SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
http://saucis.sakarya.edu.tr/
e-ISSN: 2636-8129
Publisher: Sakarya University
Vol. 7, No. 1, 2024, 77-91
DOI: 10.35377/saucis...1381522
SAUCIS
SAKARYA UNIVERSITY
RESEARCH ARTICLE

Systematic analysis of speech transcription modeling for reliable assessment of depression severity

Ergün Batuhan Kaynak [1], Hamdi Dibeklioğlu [2]

[1] Department of Computer Engineering, Bilkent University, Ankara, Türkiye
[2] Department of Computer Engineering, Bilkent University, Ankara, Türkiye

Corresponding author:
Ergün Batuhan Kaynak,
Department of Computer Engineering,
Bilkent University, Ankara, Turkey
ebatuhankaynak@gmail.com

Article History:
Received: 26.10.2023
Accepted: 22.03.2024
Published Online: 22.03.2024

ABSTRACT
In evaluating the severity of depression, we rigorously investigate a segmented deep learning framework that employs speech transcriptions for predicting levels of depression. Within this framework, we examine the effectiveness of well-known deep learning models for generating useful features for gauging depression. We validate the chosen models using the openly accessible Extended Distress Analysis Interview Corpus (E-DAIC) as a dataset. Through our findings and analytical commentary, we demonstrate that valuable features for depression severity estimation can be achieved without leveraging the sequential relationships among textual descriptors. Specifically, temporal aggregation of latent representations surpasses the current best-performing methods that utilize recurrent models, exhibiting an 8.8% improvement in Concordance Correlation Coefficient (CCC).

Keywords: Depression severity assessment, Text analysis, Deep learning, Speech transcription

1. Introduction

Depression is a mental disorder that negatively affects the feelings, behaviors, 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 the 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]. In this study, we use a dataset with PHQ labels to evaluate our architectures.

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% according to recent statistics by World Health Organization[1]. In light of this, many studies are conducted to further investigate the effects of the pandemic and depression [3]. 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. Advances in automatic assessment of depression severity would lead to helping over 300 million [4] people suffering from depression and even save their lives.

[1] WHO (2022). COVID-19 Pandemic [online] https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide. [accessed 09 09 2023]

Cite as: Kaynak E. B. and H. Dibeklioğlu, "Systematic analysis of speech transcription modeling for reliable assessment of depression severity", *Sakarya University Journal of Computer and Information Sciences*, vol. 7, no. 1, pp. 77-91, 2024. doi: 10.35377/saucis...1381522

To this aim, this study systematically analyzes different network architectures for depression severity assessment from speech transcriptions and discusses the implications of empirical results. Both temporal and non-temporal modeling approaches are investigated. Our experimental results show that the use of temporal pooling of latent representations, rather than recurrent modeling, provides state-of-the-art performance.

2. Related Work

Most depression severity assessment literature focuses on audiovisual modalities and their fusion. In contrast, this study focuses text modality as a single modal.

Studies show that many syntactic and statistical measures regarding language correlate with depression, such as the decrease in syntactic complexity [5] or the use of first-person pronouns [6]. Depressed individuals also display slower speech rates and longer pauses [7, 8]. 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 [9]. Ye et al. choose to use the top 10 most frequent words to differentiate between healthy and depressed groups [10].

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 [11, 12] use Word2vec [13] and its variants to extract representations. Recently, more powerful sentence embeddings are utilized with Universal Sentence Encoder [14-16] or BERT [17] 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 [15, 16, 18, 19]. Differently, Yang et al. [12] format the text as a two-dimensional matrix of words and embeddings and process it using a TextCNN [20] variant with k-Max-pooling. This study also investigates the performance of temporal architectures on modeling text representations. Contrary to the literature, we also propose non-temporal modeling of sentences.

Even though it is not used as much as audiovisual modalities, text modality has several advantages. 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 [21]. Singh et al. show that for attention enabled ensemble learning models are effective methods of detecting depression symptoms on social media data, such as ones from Reddit and Twitter [22]. Some studies also combine text with social media metadata, such as posting habits [23, 24]. Even in such recent studies, non-deep learning based models, such as decision trees [23] are utilized. We also see this in the case of Liu et al., in which case the authors use an ensemble of support vector machine, naive bayes and regression approaches to detect depression cues [24]. 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.

