality, physical activity, guilt, and anxiety. The expert is expected to score the individual on these cues to understand their depression severity. As another means of assessment, individuals are also

asked to self-assess using simple questionnaires, such as Physical Health Questionnaire Depression Scale (PHQ) [2]. Due to the subjectivity of feelings, these self-assessments can be unreliable.

The recent COVID-19 pandemic acted as a figurative breeding ground for depression. With many people stuck in their homes, deprived of their daily routines, anxiety and depression increased by %25 [3] according to recent statistics by World Health Organization. In light of this, many studies are conducted to further investigate the effects of the pandemic and depression [4, 5, 6]. This recent surge in depression, along with the discussed challenges, makes it clear that the need for good automated detection of depression severity systems is more important than ever. With such systems, we can help over 300 million [7] people suffering from depression and even save their lives.

1.1 Related Work

Most depression severity assessment literature focuses on audiovisual modalities and their fusion. Before we focus on the text modality, we give a brief summary of such studies. Visual cues are very important when it comes to understanding the mental state of an individual. This is especially the case when a person does not show verbal or auditory cues by not talking. Psychological studies show that reduced nonverbal reactivity for emotions is an important marker for depression [8, 9, 10]. Prominent emotions such as joy or anger are more easily modeled and shed light on the mental state of a person, while subtle changes or distortions in facial expressions such as blinking, lip movements, or twitching require a more complex computational model. Due to the importance of temporal dynamics in understanding such cues, temporal modeling of videos is favoured over single images [11]. Acoustics-wise, we know that depressed individuals have lower fundamental frequency f_0 and longer pauses between their answer periods compared to healthy control groups [12, 13]. Similar to the visual reactivity finding, depressed people are also less loud and have reduced pitch compared to healthy people [14, 15, 16, 17]. The frequency of low audiovisual functionality

cues is also significantly correlated with the diagnosis of depression [18].

The emergence of visual depression detection happened with the creation of active appearance models (AAM) using facial action coding system action units (FACS-AU) or region units (FACS-RU) [19, 20, 21]. These statistical models analysed handcrafted features, such as the activation for depression-related action units, the duration of activations, and onset-offset dynamics. A big shortcoming of this approach is that AAMs are subject-dependent and they need to be reconfigured for each participant. Early audio analysis attempts are also simple and use mostly handcrafted features like pause duration and f_0, and classify using logistic regression [22]. For both visual and audio modalities, more recent studies use convolutional neural networks (CNN) [23] to create representations from raw data and process them through recurrent neural networks (RNN) to extract temporal information [24]. Long short-term memory (LSTM) [25] and gated recurrent unit (GRU) [26] architectures are more widely used as they address the vanishing gradients problem that RNN suffers from. Recurrent processing can also be applied to raw features [27]. It is also possible to use CNNs for temporal learning, as in the case of [28] where the authors use two separate CNN architectures: one for the analysis of still frames and the other for learning temporal dynamics via optic flow images. Chen et al. also demonstrate this concept by creating fixed-size spectral heatmaps of facial attributes and processing them with a CNN for joint spatiotemporal training [29]. Use of public frameworks for feature extraction, such as OpenSMILE [30] for audio and OpenFace [31] for vision, is also utilized [32, 33, 34]. Following the increased popularity of depression severity assessment task, novel attempts start to emerge. Sun et al. make use of the transformer [35] architecture [34] for temporal modeling. As another approach for temporal processing, [29] shows that it is better for depression prediction to process features with no clear spatial correlations (such as AUs, gaze, and head pose) using CNNs, while processing features with strict spatial distributions, such as facial landmarks, using graph convolutional networks (GCN) [36]. If the reader wishes to learn more about audiovisual modalities and their use in depression severity assessment, we point them towards the reviews on the subject [37, 38, 39].

Looking at the dimensionality in which they are used, we see that audiovisual

modalities require a monumental amount of computational resources to process. To overcome these problems, most studies process these modalities with lower sampling rates [19, 40](e.g. averaging 30 frames of video or 100 samples of audio per second to obtain a representation for each second). While this method helps with the resource problem, it is unclear whether these features are enough to detect subtle cues and how much information is lost.

Due to the multi-modal nature of the problem, studies rarely use a single modality if they have access to more. An increase in accessible multi-modal datasets contributed to the emergence of multi-modal fusion networks for depression severity assessment. The majority of the studies apply fusion of multiple modalities to increase their performance. Fusion is usually applied on higher levels of the network after each modality is processed via individual networks. Final Representations of modalities are either concatenated [27, 33, 41] and regressed to a final output, or the fusion simply happens after scores from each modality are regressed using adaptive late-fusion of scores [34]. Yin et al. [40] take an earlier fusion approach by first individually processing each modality in small clips and then applying the concatenation of modalities before these clip summaries are temporally processed again. After proposing a multi-modal, multi-level attention network, Ray et al. point out that using only the text modality part of the network without any fusion achieves better generalization [42].

Feature importances for mental health models can be interpreted via explanation methods. Using the model-agnostic Kernel SHAP method [43], Baki et al. [44] calculate feature importance for their study on mania-level classification of bipolar disorder patients. For D features and c classes, Kernel SHAP produces a $D \times c$ matrix for each participant that signifies relative feature importances for each class. These values can then be analyzed to understand which feature is more descriptive for a class.

Studies show that many syntactic and statistical measures regarding language correlate with depression, such as the decrease in syntactic complexity [45] or the use of first-person pronouns [46]. However, compared to other modalities, fewer studies use text modality. When text modality is used, it is usually used as an

additional modality rather than the main focus. Due to this lack of focus on text, most studies use rudimentary processing methods. Kaya et al. create 42 functionals using low-level descriptors such as word count and speech duration, along with a bag of words representation for each participant using term frequencies. Both these text-based features are then evaluated both by themselves and the use of weighted fusion networks [47]. Ye et al. choose to use the top 10 most frequent words to differentiate between healthy and depressed groups [48].

Deep learning based natural language processing embeddings are becoming more and more popular. Consequently, depression assessment networks also started utilizing these high performance semantic information descriptors. Studies [49, 50, 51] use Word2vec [52] and its variants such as Wikipedia2Vec [53] or Doc2Vec [54] to extract representations. Recently, more powerful sentence embeddings are utilized with Universal Sentence Encoder [55, 42, 56, 33] or BERT [57] models. These embeddings are usually used without finetuning the embedding network. An overwhelming majority, if not all, of recent deep learning networks process word and sentence embeddings using a recurrent architecture to explore temporal relationships [42, 41, 33, 27]. Differently, Yang et al. [58] format the text as a two-dimensional matrix of words and embeddings and process it using a TextCNN [59] variant with k-Max-pooling.

Depression data is not as easily obtainable compared to most computer science tasks. This is true for both initial data acquisition from therapy sessions and for sharing the data with other researchers. Privacy and ethics are the main concerns due to the nature of the data. Clinical data is usually more subjected to restrictions regarding sharing, and most of them cannot be shared even if data is not in its raw format (e.g. numeric frame representation vectors instead of an unprocessed image). This slows down research on depression since the studies cannot be replicated or compared directly. Due to these restrictions on clinical data, most depression research uses public datasets. Most of these datasets come from challenges and provide a space for researchers to compare different models. While some datasets contain only text modality [60, 61] (i.e. conversation transcripts), most of them contain audiovisual features [62, 63]. More recent datasets contain all three modalities [64, 32].

1.2 Motivation

The majority of the studies in the literature use audiovisual features to tackle the depression severity assessment problem. The use of text modality is comparatively lower. Usually, text is used as a supporting modality in fusion-based methods, using architectures that are relatively less complex than audio and video modalities. Consequently, the effects of text modality are not studied well. Text modality has several advantages over other modalities. Channels like social media and messaging apps contain an abundance of text data. Although audiovisual modalities are also present in such channels, they are mostly used for other purposes and do not shed light on an individual's psyche as much as text modality does. Several studies document the potential of the social media domain [65, 66, 67]. Another property of text modality is that it is significantly harder to identify a person using non-handwritten text only. In contrast, a short audio recording or a single image can be enough to identify an individual. Such arguments create great motivation for the use of text modality for depression analysis.

Motivated by such reasons, this thesis focuses on the analysis of text modality for the depression severity assessment task.

1.3 Outline of the Thesis

The outline of the rest of the thesis is as follows: In Chapter 2, the proposed modules for our depression severity assessment system are presented. Chapter 3 details the dataset used and explains the experimental setup used for ablation testing of modules and the final system. This chapter also reports our experimental results, analyzes them, and compares them to other studies in the literature. Finally, the thesis is concluded by providing a summary of the work, and possible future directions in Chapter 4.

Chapter 2

Methodology

This thesis introduces architectures from literature as modular components. The effects of these components are analyzed through ablation studies, and several architectures are proposed as a result.

2.1 Overview

Figure 2.4 illustrates the schema for the proposed modular pipeline. Remaining sections within this chapter detail the modules that will be used during experiments along with where they fit in our pipeline. Each module touches on a different consideration that emerges while building a regression network. Methods within each module have very similar, if not the same, input and output dimensions and considerations. This results in a highly modular experiment setup where each method of a module can be seamlessly interchanged with one another. After all modules are introduced, we present our proposed architectures in Section 2.6. These architectures are formed according to ablation studies performed in Chapter 3.

Main processing flow in this thesis is as follows: First, each sentence for a

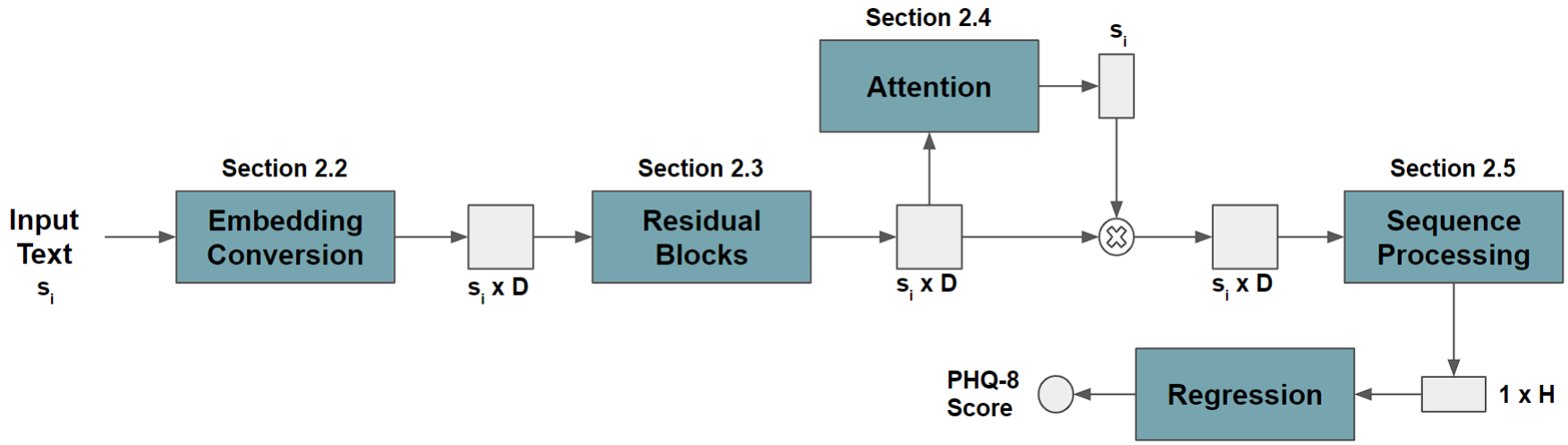


Figure 2.1: Overview of the proposed pipeline. Data representations at several stages are depicted with their shapes. Shapes are given for a single participant and not batched input. For participant p_i, s_i is the number of sentences, D is the size of the sentence embedding (depends on embedding choice), and H is the output dimension of the sequence processing module ($H = D$ for non-temporal modules, but it is a hyperparameter for temporal ones). More information about each modular component can be found in their respective sections.

participant is converted from text to a numeric vector representation. It is required for any form of natural language processing algorithm to apply this step first (unless we are considering a rule-based approach based on the actual string content). Section 2.2 describes the method we follow to achieve this conversion. These sentence embeddings can go through an individual processing step using residual blocks (Section 2.3) and/or the attention module (Section 2.4). These modules allow us to learn an intermediate representation before modeling the sequences as a whole. Our next task is to reduce these variable number of intermediate representations into a single representation, i.e. a summary of the participant. To this aim, Section 2.5 discusses both temporal and non-temporal summarization methods. Temporal methods use the order information of each sentence (i.e. each sentence is processed within the context of its preceding sentences), while non-temporal methods are not concerned with when the sentence is uttered. Ultimately, the summary representation for each participant is regressed into a single value, which is our PHQ-8 prediction, using linear regression layers.

2.2 Transcript Representation

To process text data with any deep learning architecture, it is necessary first to convert them into numerical representations. Historically, handcrafted algorithms are used to map words or sentences into numeric vectors. More recently, deep learning based architectures are used. This is a required step for our architectures.

For participant p_i, we obtain a sequence of sentence embeddings $P^{(i)} \in s_i \times D$, where s_i is the number of sentences and D is the embedding size. We define the single sentence embedding uttered within time period t as x_t. A summary of the embeddings can be found in Table 2.1. Note that their results are presented for the sake of completeness, and since the embeddings do not necessarily use the same datasets during their evaluation, they are not comparable. Embedding size D depends on the choice of embedding (see Table 2.1), but it does not create any considerations while building our architectures.

Table 2.1: Names and dimensions for different embeddings that are analyzed in this thesis.

