UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

Transformers in Depression Detection from Semi-Structured Psychological Interviews

Author:
Iker HONORATO LÓPEZ

Supervisor:
Jordi VITRIÀ
Javi JIMÉNEZ
Alberto COCA

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

June 30, 2023

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Transformers in Depression Detection from Semi-Structured Psychological Interviews

by Iker HONORATO LÓPEZ

The expansive adoption of Transformer models across the Machine Learning landscape is undeniable, and health is not an exception. This study undertakes a rigorous exploration of the efficacy of these novel architectures in discerning depression indicators from semi-structured psychological interviews. A key focus of this study is the extrapolation of the pre-training knowledge inherent in these models, and the comparison with traditional state-of-the-art Machine Learning models. In doing so, the thesis proposes a comprehensive framework designed to facilitate objective comparison. The study extends its inquiry into the differential performance of text and speech modalities, in isolation and combination, within the context of depression detection. Moreover, this research delves into the importance of topical relevance in the detection process, culminating in an evaluative discussion of crucial themes integral to accurate depression detection. Ultimately, this thesis contributes to the deepening understanding of the complex interplay between Transformer models, modality use, and topic importance in the realm of depression detection.

Acknowledgements

To flatmates for keeping me nourished when the deadline was getting close and to the people in AcceXible, for helping me get this project through.

Contents

Chapter 1

Introduction

1.1 Motivation

Major Depressive Disorder (MDD) is a leading cause of disability and a common worldwide mental health issue with high socioeconomic impact and the principal cause leading to suicide. According to World Health Organization (WHO) in 2017 an estimate of more than 300 million people worldwide suffered from depression, which is roughly 4.4 percent of the global population (WHO, 2017). Some of its symptoms might include (American Psychiatric Association, 2013):

1. Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day.
2. Significant weight loss when not dieting or weight gain, or decrease or increase in appetite nearly every day.
3. A slowing down of thought and a reduction of physical movement (observable by others, not merely subjective feelings of restlessness or being slowed down).
4. Feelings of worthlessness or excessive or inappropriate guilt nearly every day.
5. Diminished ability to think or concentrate, or indecisiveness, nearly every day.
6. Recurrent thoughts of death, recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.

Early diagnosis is important in the management of any patient. However, a large number of individuals with Major Depressive Disorder – nearly one in four – remain undetected, with less than half receiving appropriate treatment (Epstein et al., 2010). This gap in healthcare delivery may be attributed, in part, due to the lack of screening power in the field.

Among the prevailing methodologies employed to evaluate the severity of depression, the Patient Health Questionnaire (PHQ-8) is frequently utilized (Kroenke et al., 2009). From it we can extract the following intervals depending on the score on the test:

- 0-4, no significant depressive symptoms
- 5-9, presents mild depressive symptoms
- 10-14, moderate depressive symptoms, and the threshold for MDD
- 15-19, moderately severe symptomatology
- 20-24, severe symptoms

Moreover, the intention behind its questions is transparent, as illustrated in **Figure 1.1**. Nevertheless, it is crucial to recognize that patients' responses might be skewed due to an inclination to conform to societal stereotypes linked with depression (Falicov, 2003).

PHQ-8

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 ____ = ___ + ___ + ___)

From the Primary Care Evaluation of Mental Disorders Patient Health Questionnaire (PRIME-MDPHQ). The PHQ was developed by Drs. Robert L. Spitzer, Janet B.W. Williams, Kurt Kroenke and colleagues. For research in formation, contact Dr. Spitzera trls8@columbia.edu. PRIME-MD® is a trademark of Pfizer Inc. Copyright© 1999 Pfizer Inc. All rights reserved. Reproduced with permission

FIGURE 1.1: PHQ-8 questionnaire

However, it is also commonly known that psychologists are able to perceive depression traits based on open conversation with the patients (Lord, 1921), which is the reason why a semi-structured clinical interview where the physician determines the symptoms of a patient remains the principal diagnose methodology (Levis et al., 2018), since this modality is less susceptible to the previously mentioned social bias. Additionally, there is evidence that spontaneous language is an important factor when it comes to MDD early diagnosis, on an analysis in the effect on some speech features, it can be found that there are distinctive traits between read speech and impromptu conversation (Alghowinem et al., 2013), making the latter more capable.