3. Methodology

In this study, we propose a modular natural language processing pipeline to predict a PHQ-8 score, given sequential sentences of an individual. Figure 1 illustrates the schema for the proposed modular pipeline. Remaining subsections within this section 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 investigated architectures in Section 3.2. These architectures are formed according to ablation studies performed in Section 4.

Main processing flow of our pipeline is as follows: First, each sentence for a 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 3.1.1 describes the method we follow to achieve this conversion. These sentence embeddings can go through an individual processing step using residual blocks (Section 3.1.2) and/or the attention module (Section 3.1.3). 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 3.1.4 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 prediction, using linear regression layers.

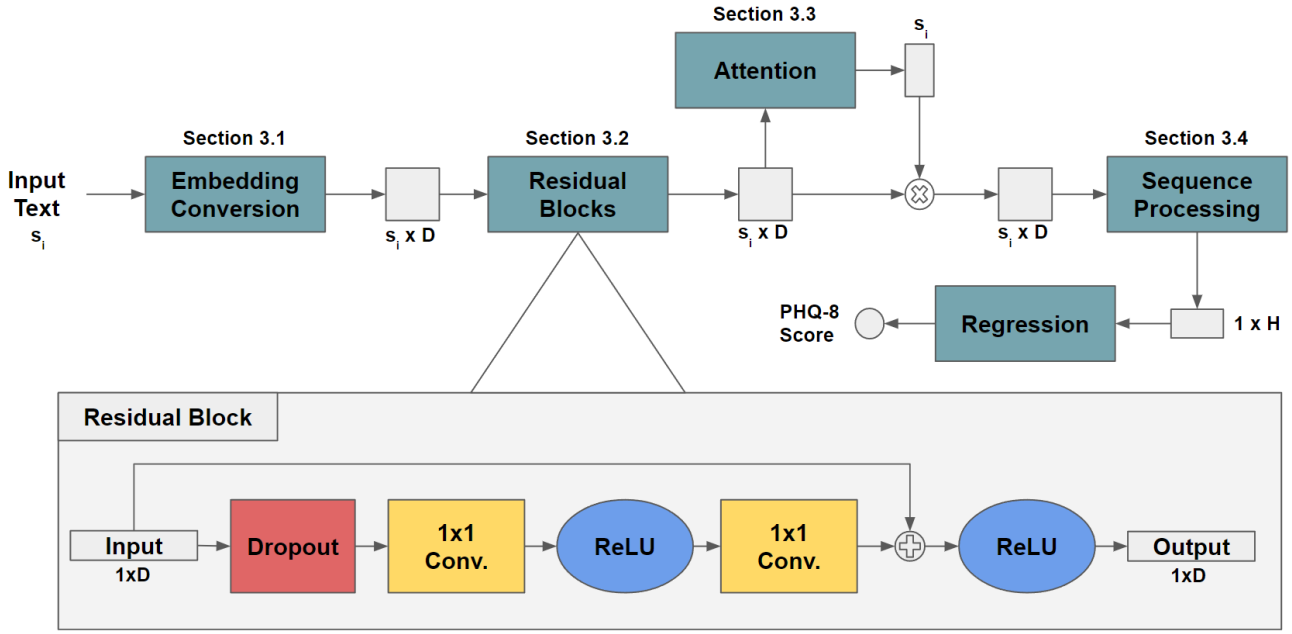


Figure 1: Overview of the proposed pipeline and the residual blocks. Data representations at several stages are depicted with their shapes. Shapes are given for a single participant and not batched input. For participant p_i , s_i is the number of sentences, D is the size of the sentence embedding (depends on embedding choice). H is the output dimension of the sequence processing module ($H = D$ for non-temporal modules, but it is a hyperparameter for temporal ones). Dropout is used at the beginning of each block for regularization. Two 1x1 convolutions with ReLU activation are used to process sentence embeddings and the skip connection happens after the second convolution, before activation.