Embedding	Embedding Size (D)
all-mpnet-base-v2 [68]	768
SiEBERT [69]	1024
CardiffNLP-Sent [70]	768

We start by introducing our main embedding, all-mpnet-base-v2 [68]. This embedding is a finetuned version of Microsoft's mpnet-base model [71]. Finetuning was applied with a contrastive loss objective using over 1 billion training pairs. Among other embeddings from the same framework, all-mpnet-base-v2 has the best average performance on 14 diverse sentence embedding performance tasks and 6 various semantic search tasks. Due to its performance and popularity, we believe this embedding is a good starting point for our ablation studies. We also introduce two sentiment-based embeddings. SiEBERT [69] model finetunes the RoBERTa [72] network (which was trained with 160 GB of uncompressed text data) on 15 datasets on sentiment classification task, and achieves an average of %93.2 accuracy across all datasets. Second sentiment network is created by CardiffNLP research group, and it is another RoBERTa-based network. The network is trained using 58 million tweets and finetuned with the TweetEval sentiment analysis benchmark dataset [70]. The authors report 72.9 macro-averaged recall on their test set for this embedding. We include this network to examine the effects of using tweets, a more casual form of text, in depression assessment.

It should be noted that these architectures are not appended to our end-to-end depression severity assessment network. To elaborate, output sentence embeddings are frozen, and the error from our PHQ-8 value prediction is not propagated back to these networks. Embeddings for every sentence are the same throughout training and validation procedures. This method was chosen to reduce computation and memory costs. Performance differences in using different embeddings are discussed after model selection in Section 3.3.4.

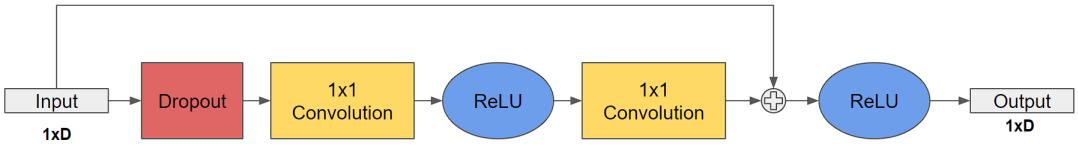


Figure 2.2: Architecture for a single residual block. These blocks can be appended end to end as much as required.

2.3 Residual Blocks

Residual blocks are the next module after converting sentences to sentence embeddings. Here, each sentence embedding x_t is processed through a variable amount of connected residual blocks. Diagram for a single residual block is depicted in Figure 2.2. These residual blocks use the residual learning idea from [73], where we add the block input to the output of that block. This skip-connection helps the network by both reducing vanishing gradients and reducing the possibility that new blocks degrade previously learned information. Since this module is optional, output representations are also called x_t for ease of notation.

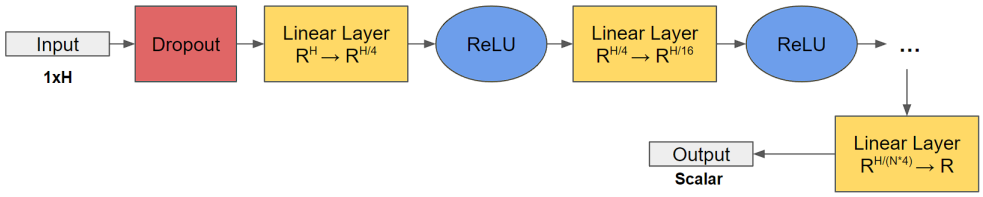


Figure 2.3: Architecture for the attention/regressor module.

2.4 Attention

In this section, we present our attention mechanism that can be used to introduce an additional scaling in between modules for intermediate vectors. Attention

weight a_t is calculated by regressing a scalar value from each intermediate representation x_t. This regressor, depicted in Figure 2.3, is identical to the one we utilize to regress a PHQ-8 score after sequence processing. x_t is then scaled with their respective attention score a_t before being pooled into a summary representation. We include a dropout layer before linear layers for regularization. Each linear layer reduces the input dimension by a multiple of 4 (i.e. $\mathbb{R}^d \rightarrow \mathbb{R}^{\frac{d}{4}}$). The last layer outputs a single scalar no matter the input dimension. Output attention weights can be normalized to the 0-1 range using the sigmoid function or min-max normalization:

$$a_t = \frac{a_t - \min(a_1, a_2, ..., a_t)}{\max(a_1, a_2, ..., a_t) - \min(a_1, a_2, ..., a_t)}$$

We use such normalization functions to better regularize our network weights, and also provide contextual information across representations for a given participant.

2.5 Sequence Processing

Regressing a single scalar PHQ-8 value requires some form of summarization of the many sentences a participant utters during their interviews. This section details such summarization methods.

2.5.1 Temporal Modeling of Sequences

2.5.1.1 Single-Level Gated Recurrent Unit

We use single-layer bidirectional GRUs to temporally model sentence embeddings. GRU architecture is selected due to its documented performance on temporal

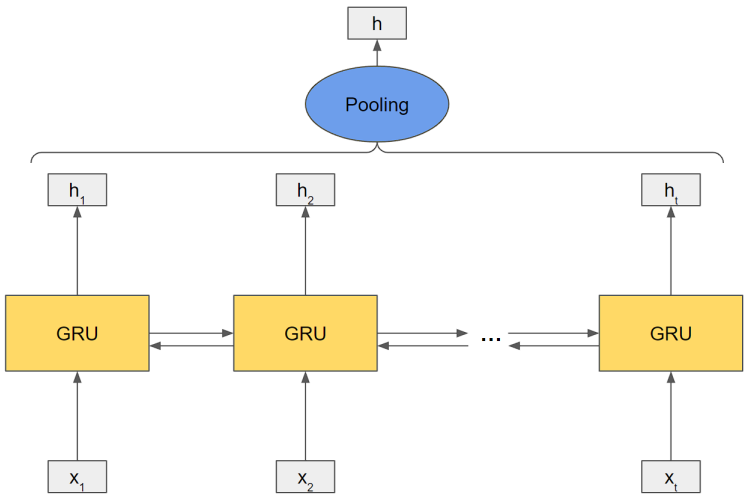


Figure 2.4: Unrolled architecture for the single-level GRU. Each GRU box depicts the same network at different timesteps. Output for each time step is stored and pooled into a single representation.

problems over RNNs and LSTMs, and due to their ability to process variable length sequences. Bidirectionality incorporates information from both directions (forward and backward) of the sequence. We define the forward direction of our single-layer GRU using the following equations:

$$
\begin{aligned}
z_t &= \text{sigmoid}(\mathbf{W}_z x_t + \mathbf{U}_z \overrightarrow{h}_{t-1}) \\
r_t &= \text{sigmoid}(\mathbf{W}_r x_t + \mathbf{U}_r \overrightarrow{h}_{t-1}) \\
c_t &= \tanh(\mathbf{W}_c x_t + \mathbf{U}_c (r_t \odot \overrightarrow{h}_{t-1})) \\
\overrightarrow{h}_t &= (1 - z_t) \overrightarrow{h}_{t-1} + z_t c_t
\end{aligned}
$$

Where r_t and z_t are the reset and update gates, respectively. The activation of the hidden state $\overrightarrow{h}_t$ at time t is the linear interpolation between previous activation $\overrightarrow{h}_{t-1}$ and the candidate activation c_t. Weight matrices $\mathbf{W}$ and $\mathbf{U}$ with subscripts z, r, c are the parameters of the GRU. Each subscript defines another translation from the input sentence embedding x_t. $\odot$ is the operation for elementwise multiplication.

We concatenate the hidden states of the forward and the backward passes for each t as:

$$h_t = \begin{bmatrix} \overrightarrow{h}_t \\ \overleftarrow{h}_t \end{bmatrix}.$$

We reduce the resulting s_i many timesteps to a single vector using either last-pooling, mean-pooling, or max-pooling. Last-pooling simply assigns the output to the last hidden state; mean-pooling takes the average of all hidden states over the t dimension, while max-pooling takes the maximum over the t dimension. Formally, the output $h^{(i)} \in \mathbb{R}^H$ for $P^{(i)}$ is obtained by:

$$\begin{aligned} \text{LAST\_POOL} \rightarrow h^{(i)} &= h_{s_i - 1} \\ \text{MEAN\_POOL} \rightarrow h^{(i)} &= \frac{1}{s_i} \sum_{t=1}^{s_i} h_t \\ \text{MAX\_POOL} \rightarrow h^{(i)} &= \max_t [h_1; h_2; ...; h_t] \end{aligned}$$

2.5.1.2 Hierarchical Gated Recurrent Unit

We also propose a hierarchical recurrent network model. This network processes our temporal sequences in two levels instead of a single level as in Section 2.5.1.1. Sequence of each participant, $P^{(i)}$, is separated into chunks with window length C_w. These chunks can be overlapping and non-overlapping, determined by the overlap amount C_o. Given C_w and C_o, a sequence of length s_i can be separated into $C_n = ceil(\frac{s_i - C_o}{C_w - C_o})$ chunks of length C_w. Non-overlapping chunks are created when $C_o = 0$. If s_i is not divisible by C_w, the last chunk will have $s_i \pmod{C_w}$ elements in it. As seen in Figure 2.5, sequence chunks are first processed with a single-level GRU (Section 2.5.1.1) to create an intermediate representation $h_{1,i}$. Then, the intermediate representations are summarized with another single-level GRU on the second level. Note that all chunks are processed with the same

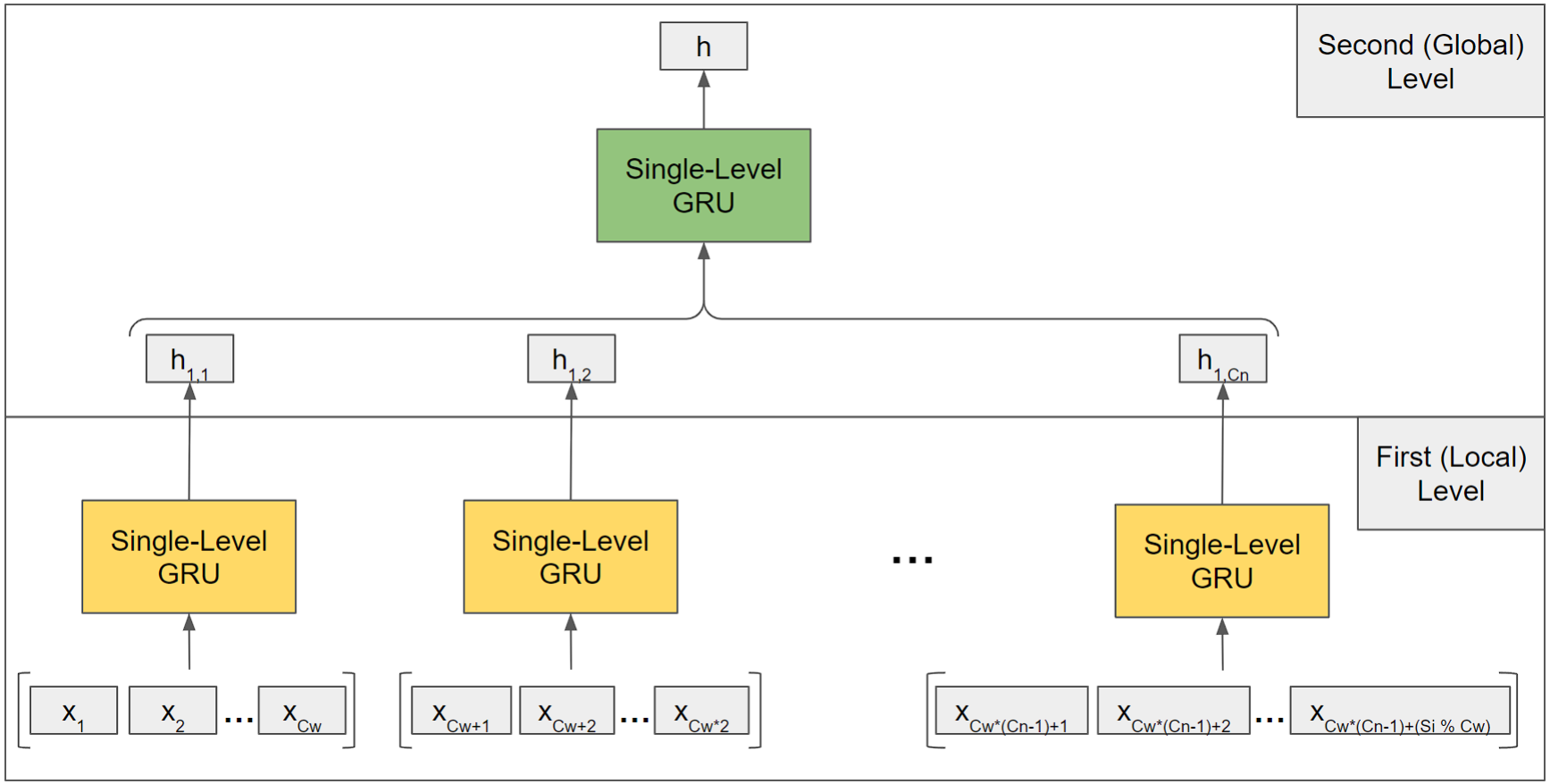


Figure 2.5: Unrolled architecture for the hierarchical GRU. Each GRU box in the first level depicts the same single-level GRU network at different timesteps. GRU at the second level uses pooled results from the first level as input.

single-level GRU on the first level, and the GRUs on different levels do not share weights.

2.5.2 Non-Temporal Modeling of Sequences

Temporal methods are not the only way to process ordered sequences. Temporal architectures have the inherent assumption that there is generalizable information to be found in the order in which we find our sentence embeddings and process each sentence embedding within the context of previous ones. While such methods are widely used in depression severity assessment literature, the idea that sentences can be good indicators by themselves has not been explored much. Theoretically, if sentences themselves are sufficient, additional context information could even be causing overfitting or significant prediction error for participants with underrepresented PHQ-8 scores.

The proposed non-temporal sequence processing module aims to reduce a sequence $P^{(i)}$ with a variable number of sentence embeddings into a single vector. To this aim, we employ several pooling methods. Similar to Section 2.5.1.1, mean-pooling takes the average of sentence representations, while max-pooling filters the maximum activations along the t dimension. Formally, the output $h^{(i)} \in \mathbb{R}^D$ for $P^{(i)}$ is obtained by:

$$\text{MEAN\_POOL} \rightarrow h^{(i)} = \frac{1}{s_i} \sum_{t=1}^{s_i} x_t$$

$$\text{MAX\_POOL} \rightarrow h^{(i)} = \max_t [x_1; x_2; ...; x_t]$$

2.6 Overview of Investigated Architectures

Through our experiments and ablation studies, we combine and examine the sub-modules detailed in this chapter. Some of these complex models are proposed as

good candidate architectures for the depression severity assessment task, while others will be disqualified through our validation process and discussions on behavioral aspects and generalization to real-world scenarios. This section reveals the details of such candidate architectures. Table 2.2 provides a brief summary of these architectures. Abbreviations for the names of our architectures are as the following:

- NT-MEAN: **N**on-**T**emporal model using **MEAN**-pooling
- NT-MEAN-ATT: **N**on-**T**emporal model using **MEAN**-pooling with **A**ttention
- TS-MEAN: **T**emporal model using **S**ingle-Level GRU with **MEAN**-pooling
- TH-MEAN.MAX: **T**emporal model using **H**ierarchical GRU with **MEAN**-pooling on first level and **MAX**-pooling on second level

Table 2.2: Table of attributes for our proposed models.

Model Name	Temporality	Attention
NT-MEAN	Non-Temporal	No
NT-MEAN-ATT	Non-Temporal	Yes
TS-MEAN	Temporal	No
TH-MEAN.MAX	Temporal	No

2.6.1 NT-MEAN and NT-MEAN-ATT

NT-MEAN is our simplest model. Sentence embeddings are passed through residual blocks. These output representations $x_t \in \mathbb{R}^{s_i \times D}$ are then averaged for each participant to obtain a summary representation for that participant. Resulting summary representation of shape $1 \times D$ is regressed with linear layers to obtain the final PHQ-8 score. NT-MEAN-ATT is a non-temporal model similar to NT-MEAN. Their difference lies in the addition of feed forward attention module. This attention module calculates attention scores a_t for each x_t. x_t is then scaled

with their respective attention score before being pooled into a summary representation. Both NT-MEAN and NT-MEAN-ATT architectures are presented in Figure 2.6.

2.6.2 TS-MEAN and TH-MEAN.MAX

TS-MEAN and TH-MEAN.MAX are both temporal models, meaning that they both use a recurrent architecture, namely a GRU, to process intermediate representations x_t created by the residual blocks. TS-MEAN uses a single level GRU followed by mean-pooling to create a summary representation of shape $1 \times H$. On the other hand, TH-MEAN.MAX utilizes a hierarchical GRU with mean-pooling in the first level and max-pooling in the second level to create the same representation. As with our non-temporal models, these output summaries are regressed to PHQ-8 scores. We present these architectures in Figure 2.7.

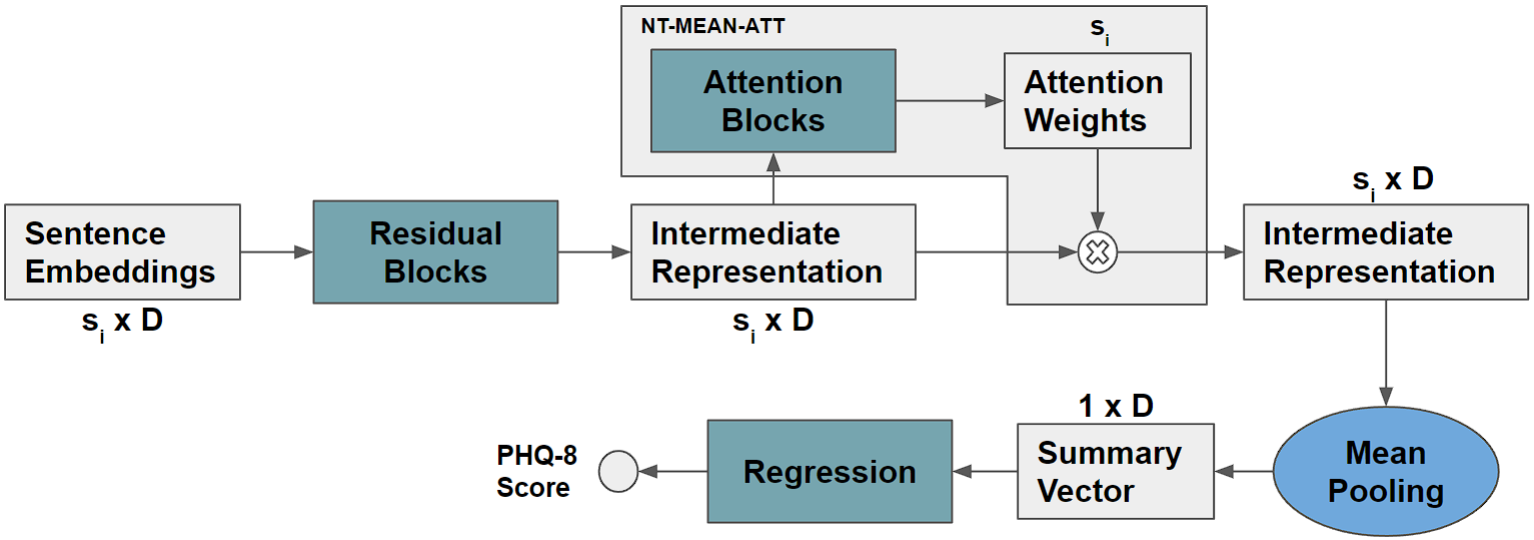


Figure 2.6: Diagram for NT-MEAN and NT-MEAN-ATT architectures. NT-MEAN-ATT additionally applies attention weights after processing the sentence embeddings via residual blocks.

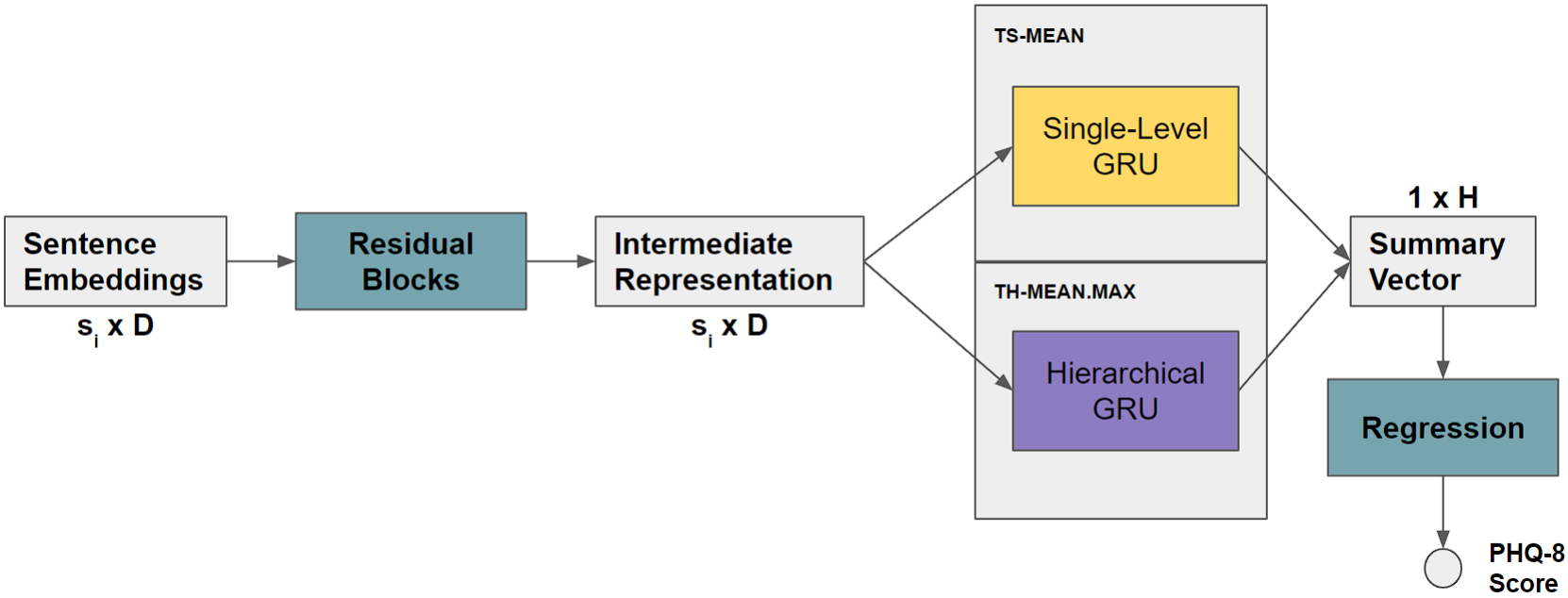


Figure 2.7: Diagram for TS-MEAN and TH-MEAN.MAX architectures. TS-MEAN processes the intermediate representation via a single-level GRU, while TH-MEAN.MAX uses a hierarchical one.

Chapter 3

Experiments and Results

3.1 Dataset

We use the Extended Distress Analysis Interview Corpus dataset (E-DAIC) dataset, which is the dataset for the Detecting Depression with AI Sub-challenge (DDS) during The Audio/Visual Emotion Challenge (AVEC) 2019 Workshop and Challenge [32]. The dataset contains 275 interviews with unique participants, and it is collected in an effort to create an AI agent that can interview and identify mental health problems. The interviews consist of a participant's dialogue with an interviewer. The interviewer is either a human or a fully automated AI. Regardless of the nature of the interviewer, the participant sees an animated virtual avatar on the screen in front of them. Among the 275, 163 subjects are used for training purposes, while validation and test splits contain 56 each. The splits are balanced in terms of age, gender distribution, and PHQ-8 scores. An important difference between splits is that the test split contains interviews where the interviewer is always the automated AI, while training and validation splits contain a mix of both human and AI interviewers. We use the predetermined splits during our experiments.

The dataset contains four video and six audio features along with raw audio

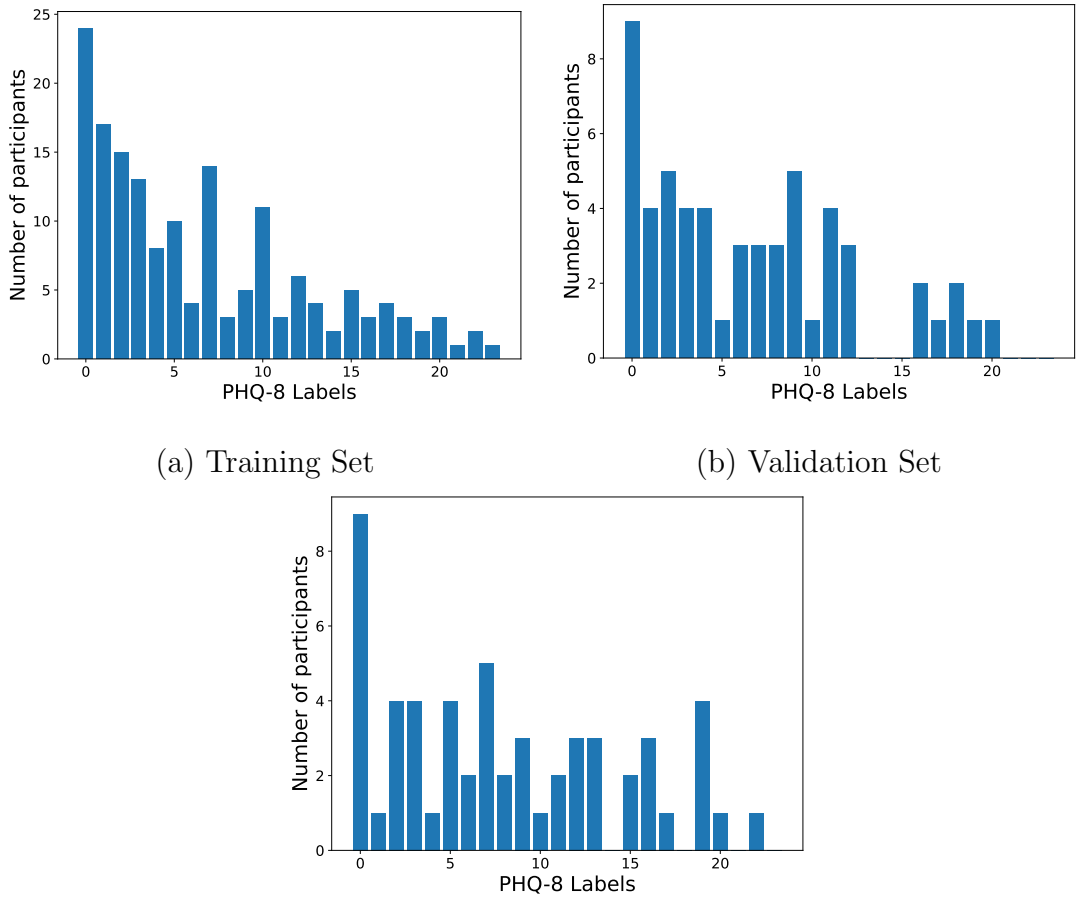
Over the last 2 weeks, how often have you been bothered by any of the following problems? (Use "✔" to indicate your answer)	**Not at all**	**Several days**	**More than half the days**	**Nearly every day**
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself – or that you are a failure or have let yourself or your family down	0	1	2	3
7. Troubleconcentratingonthings,suchas reading thenewspaperorwatchingtelevision.	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed? Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
(For office coding: Total Score ____ = ___ + ___ + ___)				

Figure 3.1: Questions within PHQ-8. A score is assigned for eight different items depending on how frequently the individual is impaired by that condition.

and speech transcripts. The text is transcribed using Google Cloud's speech recognition service. Due to the private nature of the data, raw video footage is not available. This thesis uses only transcribed text portion of the data. Label for each participant is a self-reported 8-item Patient Health Questionnaire (PHQ-8) score [2]. PHQ-8 is a self-assessed depression severity measure. The questionnaire provides insight into the degree of impairment an individual goes through on eight different depression cues. A higher score means it is more likely that the individual is suffering from depression. The minimum score is 0 and the maximum score is 24. The reader is referred to Figure 3.1 for details on the questionnaire.