3.1 Architectural Modules

3.1.1 Transcript Representation

To process text data, we first 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 $P^{(i)} \epsilon\, s_i \times D$, where s_i is the number of sentences and D is the embedding size. We define the single sentence embedding uttered within the time period t as x_t. Embedding size D depends on the choice of embedding, but it does not create any considerations while building our architectures.

We start by introducing our main embedding, all-mpnet-base-v2 [25]. This embedding is a finetuned version of Microsoft's mpnet-base model [26]. 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.

It should be noted that the embedding architecture is not appended to our end-to-end depression severity assessment network. To elaborate, output sentence embeddings are frozen, and the error from our 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. Due to this, it is worth bearing in mind that our performance is reliant on the performance of our chosen embedding.

3.1.2 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. These residual blocks use the residual learning idea

from [27], where we add the block input to the output of that block. This skip-connection from the input to the end of the last convolutional layer, as seen in Figure 1 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.

3.1.3 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 $a_t \epsilon R$ is calculated by regressing a scalar value from each intermediate representation x_t. This regressor is identical to the one we utilize to regress an output 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. $R^d \rightarrow R^{\frac{d}{4}}$using $a_t^i = ReLU(W_i a_t^{i-1}) + b_i$. Herein, $W_i \epsilon R^{s_i \times d}$ is the transformation matrix between layers and b_i is the bias term for the i-th layer. First a_t is the intermediate representation x_t. 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 $a_t = \frac{a_t - min(a_1, a_2, \ldots, a_t)}{max(a_1, a_2, \ldots, a_t) - min(a_1, a_2, \ldots, a_t)}$. We use such normalization functions to better regularize our network weights, and also provide contextual information across representations for a given participant.

3.1.4 Sequence Processing

Regressing a single scalar value requires some form of summarization of the many sentences a participant utters during their interviews. To this aim, we investigate recurrent and non-recurrent architectures. Transformer architecture was also considered, but our preliminary experiments showed that multi-head attention with positional encoding saturated our embeddings, and we could not reach decent performance. Therefore, it was not included in our in-depth analysis. In this section, we will provide the details of the various summarization methods we explored.

Temporal Modeling of Sequences To temporally model sentence embeddings, we use bidirectional GRUs. GRU architecture is selected due to its documented performance on temporal problems over RNNs and LSTMs, which we also replicated in our cross-validated preliminary experiments, and due to their ability to process variable length sequences. GRUs are also, on average, faster than RNN and LSTMs and use less parameters. Bidirectionality incorporates information from both directions (forward and backward) of the sequence. We define the forward direction of our GRU using the following equations:

$$\begin{aligned} z_t &= sigmoid(W_z x_t + U_z \vec{h}_{t-1}) \\ r_t &= sigmoid(W_r x_t + U_r \vec{h}_{t-1}) \\ c_t &= tanh(W_c x_t + U_c(r_t \odot \vec{h}_{t-1})) \\ \vec{h}_t &= (1 - z_t)\vec{h}_{t-1} + z_t c_t \end{aligned} \tag{1}$$

Where r_t and z_t are the reset and update gates, respectively. The activation of the hidden state $\vec{h}_t$ at time t is the linear interpolation between previous activation $\vec{h}_{t-1}$ and the candidate activation c_t. Weight matrices W and 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 element wise multiplication. We concatenate the hidden states of the forward and the backward passes for each t to obtain h_t. The resulting s_i many timesteps are reduced 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 $h^{(i)} \epsilon R^H$ for $P^{(i)}$ is obtained by:

$$\begin{aligned} \text{LAST\_POOL} &\rightarrow h^{(i)} = h_{s_i - 1} \\ \text{MEAN\_POOL} &\rightarrow h^{(i)} = \frac{1}{s_i} \sum_{t=1}^{s_i} h_t \\ \text{MAX\_POOL} &\rightarrow h^{(i)} = max[h_1; h_2; \ldots; h_t] \end{aligned} \tag{2}$$

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 3.1.4, mean-pooling takes the average of sentence representations, while max-pooling filters the maximum activations along the t dimension. Formally, the output $h^{(i)} \epsilon R^D$ for $P^{(i)}$ is obtained by:

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

$$\text{MAX\_POOL} \rightarrow h^{(i)} = max[x_1; x_2; \ldots; x_t] \quad (3)$$

3.2 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:

- NT-MEAN: **N**on-**T**emporal model using **MEAN**-pooling
- NT-MEAN-ATT: **N**on-**T**emporal model using **MEAN**-pooling with **Att**ention
- T1-MEAN: **T**emporal model using **one** GRU with **MEAN**-pooling
- T2-MEAN.MAX: **T**emporal model using **two** GRU's with **MEAN**-pooling on first level and **MAX**-pooling on second level

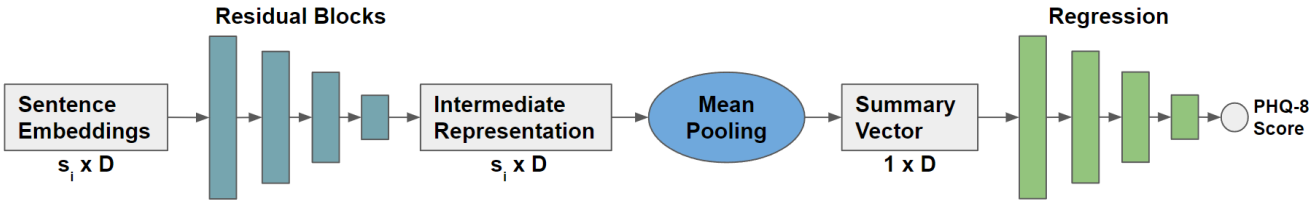


Figure 2: Overview of the NT-MEAN architecture. Shapes are given for a single participant and not batched input. For participant p_i , s_i is the number of sentences, D is the size of the sentence embedding (depends on embedding choice).

3.2.1 NT-MEAN and NT-MEAN-ATT

NT-MEAN is our simplest model. Sentence embeddings are passed through residual blocks. These output representations $x_t \epsilon R^{s_i \times D}$ 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 score. NT-MEAN-ATT is a non-temporal model similar to NT-MEAN. Their difference lies in the addition of a 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. Figure 2 shows the detailed NT-MEAN architecture.

3.2.2 T1-MEAN and T2-MEAN.MAX

T1-MEAN and T2-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. T1-MEAN uses a single level GRU followed by mean-pooling to create a summary representation of shape $1 \times H$. On the other hand, T2-MEAN.MAX utilizes a two-level 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.

4. Experimental Setup

4.1 Dataset

4.1.1 Overview

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 [28]. This is an actively used benchmark dataset for this task. 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. We use the predetermined splits during our experiments.

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

4.1.2 Challenges

Label imbalance 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 $\geq$ 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.

Transcription Noise 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 real 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.

4.2 Evaluation Criteria

Dataset used in this study was first introduced in The Audio/Visual Emotion Challenge and Workshop 2019 (AVEC 2019) [28]. Organizers of the challenge picked Lin's Concordance Correlation Coefficient (CCC) [29] 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. Since we are dealing with a sample of the total population, we use an approximation of CCC:

$$\widehat{CCC} = \frac{2S_{YX}}{S_X^2 \quad + S_Y^2 + (Y - X)^2} \tag{4}$$

Organizers choose this metric due to its invariance to scale, as well as its ability to include information on accuracy and precision [28]. We also use this metric in our training and evaluation. Organizers also propose Root Mean Square Error (RMSE) as a secondary metric. 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.

We follow the traditional training-validation-test scheme using the predetermined splits of the AVEC competition [28]. 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 and models are optimized with optuna library using a Tree-structured Parzen Estimator (TPE) for hyperparameter selection.

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 [30]

is generally detrimental for our training, and do not include it in our models. We use ReLU as our activation function of choice due to its popularity and performance over other activation functions. After our optimization and validation procedures, we opt for using four regression blocks, four linear layers for regression and 0.5 dropout probability in our proposed NT-MEAN model. All experiments are conducted on an Nvidia RTX 2080 Ti GPU.

5. Experimental Results

5.1 Temporal Modeling

Temporal models include the order of sentence representations within the sequence as additional information. We now analyze the effects of different pooling methods on these sequences, and the implications of processing overlapping sequence chunks in a two-level GRU setup. After this discussion, we pick the best-performing single-level and two-level models and evaluate them in the test set in Section 5.1.2.