Figure 3.2 provides the distribution of labels within training, validation and test splits respectively. While the splits are balanced in terms of the PHQ-8 score, there is a high imbalance of scores within each split, i.e. People considered non-depressive (PHQ-8 < 10) make up %69, %73 and %63 (ordered training, validation and test) of all data. This imbalance is increased when we compare participants with severe depression (PHQ ≥ 20) to the remainder of the data (non-depressive PHQ-8 < 10 and depressive $10 \leq$ PHQ-8 < 20). In that case,

participants with severe depression only make up %4, %2 and %7 (ordered training, validation and test) of all data.



(a) Training Set

(b) Validation Set

(c) Test Set

Figure 3.2: Distribution of PHQ-8 labels for (a) training (b) validation and (c) test sets

It should be noted that the transcriptions of sessions are not perfect. There are many sentences that do not exactly match the raw audio, and sometimes the voice of the therapy AI or a technician is also transcribed as sentences from the participant. There are also cases where sentence breaks are not recognized, and several sentences are transcribed as a single long sentence. Since it is not feasible to dynamically correct these mistakes, we left the faulty data in its original state.

The transcribed text also contains a confidence level (a reel number between 0 and 1) for each transcribed sentence. We empirically see that inclusion of this value in our training is generally detrimental to performance. Manual inspection of the dataset shows us that the confidence level is not very reliable, as it often gives low confidence to correctly transcribed words while giving high confidence for bad transcriptions. In light of these inspections, we opt not to use this information.

3.2 Experimental Setup

3.2.1 Evaluation Criteria

Dataset used in this thesis was first introduced in The Audio/Visual Emotion Challenge and Workshop 2019 (AVEC 2019) [32]. Organizers of the challenge picked Lin's Concordance Correlation Coefficient (CCC) [74] as the evaluation metric. CCC is a statistical measure of how well a set of predictions compares to the ground truth labels. Since it is a correlation measure, the value of CCC ranges from -1 to 1, where 1 signifies complete correlation between two sets. More precisely:

$$-1 \leq -|\rho| \leq CCC \leq |\rho| \leq 1$$

where ρ is Pearson's correlation coefficient. For the set of predictions x and corresponding truth values y, CCC is formally defined as:

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

where μ and σ^2 are respectively the population average and standard deviation

for their corresponding variables. Since we are dealing with a sample of the total population, we use an approximation of CCC:

$$\hat{CCC} = \frac{2S_{YX}}{S_X^2 + S_Y^2 + (\bar{Y} - \bar{X})^2}$$

Organizers choose this metric due to its invariance to scale, as well as its ability to include information on accuracy and precision [32]. We also use this metric in our training and evaluation.

Organizers also propose Root Mean Square Error (RMSE) as a secondary metric:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i)^2}$$

RMSE computes the numeric difference between prediction and target without any complex statistics. Taking the square of the error makes it so that higher errors are punished more. When used as a loss function, this property of RMSE can help reduce overfitting that can occur in our dataset due to label imbalance.

Alternative to these two metrics, we also propose reporting Mean Absolute Error (MAE). While CCC and RMSE have great properties during training, they cannot be easily interpreted. Even though MAE is ubiquitous within the literature for regression tasks, we see that it is scarcely reported for this dataset. We believe this metric is important to better understand and discuss our results and should also be reported for this dataset.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Y_i - X_i|$$

3.2.2 Training Setup

We follow the traditional training-validation-test scheme using the predetermined splits of the AVEC competition [32]. To reduce selection bias, and mimic the conditions of AVEC competition, we do not evaluate our models on the test set until we finalize model selection through ablation studies on the validation set. We evaluate four models in the test set during this study, to discuss and compare generalization performances. Implementations are done using PyTorch [75] and models are optimized with optuna library [76] using a Tree-structured Parzen Estimator (TPE) for hyperparameter selection. The following table details the range for our hyperparameters.

Table 3.1: Table of hyperparameters and their corresponding search ranges

Hyperparameter Name	Considered Values
Batch Size	{4, 8, 16, 32}
# Residual Blocks	{1, 2, 3, 4, 5}
Residual Block Dropout Probabilities	{0, 0.1, 0.2, 0.3, 0.4, 0.5}
# Linear Layers	{1, 2, 3, 4, 5}
Linear Layer Dropout Probability	{0, 0.1, 0.2, 0.3, 0.4, 0.5}
# GRU Layers	{1, 2}
GRU Hidden Dimension Size	{64, 128, 256, 512}
Hierarchical GRU C_w	{3, 5, 7, 10, 15}
Hierarchical GRU C_o	{1, 2, .., C_w - 1}

We use Adam optimizer with 10^{-4} learning rate. Training data is shuffled each iteration, but no augmentation is applied. The training is terminated if our validation loss doesn't improve for 25 epochs. When the training is terminated, the checkpoint with the lowest validation loss is taken as the trained model. We empirically see that Batch Normalization [77] is generally detrimental for our training, and do not include it in our models. We use ReLU [78] as our activation function of choice due to its popularity and performance over other activation functions. All experiments are conducted on an Nvidia RTX 2080 GPU.

3.3 Experimental Results

3.3.1 Temporal Modeling

Temporal models include the order of sentence representations within the sequence as additional information. This is done by processing each sentence embedding in order through recurrent architectures. This thesis uses the GRU architecture to process sequences for each participant. The output of the GRU is based on the utilized pooling method, which we will select in the following section.

3.3.1.1 Comparison of Pooling Methods

We compare three different temporal models in Table 3.6, each of which use a different pooling method to obtain a single summary vector. We use a single-level GRU in this section, meaning that the entire sequence is processed by a single GRU.

Table 3.2: Comparison of pooling methods for single-level temporal architectures

Pooling Method	Val CCC
Last	0.589
Mean	0.650
Max	0.646

Last-pooling method, which performs the worst, is generally the default configuration for a recurrent network, where the hidden state at the last time step h_t is taken as the output. The performance of this sequence reduction depends on the assumption that h_t holds the information for the entire sequence. Based on the length of the sequence and the decisions of gates within the architecture, this assumption may not hold well. Also, our knowledge of the nature of interviews and manual inspection of the data shows us that the last couple of sentences are reserved for farewells (e.g. "goodbye", "bye-bye", "okay bye") or small talk about

the interview (e.g. "a real life person is really looking at me", "I was expecting", "that was cool"). Also, as we discuss on 3.1, an operator's voice can be mistaken as the interviewee's and transcribed into text. This usually happens at the start or the end of the interview. Since hidden states hold more information on recent timesteps, these noisy data points can pollute the hidden state and, therefore, reduce the information contained within our summary vector.

Temporal model performance is significantly improved using max or mean pooling instead of last-pooling. Both mean and max pooling has been used extensively in the literature. Theoretically, max-pooling works best when the existence of certain peak values is very important for inference, and completely saturating other activations are not a problem. Conversely, mean-pooling is a better choice when losing minima and maxima is not important, but keeping the overall activation is. The slightly superior performance of mean-pooling over max-pooling in our case could mean that for a participant to be high on the PHQ-8 scale, their cues need to be apparent throughout the interview and not only during some parts of it.

3.3.1.2 Effect of Hierarchical Modeling

Our previous iteration of recurrent architecture processes the sequence as a single unit. The purpose of hierarchical modeling is to handle this process on two levels: local and global. The first-level GRU works on relatively small fixed-size chunks of the initial sequence and provides their summary to the second-level GRU, which handles the temporal processing of the summaries.

3.3.1.2.1 Non-overlapping sequences on first level Effects of pooling methods on temporal processing are given in the previous section (Section 3.3.1.1). Results for that analysis may not hold for the GRUs in our current hierarchical configuration since their objective and sequence lengths are quite different. For this reason, we reinvestigate the effects of pooling methods for each level. We also search for different window lengths, C_w, which determines the

length of the chunks for the first level. In this section, we use non-overlapping chunks in the first level.

Table 3.3: Comparison of pooling methods for hierarchical temporal architectures

First Level Pooling	Second Level Pooling	Val CCC	Best C_w
Last	Last	0.632	15
Last	Max	0.649	7
Last	Mean	0.634	7
Max	Last	0.655	10
Max	Max	0.649	5
Max	Mean	0.637	3
Mean	Last	0.658	15
Mean	Max	0.659	7
Mean	Mean	0.624	7

Results show that some configurations of the hierarchical model perform better than the single-level GRU, while others are still behind single-level with mean or max pooling. Compared to using last-pooling for a single-level GRU, using last-pooling at the first level of the hierarchy does not have any obvious detrimental effects on performance. This is possibly due to the increased information stored within the last hidden state of each chunk in the first level. The selection of C_w also seems to be dependent on the choice of the pooling method for the second level. Second level last-pooling thrives with higher C_w (10, 15) compared to max and mean pooling (3, 5, 7). Also, two of the top three results in this analysis use last-pooling in the second level, providing another evidence that last-pooling thrives with shorter sequences. These findings about last-pooling show us that temporal information about depression is not retained for a long time.

The performance of mean-pooling in the first level is also noteworthy. We see that for configurations where the second level uses either last or max-pooling, having mean-pooling in the first level is always better. This does not hold when mean-pooling is used for the second level. This could mean that the local scope is better used to understand the overall depressiveness of small conversation episodes, and the global scope does better at forming a final representation using these summaries.

3.3.1.2.2 Overlapping sequences on first level This section is a reiteration of the analysis in the previous section, but with overlapping chunks in the first level. We experiment with overlapping chunks since the clear-cut nature of non-overlapping chunks can hinder the continuity of chunks received by the second level. Due to the increase in computation costs, we reapply the analysis for the top three performing configurations from the previous section.

Table 3.4: Results of the hierarchical temporal model when chunks for the first level are created with overlaps (i.e. $C_o \neq 0$). Corresponding $C_o = 0$ results are also reported here again for ease of comparison.

First Level Pooling	Second Level Pooling	Val CCC	Best C_w	Best C_o
Max	Last	0.655	10	0
Mean	Last	0.658	15	0
Mean	Max	0.659	7	0
Max	Last	0.648	10	1
Mean	Last	0.621	15	1
Mean	Max	0.656	7	1

Table 3.4 shows that overlapping chunks are detrimental to performance. So much so that the best overlap picked by every model is 1, meaning that the optimization process tries to reduce overlap as much as possible. To unpack what is happening here, we can observe the consequences of using overlap. We notice that bigger overlaps create more summaries for the second level to process. In the worst case, where $C_o = C_w - 1$, the sequence for the second level is almost as long as the initial sequence. As we can expect, last-pooling takes the biggest hit from longer sequences. It seems that even if overlapping chunks can create a smoother summary transition for the second level, this does not help create a better final representation.

This further supports the point we made in Section 3.3.1.2.1, where we argued that the first level is responsible for picking summaries while the second level selects distinct summaries to create the final representation. Overlapping chunks cause summaries to be too similar, and therefore this reduced salience makes the selection process in the second level harder.

3.3.1.3 Best-Performing Temporal Models

When we compare our single-level and hierarchical GRU experiments, we see that the models are relatively close in performance. The best performing single-level model beats 6 of the 9 hierarchical configurations. To better understand the effects of using a two-level approach, we examine the best-performing model from both single-level and hierarchical experiments in Table 3.5.

Table 3.5: Results for best performing temporal models. Results are given for three metrics in both validation and test sets.

	——Validation——			———Test———		
Model	CCC	MAE	RMSE	CCC	MAE	RMSE
TS-MEAN	0.650	3.393	4.66	0.598	5.232	6.656
TH-MEAN.MAX	0.659	3.321	4.33	0.572	4.464	5.616

Table 3.5 shows us that the hierarchical GRU model, TH-MEAN.MAX, performs better than our single-level GRU, TS-MEAN, in all metrics except test set CCC. Even so, both models have severe generalization problems.

In this section, we examined the temporal dynamics regarding depression cues. Our findings show that there is an episodic nature to the interviews we are analyzing. This can also be seen when manually inspecting the dataset. Compared to longitudinal clinic therapy sessions, the conversations within our dataset seem to move between topics. The summaries for these episodes can be temporally analyzed to obtain good predictions. While learning such relationships result in good models, it is unclear how much of our performance can be ascribed to them. This makes us question the reliability and benefit of temporal architectures. Indeed, if we think about our data, if we know that the participant said: "I haven't been happy at my jobs for at least 10 years", do we need to relate that information to the sentence "New York"? (sentences taken from a participant within our dataset). There is no denying the importance of contextual information, especially for audiovisual modalities [22]. However, we believe they are not as strong for text modality. Given more data, it may be possible to form contextual relations. But for our case, forcing the model to create such relations might result in

noise most of the time.

3.3.2 Non-Temporal Modeling

Following our findings regarding temporal dynamics, we experiment with the simpler non-temporal approach. In this approach, each sentence embedding is passed through several residual blocks before they are pooled into a single vector. The selection of the pooling method and the number of residual blocks are optimization considerations. Note that we opt for using at least one residual block. This is because we want to introduce trainable parameters before any pooling operation is done on the sequence. Otherwise, error gradients cannot backpropagate to individual sentences, and the pooling operations (which also do not contain trainable parameters) give deterministic results.