Table 1: Comparison of pooling methods for temporal architectures. Having no pooling on the second level means that the model utilizes a single-level GRU.

First Level Pooling	Last	Last	Last	Last	Max	Max	Max	Max	Mean	Mean	Mean	Mean
Second Level Pooling	-	Last	Max	Mean	-	Last	Max	Mean	-	Last	Max	Mean
Validation CCC	0.589	0.632	0.649	0.634	0.646	0.655	0.649	0.637	0.650	0.658	0.659	0.624

5.1 Assessment of the Number of Recurrent Layers and Pooling Methods

Results in Table 1 show that some configurations of the two-level 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. Also, two of the top three results in this analysis use last-pooling in the second level, providing 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.

Table 2: Results for different recurrent structures. Results are given for three pooling methods in the validation set.

	Pooling Method		
Recurrent Structure	Last	Max	Mean
GRU	**0.589**	**0.646**	**0.650**
LSTM	0.557	0.625	0.643
RNN	0.363	0.614	0.603

In the single-level setup, we compare three different temporal models, each of which uses a different pooling method to obtain a single summary vector. With the last-pooling method, the performance of the sequence reduction depends on the assumption that h_t holds the information for the entire sequence. This assumption may not hold well based on the length of the sequence and the decisions of gates within the architecture. 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 discussed in Section 4.1.2, 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 is 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 average pooling over maximum pooling in the proposed case is because for a participant to be high on the PHQ-8 scale, cues need to be salient throughout the entire interview, not just in parts of the interview. To make sure we are covering a wider range of temporal models when we talk about them, we also include experiments conducted on RNN and LSTM architectures (Table 2). GRU outperforms the other architectures in every pooling configuration. For brevity, we do not go into details for them.

Table 3: Results for best performing temporal (T1-MEAN and T2-MEAN.MAX) and non-temporal (NT-MEAN and NT-MEAN-ATT) models. Results are given for three metrics in both validation and test sets.

Model	Validation			Test		
	CCC	MAE	RMSE	CCC	MAE	RMSE
T1-MEAN	0.650	3.393	4.66	0.598	5.232	6.656
T2-MEAN.MAX	0.659	3.321	4.33	0.572	4.464	5.616
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

5.1.2 Best-Performing Temporal Models

When we compare our single-level and two-level 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.

In this section, we examined the temporal dynamics regarding depression cues. While learning such relationships result in good models, it is unclear how much of our performance can be ascribed to temporal dynamics. 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 [31]. 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.

5.2 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. Similar to the temporal models, we start by experimenting with different pooling methods. The next subsection uses different attention methods and comments on their differences. We again finalize our discussions by evaluating the best-performing models from each subsection.

5.2.1 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. In Section 5.1, 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, and achieve 0.673 for mean-pooling and 0.629 for max-pooling in terms of validation CCC. Observing the performance of mean-pooling, it appears that the exclusion of temporal information leads to a performance increase. As with the temporal pooling experiments in Section 5.1.1, 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 5.1.1. It seems that individual high activations are less impactful while forming a summary vector compared to computing the overall activations.

5.2.2 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 4: 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 in Table 4 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.

Sigmoid function introduces additional non-linearity to our network, and it could give saturated weights for some embeddings due to its shape. Although it has such qualities, it 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. Even though AWM with min-max normalization is the best among weighted models, it is still behind simple mean-pooling (Section 5.2.1). 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 5.1. 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 5: 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

Table 5 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.

5.2.3 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 to 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 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. Four residual blocks with 0.4 dropout probability, followed by four linear layers with 0.5 dropout probability to obtain the best-performing NT-MEAN model.

5.3 Experiment on Word Count per Sentence

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.

During our analysis of feature importance in Section 5.2.2, we observe that the length of individual sentences differs for different participants. As we recall from Table 5, 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 re-evaluate 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.