3.3.2.1 Comparison of Pooling Methods

We compare two different non-temporal models, each of which uses a different pooling method to obtain a single summary vector.

Table 3.6: Comparison of pooling methods for non-temporal architectures

Pooling Method	Val CCC
Mean	0.673
Max	0.629

In Section 3.3.1, we hypothesize that temporal information could hinder performance. To this aim, we discard recurrent modules from our architecture and replace them with simple pooling operations. Observing the performance of mean-pooling in Table 3.6, it appears that the exclusion of temporal information leads to a performance increase. As with the temporal pooling experiments in Section 3.3.1.1, mean-pooling is superior, this time with a bigger margin compared to max-pooling. Observations we can make here regarding the comparison of max and mean pooling are similar to the ones made in Section 3.3.1.1. It seems that

individual high activations are less impactful while forming a summary vector compared to computing the overall activations.

3.3.2.2 Effect of Weighting Embeddings

As per our discussions, mean and max pooling both make different assumptions on the relative weights of sentence representations. The reason for the performance of mean-pooling is unclear. Since we know that not every sentence is a cue for depression, we would expect such sentences to provide noise to the averaging process. For this reason, incorporating other modules to have a better representation selection process could result in a better average summary. To this aim, we experiment with Softmax Weighted Mean-Pooling and Attention Weighted Mean-Pooling (named SWM and AWM, respectively).

SWM simply takes the softmax of intermediate representations. Softmax values are then multiplied with their corresponding representations to scale them before mean-pooling is applied. This incorporates feature importance to each representation and by proxy to the summary vector. AWM technique calculates an attention score a_t for each representation. a_t for each sentence representation can be any real number. Since this can cause scaling instabilities, we also experiment with applying two normalization techniques before we multiply it with its respective sentence embedding: min-max normalizing a_t to the range $0 - 1$ and passing a_t through a sigmoid function. As with SWM, each a_t is multiplied with its corresponding representation before being pooled to create a summary representation.

Table 3.7: Comparison of different representation weighing methods for non-temporal architectures

Pooling Method	Val CCC
SWM	0.581
AWM /wo Norm	0.642
AWM /w Min-Max Norm	0.654
AWM /w Sigmoid Norm	0.645

Our results show that AWM with min-max normalization does better than the alternatives, and AWM in general performs better than SWM. Since AWM contains an additional network to compute individual weights for each representation, each weight is calculated independently of the other embeddings within the sequence. Weights from Softmax, on the other hand, depend highly on the length of the sequence. To elaborate, the same representation can have a very different softmax weight for different participants since softmax distributes a probability of 1 among all representations of that participant. While this can create better relative weights within the sequence, it is a source of high variance for the model in general.

Although sigmoid introduces additional non-linearity to our network, it could give saturated weights for some embeddings due to its shape. It also doesn't have a way of incorporating information from other representations within the sequence. One could argue that min-max normalization is better in that regard, as it does a better job of distributing weights linearly among other representations.

Although AWM with min-max normalization is the best among weighted models, it is still behind simple mean-pooling (Section 3.3.2.1). Scaling each representation with learned weights seems to result in less representative embeddings. The reason could be similar to the arguments we made for temporal information in Section 3.3.1. There, we argued that while temporality could be informative by incorporating contextual information, it could cause more noise than information. While we use a relatively simple way to add context information in this section, it still causes noise to our model.

Since we argued that some sort of selection should happen for a good model, we analyze our non-temporal mean-pooling model (NT-MEAN) to better understand its inner workings. To this aim, we come up with a way to relatively weigh the intermediate representations used by a trained NT-MEAN model. We opt for using a measure of magnitude. Namely, we take the average of each representation over the embedding dimension to obtain s_i magnitude averages. These averages are then min-max normalized to the 0-1 range. These weights are called feature importance for a given representation. These weights are not directly a measure

of depression per se, but rather signify how much a sentence is deemed important for giving a PHQ-8 prediction. Whether the model prediction is high or not depends on the interaction of residual block outputs with the linear regression head, and is not easily interpretable.

Table 3.8 shows the PHQ-8 prediction from NT-MEAN model side-by-side with the PHQ-8 label for the participant. Then, it lists the highest and lowest three feature importance weights for each participant. These weights are presented alongside their corresponding original text. All participants are from the test set. We see that sentences with high feature importance weights are indeed good indicators for either depressive or healthy behaivour. Sentences with lower feature importance seem to be either neutral or longer and more convoluted sentences. The weights are not perfect; as we can see, several sentences depicting unfortunate life events have low relative weights. Nonetheless, this shows that there is an inherent selection process. The context information is possibly applied by using mean-pooling: If the average of representations is more leaning toward a behaviour (i.e. either depressive or healthy), this means that the participant contains relatively more important representations that point to that behavior.

Table 3.8: Feature importances assigned by NT-MEAN model, along with the corresponding raw sentence data.

Prediction	PHQ-8	Importance	Corresponding Text
13	7	1.0000	• stressed out
		0.9214	• yeah I would say for the past several months
		0.8391	• I can't function as well
		0.0002	• take my dog for a walk
		0.0002	• while I was in a car accident where a drunk driver hit me and I had to
		0.0000	• I like the weather I like the beach
21	15	1.0000	• I don't know I I I developed anxiety and I freaked out you know if I think I'm going to run out of gas I get short of breath and
		0.9127	• sometimes I just give up and I don't even try anymore
		0.8360	• hello I've lost all the ability to trust and I'm numb to all feelings partly
		0.0012	• I I guess I could erase my big pts State when I was 18 a serial killer
		0.0012	• I love the weather people are generally more friendly than where I've lived on the East Coast the scenery the environment the beach the mountains the
		0.0000	• that's not my PTSD thing though if you're wondering
0	2	1.0000	• I've been feeling fine
		0.5020	• I'm pretty easy over the last two to three weeks I think there was one night or I had so much on my mind I just find it hard to fall asleep but in general I do sleep well
		0.4371	• ragging
		0.0017	• I wish that I would argue with my husband less especially in front of our kids
		0.0011	• one of my most memorable experiences in terms of travel I guess was the time that my luggage got lost in front of Vallarta and I spent the week I'm wearing my husband shorts and t-shirt
		0.0000	• I've been feeling fine this summer the work stress is still there but my kids are out of school so our household is a lot more relaxed

3.3.2.3 Best-Performing Non-Temporal Models

We conclude our non-temporal model analysis by proposing two networks. Mean-pooling network without attention achieves the best validation score among non-temporal methods. AWM /w MinMax Norm is the best-performing weighted model. We previously argued that an attention-based model could generally find good weights for embeddings, but mislead the model on edge cases. While we show that attention is not required for good representation selection, we believe it is possible for such a model could be less susceptible to overfitting and have good generalization. We also check the MAE and RMSE metrics for these two models in the validation set and observe that they are very similar. Due to these reasons, we believe the attention model should also be evaluated with the test partition and its results should be discussed.

Table 3.9: Results for best performing non-temporal models. Results are given for three metrics in both validation and test sets.

	Validation			Test		
Model	CCC	MAE	RMSE	CCC	MAE	RMSE
NT-MEAN	0.673	3.214	4.217	0.729	3.304	4.353
NT-MEAN-ATT	0.654	3.269	4.412	0.708	3.801	4.925

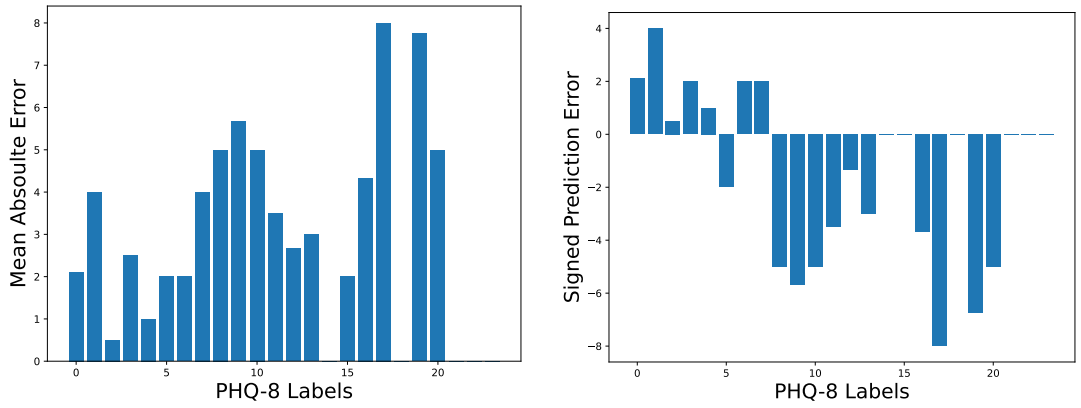
Table 3.9 presents our non-temporal model results. Both models achieve good generalization across all three metrics. NT-MEAN outperforms NT-MEAN-ATT in each metric, especially so in terms of generalization to the test set, providing evidence that the attention module in NT-MEAN-ATT is detrimental to performance. With that being said, NT-MEAN-ATT has better generalization compared to our best-performing temporal models. This provides evidence that incorporating contextual information via recurrent architectures could prove challenging, and simpler methods can perform better. Since NT-MEAN achieves the best among all investigated architectures, we also provide the detailed network hyperparameters for it in Table 3.10.

We also present the prediction error for NT-MEAN in two ways in Figure 3.3. We use MAE to understand the average error for specific labels. We can see

Table 3.10: Table of hyperparameters for the best performing NT-MEAN model

Hyperparameter Name	Selection
Embedding	all-mpnet-base-v2
# Residual Blocks	4
Residual Block Dropout Probabilities	0.4
# Linear Layers	4
Linear Layer Dropout Probability	0.5

that the model is better at detecting healthy individuals compared to depressive individuals. Our error rates have a peak around the healthy-depressive class cutoff label 10, which shows us that such individuals prove more challenging to assess. We also observe the signed prediction error in Figure 3.3-b. This error measure calculates $y_{pred} - target$ compared to the $|y_{pred} - target|$ calculation for MAE. This helps us see if our prediction is higher or lower than the target for specific PHQ-8 targets. We can see that until PHQ-8=8, we are, on average, giving higher predictions than the target, while after that point, it is the opposite. The reason for the sharp change is not clear, but it could be related to the small peak near the healthy-depression cut-off we see in Figure 3.3-a. The model may be learning a harsher threshold around such challenging individuals.



(a) Mean Absolute Error (b) Signed Prediction Error

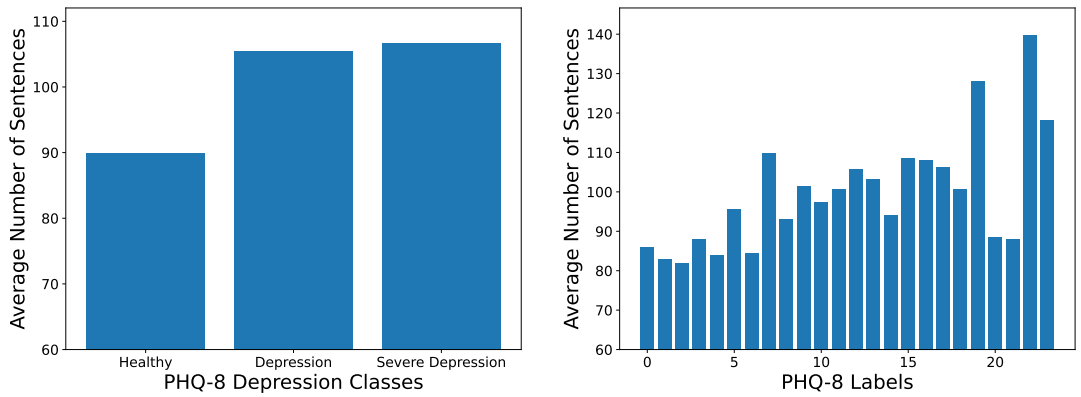
Figure 3.3: Average number of sentences per (a) depression class and (b) PHQ-8 label

3.3.3 Sentence Statistics for Depression Severity Assessment

Our experiments thus far take the semantic meaning of sentences to predict depressive behaviour. In this section, we focus solely on statistics regarding sentences rather than their meaning. During our literature reviews, we find that text modality is not well analyzed in the literature. We aim to expand the literature by connecting statistical findings with learning-based results. Every experiment is performed by running inference on already trained models, unless otherwise stated.

3.3.3.1 Sentence Count per Interview

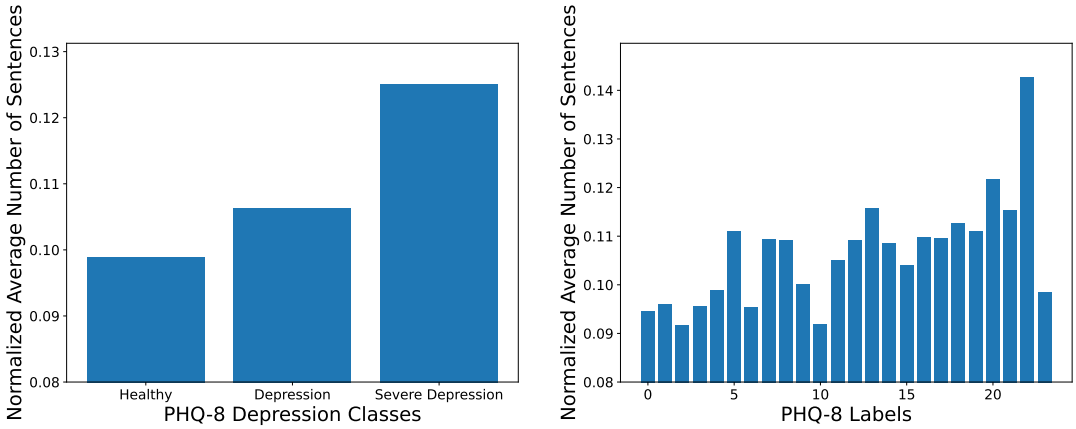
We first focus on the number of sentences uttered during an interview. Class-based inspections of the average number of sentences for depression and severe depression are quite similar, while healthy participants form fewer sentences (Fig. 3.4-a). Figure 3.4-b shows a more detailed distribution. Although the transition is smooth, we can see that the average number of sentences start to increase around PHQ-8 label 7 and 8.