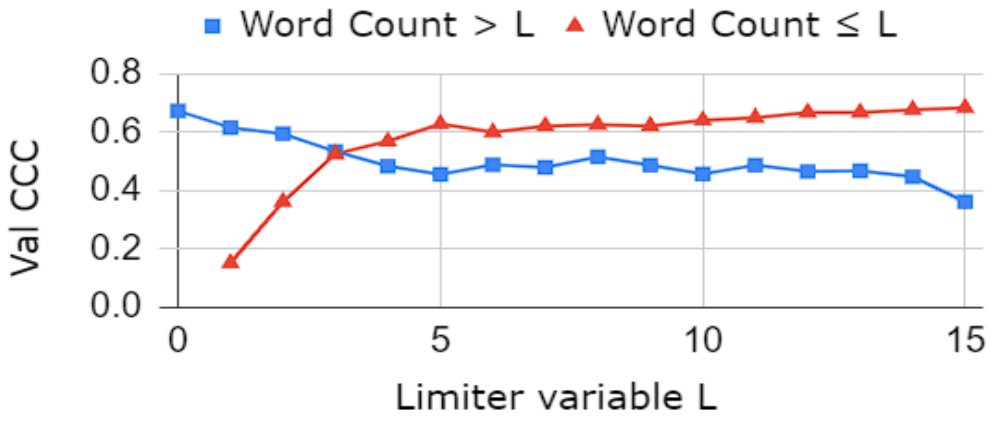


Figure 3: Additive (Red-Triangle) and Subtractive (Blue-Square) experiments on word count, using NT-MEAN model. For Limiter variable L, Additive experiment evaluates NT-MEAN on all sentences with word count ≤ L . Subtractive experiments do the same for word count > L.

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 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.

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.

Coincidentally, sentences with at most 6 words (L = 6) make up 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 sentences.

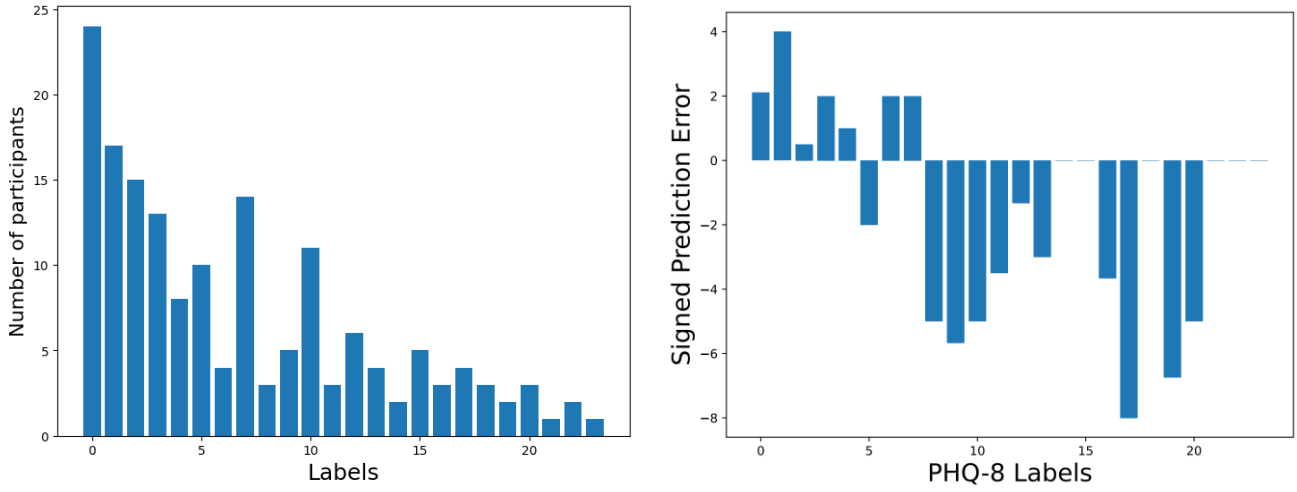
Our discussions up to this point originated from a trained model. We examine the applicability of our argument to learning by training two models. Each of these models perform their training and validation using half of the data, one uses sentences such that $|S| > 6$ while the other uses $|S| \leq 6$. We observe that these models achieve a CCC score of 0.568 and 0.575 respectively. This means that both models perform significantly worse than NT-MEAN (0.673), 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. Per our previous finding in this section, one would expect the $|S| \leq 6$ model to perform better. We argue that not including longer 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 [32, 33]. Some architectures purposefully focus their training on optimizing hard examples [34], most popularly in the case of most novel deep metric learning networks [35]. In our case, longer 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 5. 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.