(a) Depression class (b) PHQ-8 label

Figure 3.4: Average number of sentences per (a) depression class and (b) PHQ-8 label

We recall that interview lengths are also variable. Looking at these figures, it is unclear if depressed people utter more sentences because their interviews last longer, or because they simply form more sentences. To inspect this idea, we reanalyze our data by normalizing the number of sentences for each participant by the length of their interview (in seconds). Figure 3.5-a shows that class differences for severe depression is more distinct then its non-normalized counterpart. Both normalized and non-normalized analysis methods agree that depressed people form slightly more sentences.



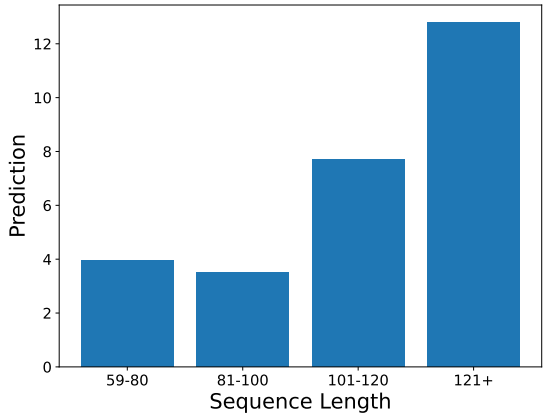
(a) Depression class

(b) PHQ-8 label

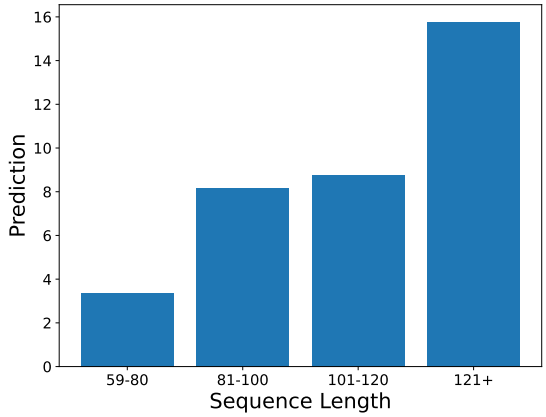
Figure 3.5: Average number of sentences per (a) depression class and (b) PHQ-8 label, normalized by interview length

It should be noted that the number of participants with severe depression (PHQ-8 $\geq$ 20) is very low, and the averages could be more biased compared to other PHQ-8 labels in our analysis.

Compared to audiovisual modalities, text modality inherently contains information regarding the amount of sentences if the text processing is done at the sentence embedding level (as we do in this thesis). The reason is that every sentence, regardless of the number of words they contain, is reduced to a single embedding. Audiovisual modalities contain information regarding the lengths of the interview (i.e. the number of datapoints is a function of interview length and the sampling rate of the modality), but it is non-trivial for a network to obtain



(a) Validation Set



(b)Test Set

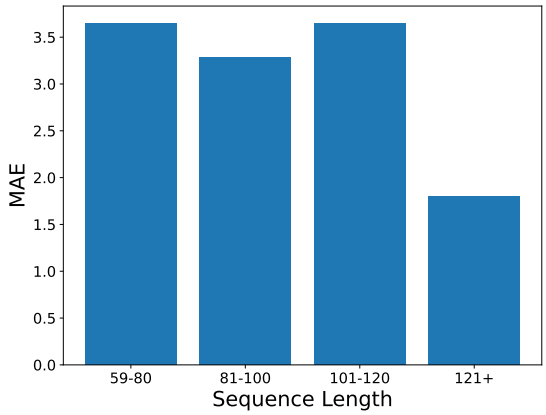
Figure 3.6: Visualization of predictions per sequence length for TS-MEAN model. The value of each bin is the average prediction for participants with that sequence length in (a) validation and (b) test sets.

sentence count information from such complex data.

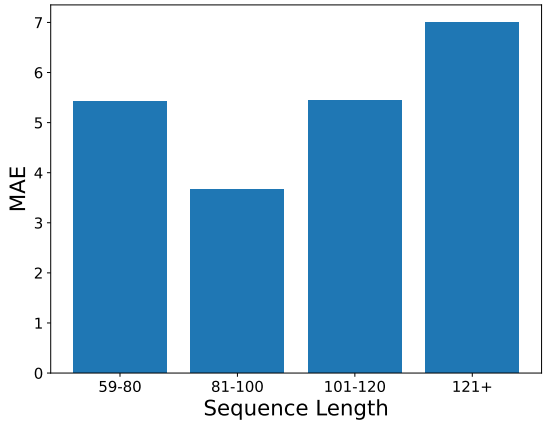
The availability of such an important marker can be part of the explanation for why text modality performs better than other modalities. But before we commit to that claim, we now provide further analysis using our trained models. Our proposed TS-MEAN model is capable of processing sequences and therefore, it is capable of learning the correlation between sequence length and depression.

In our first experiment, we examine the relationship between sequence length and model prediction. To this aim, the sequence lengths of each participant are

separated into four bins. The value of each bin is the average of corresponding predictions for participants with that sequence length. Figure 3.6 shows that there is indeed a correlation between sequence length and the predictions made by TS-MEAN. Average predictions made for the 121+ bin are more than three times the predictions for the 59-80 bin for the validation set (Fig. 3.6-a). For the test set (Fig. 3.6-b), the ratio between the same bins is more than 5-fold.



(a) Validation Set

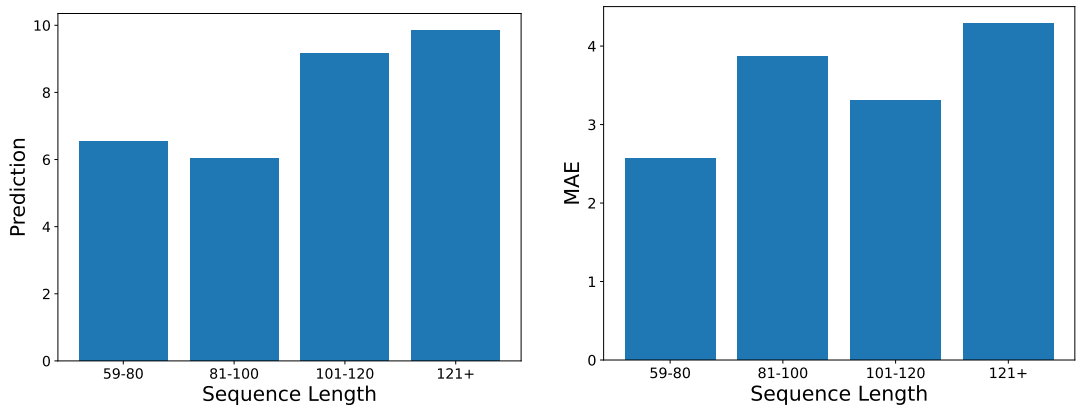


(b) Test Set

Figure 3.7: Visualization of MAE per sequence length for TS-MEAN model. The value of each bin is the average MAE for participants with that sequence length in (a) validation and (b) test sets.

While this provides evidence that our model is learning well, error values show

that this is not the case. Figure 3.7 has a similar setup to Figure 3.6, except we now show the MAE in the y-axis instead of model prediction (i.e. $|\hat{y} - y|$ instead of $\hat{y}$). At first glance, we see that the 121+ bin has a low MAE (Figure 3.7-a). This is an expected result of having high PHQ-8 predictions for the 121+ bin. What we do not expect is that while the 59-80 bin had expected prediction values, it has the highest MAE. Another contradictory finding is in Figure 3.7-b, where the MAE for 49-80 and 121+ bins are the highest ones, even though the model predictions on this set were in line with our hypothesis (Fig. 3.6-b). Comparing the error on 121+ bin for validation and test sets, we can argue that the model has bias in its predictions on high sequence lengths that do not generalize well.



(a) Predictions (b) MAE

Figure 3.8: Visualization of predictions per sequence length for NT-MEAN-ATT model. The value of each bin is the average of corresponding (a) model prediction or (b) MAE, for participants with that sequence length.

We conduct an additional experiment using our non-temporal NT-MEAN-ATT model. This model has a similar performance in the validation set with TS-MEAN (0.654 and 0.650, respectively). Consequently, changes we can see in terms of prediction and error values will be more due to the differences in learned patterns rather than differences in performance. Prediction and MAE values calculated using NT-MEAN-ATT model are presented in Figures 3.8-a and 3.8-b respectively. We see that the correlation between sequence length and model predictions is significantly less. Similarly, MAE is more distributed and does not

show a specific bias towards a certain bin. Since the non-temporal averaging we do in NT-MEAN-ATT is not capable of capturing the sequence length statistic, it appears that the model can train and generalize better by ignoring this statistic.

In this section, we pointed out a statistical finding regarding the number of sentences uttered during an interview. This finding suggested that the number of sentences is lower for healthy people and higher for depressed people. We then use a temporal model to see if this finding can be captured via learning systems. Although neural models are quite complex and it is non-trivial to pinpoint the exact inner workings of the systems, we believe our results provide good evidence that this statistic can be captured, but causes overfit behaviour.

3.3.3.2 Word Count per Sentence

During our analysis of feature importance in Section 3.3.2.2, we observe that the length of individual sentences differs for different participants. As we recall from 3.8, high feature importance can be assigned to sentences with only 2 words, as well as to sentences with around 30 words. Before we analyze the relationship of word count with depression classes, we look at its effect on general performance. Using the trained version of our model NT-MEAN, we reevaluate the validation set. The model is not trained again or finetuned for each individual word count configuration, but only reevaluated. Manual observation of the dataset shows that in cases where there is not much time between sentences, speech-to-text AI transcribes some answers by the participant as long sentences. Separation of these sentences is non-trivial, and we believe that the benefits of keeping such sentences as they are outweigh potential problems that can occur if we are to separate them.

We conduct two experiments: subtractive and additive. Subtractive word count experiment includes all sentences whose word counts are bigger than our word count limiter variable L. Formally, a sentence S is included in evaluation if the predicate $|S| > L$ is true, where $|S|$ is the number of words in S. Conversely, additive experiments include sentences that obey $|S| \leq L$. Figure 3.9 shows

the results of these experiments. We only experiment up to $L = 15$ because subtractive experiments become noisy after that point, due to the lack of such long sentences for each participant.

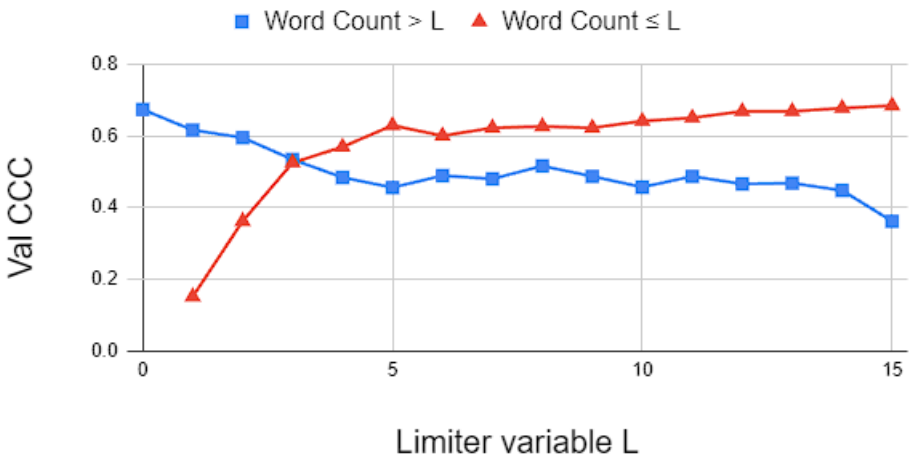


Figure 3.9: Additive (Red-Triangle) and Subtractive (Blue-Square) experiments on word count, using NT-MEAN model. For Limiter variable L, Additive experiment evaluates NT-MEAN on all sentences with word count $\leq L$. Subtractive experiments do the same for word count $> L$.

Our performance drops continuously when we do not include sentences with word counts up to but excluding 5. After that point, the performance has fluctuating changes, but they are not as close to the change and consistency we see with the first 5. Similarly, performance continuously increases as we include sentences with word counts up to and including 5, but the rest is not consistent. This analysis shows that model performance highly depends on shorter sentences. Before we can clearly understand the effect of individual sentences, we examine the percentages of sentences with a certain number of words in our data.