We stress that these findings are dependent on the dynamics of this dataset, and future work should be conducted to better understand this behaviour. We are also making our arguments based solely on analytical findings, and the argument holds only for a specific part of the pipeline, the inference. In that sense, the reliance argument on shorter sentences is purely for computational purposes, and not necessarily an argument on the linguistic properties of the disease. With that being said, longer pauses between words and sentences that we observe with depressed patients [7, 8] could mean that the automatic transcription tool creates shorter, less convoluted sentences, and these sentences are more impactful during inference.

5.4 Experiment on the Effects of Imbalanced Data

The scarcity of data for highly depressed individuals in the dataset is clearly a challenge. This problem also applies to scores in the 10 $\leq$ PHQ-8 < 20 segment due to label imbalance, and makes overfitting towards the scores in the PHQ-8 < 10 segment a possibility. However, since our problem is modeled as a regression problem, we believe this issue is applicable for every PHQ-8 score. To demonstrate that the dataset is not suffering from high degrees of overfitting, we present 4. This figure shows (a) the PHQ-8 distributions in the training set and (b) the labeled prediction errors in the test set. With this figure, we hope to observe the generalization performance in the test set for PHQ-8 scores that have the potential to be memorized during training. We would like to draw your attention to a few points in these figures. First, although the PHQ-8 < 10 segment constitutes a very large portion of the data, it also varies within itself. For instance, there are many underrepresented PHQ-8 scores in this segment. Specifically, scores like PHQ-8=6 and PHQ-8=8 have much less data compared to PHQ-8=1. The error margins for these individuals vary. For example, PHQ-8=1, which has 24 representatives, and PHQ-8=6, which has 4 representatives, have the same average error margin. PHQ-8=12, which has 6 representatives, has achieved a better error margin than both of these groups. Another example is that PHQ-8=22, which has two representatives, was predicted with a 0 error margin in the test set (since the error margin is 0, it did not form a line

in the graph; there is one data point in the test set that was predicted with a 0 error for the PHQ-8=22 score). We do not believe we've encountered a situation that would suggest a likelihood of memorization between the number of representatives and error margins.



(a) (b)

Figure 4: Figures depicted to show the effect of (a) the number of PHQ-8 representatives during training on the (b) generalization error. Our argument is that the imbalance in the data distribution does not lead to significant memorization

Table 6: Details regarding the modalities and performance of other, most recent studies in the literature that use AVEC2019 dataset, to the best of our knowledge. Modalities are abbreviated as A = Audio, V = Vision, T = Text.

Model	Modalities	Val CCC	Val RMSE	Val MAE	Test CCC	Test RMSE	Test MAE
NT-MEAN (Proposed)	T	0.673	4.22	3.21	**0.729**	**4.35**	**3.30**
Fang et al. (2023) [36]	AVT	-	-	-	-	5.17	-
Wang et al. (2022) [37]	AVT	-	4.03	3.05	-	-	-
Han et al. (2022) [38]	A	-	5.56	4.65	-	6.29	5.38
Sun et al. (2022) [19]	AVT	-	-	-	0.583	-	4.37
Saggu et al. (2022) [16]	AVT	0.662	4.32	-	0.457	5.36	-
Sun et al. (2021) [39]	AV	0.733	3.78	-	-	-	-
Yin et al. (2019) [40]	AVT	0.402	4.94	-	0.442	5.50	-
Makiuchi et al. (2019) [18]	AT	0.696	3.86	-	0.403	6.11	-
Kaya et al. (2019) [9]	AT	0.481	-	-	0.344	-	-
Ray et al. (2019) [15]	AVT	-	4.37	-	0.670	4.73	4.02
Ringeval et al. (2019) [28]	AV	0.336	-	-	0.111	-	-

5.5 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 6 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. [15] on all metrics. The relative improvements are 8.8% for CCC, 8.7% for RMSE, and 21.8% for MAE.

6. Conclusion

In this paper, we have proposed temporal and non-temporal architectures to predict PHQ-8 depression scores. Compared to the 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.

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 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 shorter sentences usually focus on a very specific semantic information or emotion and are therefore better captured by the sentence embeddings. We should stress that this is an analytical finding regarding only our inference step and could depend highly on the dynamics of the dataset, and we are not making linguistic arguments about the disease. More work is needed for a more solid understanding of this occurrence.

Motivated by the properties of text modality, we hope that the discussions we have started and improvements we have proposed in this paper 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.

References

[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] 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, no. 11, 2020.

[4] World Health Organization. “Depression and other common mental disorders: global health estimates” *Technical Report*, 2017. License: CC BY-NC-SA 3.0 IGO.

[5] J. Zinken, K. Zinken, J. Clare 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.

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

[7] M. P. Caligiuri and J. Ellwanger, “Motor and cognitive aspects of motor retardation in depression,” *Journal of Affective Disorders*, vol. 57, no. 1, pp. 83–93, 2000.

[8] Psychomotor symptoms of depression. *American Journal of Psychiatry*, vol. 154, no. 1, pp. 4–17, 1997. PMID: 8988952.

[9] H. Kaya et al. 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, page 27–35, New York, NY, USA, 2019. Association for Computing Machinery.

[10] 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*, 295:904–913, 2021.

[11] 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.

[12] 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*, 1237:032035, 06 2019.

[13] 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, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

[14] D. Cer et al. “Universal sentence encoder for English”, In *Proceedings of the 2018 Conference on Empirical*

Methods in Natural Language Processing: System Demonstrations, pages 169–174, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[15] A. Ray, S. Kumar, R. Reddy, P. Mukherjee and Ritu 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, pp. 81–88, New York, NY, USA, 2019. Association for Computing Machinery.

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

[17] 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), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

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

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

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

[21] S. Yadav, J. Chauhan, J. Prakash 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*, pages 696–709, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[22] J. Singh, N. Singh, M. M. Fouda, L. Saba and J. S. Suri, "Attention-enabled ensemble deep learning models and their validation for depression detection: A domain adoption paradigm," *Diagnostics*, vol. 13, no. 12, 2023.

[23] L. Tong et al. "Cost-sensitive boosting pruning trees for depression detection on twitter", *IEEE Transactions on Affective Computing*, pages 1–1, 2022.

[24] J. Liu and M. Shi, "A hybrid feature selection and ensemble approach to identify depressed users in online social media", *Frontiers in Psychology*, 12, 01 2022.

[25] 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.

[26] 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.

[27] 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),* pages 770–778, 2016.

[28] F. Ringeval et al. "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,* page 3–12, New York, NY, USA, 2019. Association for Computing Machinery.

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

[30] 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, page 448–456. JMLR.org, 2015.

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

[32] 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.

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

[34] Q. Dong, S. Gong and X. Zhu, "Class rectification hard mining for imbalanced deep learning", In *2017 IEEE*

International Conference on Computer Vision (ICCV), pages 1869–1878, 2017.

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

[36] M. Fang, S. Peng, Y. Liang, C.-C. Hung and S. Liu, "A multimodal fusion model with multi-level attention mechanism for depression detection", *Biomedical Signal Processing and Control*, 82:104561, 2023.

[37] C. Wang, D. Liu, K. Tao, X. Cui, G. Wang, Y. Zhao, Z. Liu, "A multi-modal feature layer fusion model for assessment of depression based on attention mechanisms". In *2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–6, 2022.

[38] Z. Han, Y. Shang, Z. Shao, J. Liu, G. Guo, T. Liu, H. Ding, Q. Hu, "Spatial-temporal feature networkfor speech-based depression recognition", *IEEE Transactions on Cognitive and Developmental Systems*, pages 1–1, 2023.

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

Conflict of Interest Notice
The authors declare that there is no conflict of interest regarding the publication of this paper.

Support/Supporting Organizations
Not applicable.

Ethical Approval and Informed Consent
It is declared that during the preparation process of this study, scientific and ethical principles were followed, and all the studies benefited from are stated in the bibliography. Since the data used in this research is publicly available, an ethics committee approval is not required / not applicable.

Availability of data and material
Not applicable.

Plagiarism Statement
This article has been scanned by iThenticate ™.