Figure 3.10 tells us that sentences with at most 6 words ($L = 6$) make up, coincidentally, almost %50 of our data. Since this point separates our data equally, we compare our two experiments using this data point. Comparison of additive and subtractive experiments show that at point $L = 6$, additive experiment with CCC value of 0.6 performs significantly better than its subtractive counterpart with CCC value of 0.489. Since both these experiments contain very similar amounts of data, we argue that the difference likely depends on whether we include shorter

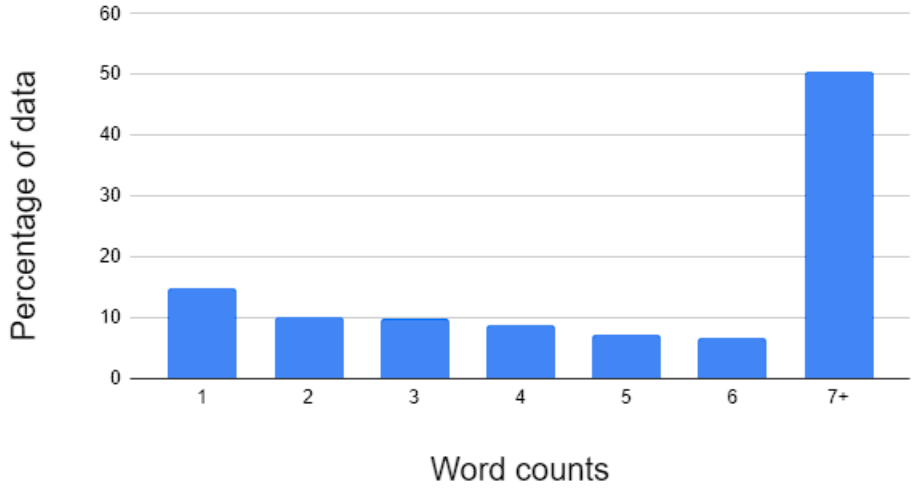


Figure 3.10: Presentation of the percentage of sentences with a given word count. Sentences with 7 and more words make up %50 of the dataset.

sentences.

Our discussions up to this point originated from a trained model. We examine their applicability to learning by training two models. Each of these models performs its training and validation using half of the data, one uses sentences such that $|S| > 6$ while the other uses $|S| \leq 6$. As depicted in Table 3.11, both models perform significantly worse than NT-MEAN, which was trained with all sentences (i.e. with $|S| > 0$). Also, both models perform similarly using their respective data. Although our data separation took into consideration the information gain from data on both sides, it seems that losing close to %50 of data for training is something we cannot ignore. Since the model with the $|S| \leq 6$ portion contains more informative sentences, one would expect this model to perform better. We argue that not including less informative sentences could be detrimental to training, irrespective of the information loss argument. It is known that some amount of noise in training is good for regularization and reduces overfit in some networks [79, 80]. Some architectures purposefully focus their training on optimizing hard examples [81], most popularly in the case of most novel deep metric learning networks [82]. In our case, less informative sentences could act as regularization by providing hard examples to the network. This way, the network does not overfit by using only specific information. In

a sense, more informative sentences are easier to learn, probably because their semantic content is more obviously a member of one class. It is no coincidence that shorter sentences have easier to learn semantic content since longer sentences are more convoluted and may contain more than one emotion. This also points out a problem with our simple word count-based approach, as the effects of longer sentences with a single emotion are completely omitted. We can see evidence of this behaviour by inspecting important sentence examples in Table 3.8. We can see that sentences that were deemed important by the model focus on a certain topic or emotion, compared to unimportant sentences. All in all, it seems that certain sentences are not very informative during inference, and they can even hinder prediction at times, but we should use as much data as possible during training. Admittedly, this finding is not out of the ordinary for neural networks, but we state it regardless to explain our findings better.

Table 3.11: Validation CCC results for models trained in different data portions. $|S| > 6$ is the NT-MEAN model, while the other two models use sentences with more than 6 words and less than or equal to 5 words, respectively.

Data Portion	Val CCC
$\|S\| > 0$	0.673
$\|S\| > 6$	0.568
$\|S\| \leq 6$	0.575

To the best of our knowledge, a comparative analysis of model reliance on word counts does not exist within the literature. Therefore, it is unclear if our findings regarding word count and information gain are specific to depression severity assessment task, or to natural language processing in general. It is known that shorter sentences are generally less convoluted in their meaning and more compressible. This could mean that more salient sentence embeddings can be created from shorter sentences, as they commonly focus on a specific topic and emotion. This observation can bridge our findings about depression with the broader natural language processing literature.

3.3.4 Assessment of Sentence Embeddings

In Section 3.2.2, we present our initial sentence embedding and explain our reasoning for its selection. This section presents different sentence embeddings and analyzes their performance. A summary of the embeddings can be found in Table 2.1.

The use of sentiment analysis for depression severity assessment is studied within the literature, and shows promising results [83, 84, 85, 86]. Behavioral studies also agree that sentiment is a good predictor of self-reported depression [87, 88]. Most of the studies in this area focus on sentiment analysis but do not necessarily use improved sentiment embeddings. To extend upon this area, we select two sentiment networks to analyze in this section.

Table 3.12: Results for training NT-MEAN with different embeddings

Embedding	Val CCC	Test CCC
all-mpnet-base-v2 [68]	0.673	0.729
SiEBERT [69]	0.531	0.315
CardiffNLP-Sent [70]	0.654	0.568

We agree that there exists a correlation between sentiment and depression. With that being said, we should be careful how sentiment information affects training. Similar to our arguments in Section 3.3.3.2, it is possible for well-trained sentiment embeddings to present only sentiment information to the network. This can happen since frozen sentiment sentence embeddings are the only source of raw information the network gets. This can cause an overfit to sentiment information and the network can have a hard time exploring depression cues. Due to these reasons, we expect embeddings like all-mpnet-base-v2, which perform well on many semantic tasks, to perform better.

Although SiEBERT is a good sentiment analysis network in its own regard, we see that it performs significantly worse compared to CardiffNLP-Sent in NT-MEAN model. We do not think this is due to differences in training for the embeddings since SiEBERT is finetuned on a well-trained RoBERTa checkpoint.

Comparatively, CardiffNLP-Sent is trained on fewer data. It is possible that tweet data may be closer in nature to the interview data we use. People often use Twitter to share their thoughts or state of mind. Previous studies on depression analysis on social media also agree with this observation [67]. We compare this to the more formal datasets used by RoBERTa checkpoint (News articles, Wikipedia) or SiEBERT finetuning datasets that include less psychological expressions (reviews for DVDs, businesses, electronics etc.). This argument can also explain the performance of all-mpnet-base-v2, since more than %60 of its training data comes from Reddit comments.

In this section, we compare different sentence embeddings and their performance. Through our analysis, we show that embeddings that perform well on general semantic information-related tasks are better for depression severity assessment compared to sentiment embeddings. We also argue that the nature of training data also has an effect on the performance of depression models, as we see better performance from embeddings that contain sentences relatively similar to the ones we see in our depression dataset.

3.3.5 Comparison with Other Methods

We finalize our analysis by comparing the performance of our best-performing model, NT-MEAN, to other studies within the literature. Table 3.13 is a compilation of studies from the literature that use the AVEC 2019 dataset. Listed modalities describe all modalities that the corresponding study explores, and is not necessarily what their final model is based on). In the case where multiple models are proposed, the one with the higher test set performance is chosen. To increase the comparability of our model for future works, we provide both validation and test set evaluations for three metrics. To the best of our knowledge, we are the only study that does not utilize a recurrent architecture in their proposed model. Comparing our performance, we see that NT-MEAN improves the state of the art by Ray et al. [42] on all metrics. The relative improvements are %8.8 for CCC, %8.7 for RMSE, and %21.8 for MAE.

Table 3.13: Details regarding the modalities and performance of other studies in the literature, to the best of our knowledge. Modalities are abbreviated as A = Audio, V = Vision, T = Text.

Model	Modalities	Year	Val CCC	Val RMSE	Val MAE	Test CCC	Test RMSE	Test MAE
NT-MEAN	T	2022	0.673	4.22	3.21	**0.729**	**4.35**	**3.30**
Sun et al. [27]	AVT	2022	-	-	-	0.583	-	4.37
Saggu et al. [33]	AVT	2022	0.662	4.32	-	0.457	5.36	-
Sun et al. [34]	AV	2021	0.733	3.78	-	-	-	-
Yin et al. [40]	AVT	2019	0.402	4.94	-	0.442	5.50	-
Makiuchi et al. [41]	AT	2019	0.696	3.86	-	0.403	6.11	-
Kaya et al. [47]	AT	2019	0.481	-	-	0.344	-	-
Ray et al. [42]	AVT	2019	-	4.37	-	0.670	4.73	4.02
Ringeval et al. [32]	AV	2019	0.336	-	-	0.111	-	-

Chapter 4

Conclusion

In this thesis, we have proposed temporal and non-temporal architectures to predict PHQ-8 depression scores. Compared to an overwhelming majority of studies in the literature, we have only used text modality as a single modal. Our non-temporal model NT-MEAN has improved the state of the art by %8.8, using a simpler architecture. To shed more light on the inner workings of this non-temporal network, we have extracted sentences that are deemed important by the network by examining network activations. Through this, we have shown that our model successfully learns to select important representations. As we have compared temporal and non-temporal architectures, we have realized that temporal relationships of individual sentences are tenuous at best, and not using the temporal information is better for performance. Similar comparison experiments have been conducted on our attention network and residual blocks. They have also shown that residual blocks can act as scalers for sentence embeddings, as an attention network would.

We have also expanded the literature on natural language processing and depression severity assessment by presenting our empirical findings regarding participant sentence statistics, such as the number of uttered sentences and word count of sentences. We have displayed that a well-trained model shows less reliance on longer sentences. To put it in another way, longer sentences are not

as informative for depression assessment compared to shorter ones. We believe this is because longer sentences are usually more convoluted, and embeddings are more saturated in that case. Comparatively, shorter sentences usually focus on a very specific semantic information or emotion and are therefore more salient. Our work on sentences has also demonstrated that there is a correlation between the number of sentences an individual forms during their therapy session, and their depression severity. This correlation, however, is detrimental to the training of temporal models. Our discussions regarding this phenomenon can explain the generalization problems we see within the literature. Furthermore, we have provided our analysis on using different sentence embeddings and their possible relationship with the depression assessment task. We have argued that while sentiment information does have its merits, semantic information could be a better comparator of depressive behaviour. Training data used for the sentiment embeddings could also play a part in its performance.

The nature of data and its effects on our results should also be considered. We point out the sample population of our data includes veterans of the U.S. armed forces. Considering this information, the possibility of learning very specific cues arises. We argue that such a bias does not exist in our models, by inspecting the model-assigned feature importance to sentences. We observe that important sentences are mostly, if not always, general cues of depression and are not related to war trauma. We also see that PTSD-related sentences are assigned lower importance values. When considering text modality, the performance of different languages should also be considered. While some studies observe that there could be slight cultural differences in depression expression [89, 90], the effect of using different languages is not clear. As Yilmaz and Kaynak et al. [91] show for sentiment analysis task, using cross-language embeddings and training multiple languages together can improve single-language results. We believe that linguistic patterns of depression could have similarities across different languages, and a similar scheme could be beneficial for the depression severity assessment task.

We suspect that one of the biggest shortcomings of depression studies is that they are not longitudinal, but rather single-instance. Model-wise, given the low

amount of data that most depression studies have, it is very hard to create generalizable models and draw general conclusions. This is especially true if we want to analyze the behaviour of participants over a period of time, which is usually a necessity for the clinical diagnosis of mental illnesses. This is possibly one of the reasons why our non-temporal model performed better than temporal ones. Per our arguments and data inspections, we have argued that therapy sessions contain episodes of conversation on certain topics. It might be possible to leverage this information with longitudinal data or with longer individual sessions. We consider that shorter interviews still have merit and applications. As an example, an automated chatbot can ask questions to an individual on diverse topics and possibly mimic the short interactions we have in single-instance therapy sessions.

There also exists significant label imbalance within datasets, and any attempt at data augmentation techniques to reduce the imbalance is non-trivial due to the complexity and variability of the input sequences. As another discussion point, experts can have a hard time diagnosing individuals in such short periods of evaluation. Coupled with the fact that individual differences can create great variability for certain PHQ-8 ranges, some amount of noise is always expected. This is especially true for the dataset of this thesis since the labels are self-appointed. What people say and feel may be quite different, which could cause a social-desirability bias when text modality is used with self-appointed labels. In such a case, audiovisual modalities could do a better job of understanding the true feeling by capturing honest signals [92]. Specific to our dataset, we reiterate that the test partition was formed entirely by participants who were interviewed by an automated agent, i.e. the interviewer saw an artificial avatar and responded to questions presented by an artificial intelligence. We believe, as the creators of the dataset do, that the partitions can have differences in their distributions. The behaviour of participants can change drastically, resulting in a different sample population than the one used in training. Our results suggest that temporal models may be more inclined to learn patterns of conversation which are dependent on the nature of the interviewer, due to their in-order sequence processing capabilities. In contrast, the non-temporal approach has resulted in more pattern-agnostic models that focus more on the importance of individual depression cues.

It should be noted that this finding could be dependent on the nature of the data, and may not hold for general comparisons of temporal and non-temporal architectures. We have not employed any automatic or manual data cleaning. It is possible that many seemingly informative sentences were not selected by our model due to this reason, and future work could surely benefit from data cleaning attempts.

Motivated by the properties of text modality, we hope that the discussions we have started and improvements we have proposed in this thesis will open new directions for feature work on depression assessment. We have high hopes that through such conversations, we will understand this insidious illness better.

Bibliography

[1] M. Hamilton, "A rating scale for depression," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 23, no. 1, pp. 56–62, 1960.

[2] K. Kroenke, T. Strine, R. Spitzer, J. Williams, J. Berry, and A. Mokdad, "The phq-8 as a measure of current depression in the general population," *Journal of affective disorders*, vol. 114, pp. 163–73, 09 2008.

[3] "Covid-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide." https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide. Accessed: 2022-09-06.

[4] A. Gupta, P. Mathur, S. Bijawat, and A. Dadheech, "A novel work on analyzing stress and depression level of indian population during covid-19," *Recent Advances in Computer Science and Communications*, vol. 13, 11 2020.

[5] H. Seens, Z. Lu, J. Fraser, J. Macdermid, D. Walton, and R. Grewal, "An intersectional approach to identifying factors associated with anxiety and depression following the covid-19 pandemic," *Scientific Reports*, vol. 12, p. 11393, 07 2022.

[6] L. M. S. Passos, C. Murphy, R. Chen, M. Santana, and G. Passos, "Association of positive and negative feelings with anxiety and depression symptoms among computer science students during the covid-19 pandemic," in *Anais do II Simpósio Brasileiro de Educação em Computação*, (Porto Alegre, RS, Brasil), pp. 50–56, SBC, 2022.

[7] World Health Organization, "Depression and other common mental disorders: global health estimates," technical report, 2017. Licence: CC BY-NC-SA 3.0 IGO.

[8] I. Jones and M. C. Pansa, "Some nonverbal aspects of depression and schizophrenia occurring during the interview," *The Journal of Nervous and Mental Disease*, vol. 167, p. 402–409, 1979.

[9] W. Gaebel and W. Wölwer, "Facial expressivity in the course of schizophrenia and depression," *European Archives of Psychiatry and Clinical Neuroscience*, vol. 254, pp. 335–342, 2004.

[10] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency, "Automatic nonverbal behavior indicators of depression and ptsdw the effect of gender," *Journal on Multimodal User Interfaces*, vol. 9, 03 2014.

[11] Z. Ambadar, J. W. Schooler, and J. F. Cohn, "Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions," *Psychological Science*, vol. 16, no. 5, pp. 403–410, 2005. PMID: 15869701.

[12] A. Nilsonne, "Speech characteristics as indicators of depressive illness," *Acta Psychiatrica Scandinavica*, vol. 77, no. 3, p. 253–263, 1988.

[13] M. Landau, "Acoustical properties of speech as indicators of depression and suicidal risk," *Vanderbilt Undergraduate Research Journal*, vol. 4, 06 2008.

[14] J. Wang, L. Zhang, T. Liu, W. Pan, B. Hu, and T. Zhu, "Acoustic differences between healthy and depressed people: a cross-situation study," *BMC Psychiatry*, vol. 19, 10 2019.

[15] J. K. Darby, N. Simmons, and P. A. Berger, "Speech and voice parameters of depression: A pilot study," *Journal of Communication Disorders*, vol. 17, no. 2, pp. 75–85, 1984.

[16] J. Leff and E. Abberton, "Voice pitch measurements in schizophrenia and depression," *Psychological Medicine*, vol. 11, no. 4, p. 849–852, 1981.

[17] S. Kuny and H. Stassen, "Speaking behavior and voice sound characteristics in depressive patients during recovery," *Journal of Psychiatric Research*, vol. 27, no. 3, pp. 289–307, 1993.

[18] L. Flyckt, E. Hassler, L. Lotfi, I. Krakau, and G. H. Nilsson, "Clinical cues for detection of people with undiscovered depression in primary health care: a case–control study," *Primary Health Care Research Development*, vol. 15, no. 3, p. 324–330, 2014.

[19] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–7, 2009.

[20] G. Edwards, C. Taylor, and T. Cootes, "Interpreting face images using active appearance models," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 300–305, 1998.

[21] G. McIntyre, R. Göcke, M. Hyett, M. Green, and M. Breakspear, "An approach for automatically measuring facial activity in depressed subjects," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–8, 2009.

[22] H. Dibeklioglu, Z. Hammal, and J. Cohn, "Dynamic multimodal measurement of depression severity using deep autoencoding," *IEEE Journal of Biomedical and Health Informatics*, vol. PP, pp. 1–1, 03 2017.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[24] K. Chlasta, K. Wołk, and I. Krejtz, "Automated speech-based screening of depression using deep convolutional neural networks," *Procedia Computer Science*, vol. 164, pp. 618–628, 2019. CENTERIS 2019 - International Conference on ENTERprise Information Systems / ProjMAN 2019 - International Conference on Project MANagement / HCist 2019 - International

Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2019.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, p. 1735–1780, nov 1997.

[26] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning, December*, 2014.

[27] H. Sun, H. Wang, J. Liu, Y.-W. Chen, and L. Lin, "Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation," arXiv:2207.14087, 2022.

[28] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 578–584, 2018.

[29] M. Chen, X. Xiao, B. Zhang, X. Liu, and R. Lu, "Neural architecture searching for facial attributes-based depression recognition," arXiv:2201.09799, 2022.

[30] 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.

[31] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, 2016.

[32] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, AVEC '19, (New York, NY, USA), p. 3–12, Association for Computing Machinery, 2019.

[33] G. S. Saggu, K. Gupta, and K. V. Arya, "Depressnet: A multimodal hierarchical attention mechanism approach for depression detection," *International Journal of Engineering Sciences*, 2022.

[34] H. Sun, J. Liu, S. Chai, Z. Qiu, L. Lin, X. Huang, and Y.-W. Chen, "Multimodal adaptive fusion transformer network for the estimation of depression level," *Sensors (Basel, Switzerland)*, vol. 21, 2021.

[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] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, 2017.

[37] A. Pampouchidou, P. G. Simos, K. Marias, F. Meriaudeau, F. Yang, M. Pediaditis, and M. Tsiknakis, "Automatic assessment of depression based on visual cues: A systematic review," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 445–470, 2019.

[38] L. He, M. Niu, P. Tiwari, P. Marttinen, R. Su, J. Jiang, C. Guo, H. Wang, S. Ding, Z. Wang, X. Pan, and W. Dang, "Deep learning for depression recognition with audiovisual cues: A review," *Information Fusion*, vol. 80, pp. 56–86, 2022.

[39] U. Yadav and A. K. Sharma, "Review on automated depression detection from audio visual clue using sentiment analysis," in *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 1462–1467, 2021.

[40] S. Yin, C. Liang, H. Ding, and S. Wang, "A multi-modal hierarchical recurrent neural network for depression detection," *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019.

[41] M. R. 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.

[42] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level attention network using text, audio and video for depression prediction," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, AVEC '19, (New York, NY, USA), p. 81–88, Association for Computing Machinery, 2019.

[43] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 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.

[44] P. Baki, H. Kaya, E. çiftçi, H. Güleç, and A. A. Salah, "A multimodal approach for mania level prediction in bipolar disorder," *IEEE Transactions on Affective Computing*, pp. 1–13, 2022.

[45] J. Zinken, K. Zinken, J. C. Wilson, L. Butler, and T. Skinner, "Analysis of syntax and word use to predict successful participation in guided self-help for anxiety and depression," *Psychiatry Research*, vol. 179, no. 2, pp. 181–186, 2010.

[46] S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition Emotion - COGNITION EMOTION*, vol. 18, pp. 1121–1133, 12 2004.

[47] H. Kaya, D. Fedotov, D. Dresvyanskiy, M. Doyran, D. Mamontov, M. Markitantov, A. A. Akdag Salah, E. Kavcar, A. Karpov, and A. A. Salah, "Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, AVEC '19, (New York, NY, USA), p. 27–35, Association for Computing Machinery, 2019.

[48] J. Ye, Y. Yu, Q. Wang, W. Li, H. Liang, Y. Zheng, and G. Fu, "Multi-modal depression detection based on emotional audio and evaluation text," *Journal of Affective Disorders*, vol. 295, pp. 904–913, 2021.

[49] N. Aloshban, A. Esposito, and A. Vinciarelli, "What you say or how you say it? depression detection through joint modeling of linguistic and acoustic aspects of speech," *Cognitive Computation*, 02 2021.

[50] C. Yang, X. Lai, Z. Hu, Y. Liu, and P. Shen, "Depression tendency screening use text based emotional analysis technique," *Journal of Physics: Conference Series*, vol. 1237, p. 032035, 06 2019.

[51] S. Gopchandani, "Using word embeddings to explore the language of depression on twitter," Ph.D Thesis, The University of Vermont, 2019. [Online]. Available: https://www.proquest.com/dissertations-theses/using-word-embeddings-explore-language-depression/docview/2209772201/se-2.

[52] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, (Red Hook, NY, USA), p. 3111–3119, Curran Associates Inc., 2013.

[53] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Takefuji, "Wikipedia2vec: An optimized tool for learning embeddings of words and entities from wikipedia," *ArXiv*, vol. abs/1812.06280, 2018.

[54] 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.

[55] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil,

"Universal sentence encoder for English," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Brussels, Belgium), pp. 169–174, Association for Computational Linguistics, Nov. 2018.

[56] S. A. Qureshi, M. Hasanuzzaman, S. Saha, and G. Dias, "The verbal and non verbal signals of depression - combining acoustics, text and visuals for estimating depression level," *ArXiv*, vol. abs/1904.07656, 2019.

[57] 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.

[58] C. Yang, X. Lai, Z. Hu, Y. Liu, and P. Shen, "Depression tendency screening use text based emotional analysis technique," *Journal of Physics: Conference Series*, vol. 1237, p. 032035, 06 2019.

[59] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1746–1751, Association for Computational Linguistics, Oct. 2014.

[60] D. N. Milne, G. Pink, B. Hachey, and R. A. Calvo, "CLPsych 2016 shared task: Triaging content in online peer-support forums," in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, (San Diego, CA, USA), pp. 118–127, Association for Computational Linguistics, June 2016.

[61] S. Pradhan, N. Elhadad, W. Chapman, S. Manandhar, and G. Savova, "SemEval-2014 task 7: Analysis of clinical text," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, (Dublin, Ireland), pp. 54–62, Association for Computational Linguistics, Aug. 2014.

[62] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013 - the continuous audio/visual emotion and depression recognition challenge," pp. 3–10, 10 2013.

[63] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014 - 3d dimensional affect and depression recognition challenge," *AVEC 2014 - Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, Workshop of MM 2014*, pp. 3–10, 11 2014.

[64] 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), pp. 3123–3128, European Language Resources Association (ELRA), May 2014.

[65] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, (New York, NY, USA), p. 47–56, Association for Computing Machinery, 2013.

[66] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health," in *ICWSM*, 2011.

[67] S. Yadav, J. Chauhan, J. P. Sain, K. Thirunarayan, A. Sheth, and J. Schumm, "Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework," in *Proceedings of the 28th International Conference on Computational Linguistics*, (Barcelona, Spain (Online)), pp. 696–709, International Committee on Computational Linguistics, Dec. 2020.

[68] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 11 2019.

[69] 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*, 2022.

[70] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, (Online), pp. 1644–1650, Association for Computational Linguistics, Nov. 2020.

[71] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," *arXiv preprint arXiv:2004.09297*, 2020.

[72] 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.

[73] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[74] L. I. Lin, "A concordance correlation coefficient to evaluate reproducibility.," *Biometrics*, vol. 45 1, pp. 255–68, 1989.

[75] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), pp. 8024–8035, Curran Associates, Inc., 2019.

[76] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp;*

Data Mining, KDD '19, (New York, NY, USA), p. 2623–2631, Association for Computing Machinery, 2019.

[77] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, p. 448–456, JMLR.org, 2015.

[78] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[79] S. Sukhbaatar, J. Bruna, M. Paluri, L. D. Bourdev, and R. Fergus, "Training convolutional networks with noisy labels," *arXiv: Computer Vision and Pattern Recognition*, 2014.

[80] M. Zhou, T. Liu, Y. Li, D. Lin, E. Zhou, and T. Zhao, "Towards understanding the importance of noise in training neural networks," *ArXiv*, vol. abs/1909.03172, 2019.

[81] Q. Dong, S. Gong, and X. Zhu, "Class rectification hard mining for imbalanced deep learning," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1869–1878, 2017.

[82] M. Bucher, S. Herbin, and F. Jurie, "Hard negative mining for metric learning based zero-shot classification," in *ECCV Workshops*, 2016.

[83] M. R. Islam, A. Kabir, A. Ahmed, A. Kamal, H. Wang, and A. Ulhaq, "Depression detection from social network data using machine learning techniques," *Health Information Science and Systems*, vol. 6, p. 8, 08 2018.

[84] 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.

[85] A. U. Hassan, J. Hussain, M. Hussain, M. Sadiq, and S. Lee, "Sentiment analysis of social networking sites (sns) data using machine learning approach for the measurement of depression," in *2017 International Conference on*

Information and Communication Technology Convergence (ICTC), pp. 138–140, 2017.

[86] X. Tao, X. Zhou, J. Zhang, and J. Yong, "Sentiment analysis for depression detection on social networks," in *Health information science and systems*, pp. 807–810, 12 2016.

[87] J. Havigerová, J. Haviger, D. Kučera, and P. Hoffmannová, "Text-based detection of the risk of depression," *Frontiers in Psychology*, vol. 10, p. 513, 03 2019.

[88] T. Liu, J. Meyerhoff, J. C. Eichstaedt, C. J. Karr, S. M. Kaiser, K. P. Kording, D. C. Mohr, and L. H. Ungar, "The relationship between text message sentiment and self-reported depression," *Journal of Affective Disorders*, vol. 302, pp. 7–14, 2022.

[89] K. Loveys, J. Torrez, A. Fine, G. Moriarty, and G. Coppersmith, "Cross-cultural differences in language markers of depression online," pp. 78–87, 01 2018.

[90] L. Kwakkenbos, E. Arthurs, F. H. J. van den Hoogen, M. Hudson, W. G. J. M. van Lankveld, M. Baron, C. H. M. van den Ende, B. D. Thombs, and for the Canadian Scleroderma Research Group, "Cross-language measurement equivalence of the center for epidemiologic studies depression (ces-d) scale in systemic sclerosis: A comparison of canadian and dutch patients," *PLOS ONE*, vol. 8, pp. 1–8, 01 2013.

[91] S. F. Yilmaz, E. B. Kaynak, A. Koç, H. Dibeklioğlu, and S. S. Kozat, "Multi-label sentiment analysis on 100 languages with dynamic weighting for label imbalance," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2021.

[92] A. S. Pentland, *Honest Signals: How They Shape Our World*. The MIT Press, 2008.