Nonetheless this process presents clear limitations, primary due to the fact that the diagnosis largely depends on the physician's skills and the patient's willingness to cooperate (Epstein et al., 2010). Moreover, as the process is influenced by these human factors, it can be time-consuming and subjective in nature, and, as a result, the diagnostic efficacy and applicability in the wider population might not be as optimal as it is needed.

For this reason, several studies have been conducted to try to make the process rapid and efficient, by introducing virtual agents (DeVault et al., 2014, Gratch et al., 2014a) to perform the clinical interviews and serve as tools to analyze the verbal and non-verbal behaviours, which can serve as extra data to help physicians with their decisions (Gratch et al., 2013).

Moreover, one of the most promising fields of study is vocal biomarkers (Cohen, Kim, and Najolia, 2013), which is considered a unique health signal containing cognitive, neurological and physiological information. In fact, voice is a data quarry, from where acoustic, phonetic and prosodic information can be extracted, and not only it encodes this information, but also sociodemographic pieces, such as sex, age or emotional state (Hebbar, Somandepalli, and Narayanan, 2018, Reynolds et al., 2003). This has given an opening to Machine Learning (ML) to form part of the new methodologies to explore the early diagnose of depression.

1.2 Indicators

Research on spontaneous speech has been ongoing for an extended period of time, as an example (Darby, Simmons, and Berger, 1984) conducted a study on 13 hospitalized subjects aged between 27 to 67 years, who had been diagnosed with bipolar disorder or unipolar disorder. They discovered that the speech characteristics of depression typically included reduced stress, monopitch, and monoloudness. While reduced stress was evident in all cases, monopitch and monoloudness were consistently found together. It is also stated that, at that moment, there was a lack of information to make such an assumption. Nonetheless, latter studies (Moore II et al., 2008) confirm the hypothesis and also add that depressives have a slower rate of speech and relatively monotone delivery when compared with normal speaking patterns.

Moreover, various indicators can be found inside the lexicon, for instance, pauses, dubitation and constant distraction, related to mind wandering (MW) (Chaieb, Hoppe, and Fell, 2022) seem to increase in depressive patients, nonetheless, there is still little information on this topic. However, there are other signs that have been more thoroughly studied, for instance, that the insistent use of first person singular pronouns is an strong indicator of depression (Brown and Weintraub, 1984 and Edwards and Holtzman, 2017), or that people that present MDD symptoms tend to use more emotion related adjectives, with a tendency towards negative connotations (Alghowinem et al., 2013). These indicators served as basis for the development of several computerized text analysis methods such as the Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Francis, and Booth, 1999), from which we can extract a total of 93 lexicon-based features, which can be classified as follows:

- **Standard Linguistic Dimensions**: This includes purely information on tense and type of verbes, pronouns prepositions or even negations.
- **Psychological processes**: A classification on different words, which can be stratified, for instance in social, affective or cognitive among others
- **Personal concerns**: Trying to classify words inside various axis of a person's life, some examples are, work, home, along with money and others

Depression, as a mood disorder induces changes in response to emotional stimuli, so it is essential to examine if there is relationship emotion and depression speech (Reed, Sayette, and Cohn, 2007), and some studies have indeed proven the benefits of including emotion assessment in audio-based automatic depression diagnosis systems (Stasak et al., 2016). Normally, all literature found when it comes to text, or sentiment features, comes from social-media depression analysis.

Moreover, not only there are clues on the diagnosis in the way a person expresses, but also on variations in the frequency of the voice (Ozdas et al., 2000 and Trevino, Quatieri, and Malyska, 2011). **Table** 1.1 illustrates some the most common features that can be obtained from voice, extracted from openSMILE (Eyben et al., 2013), an open-source library for multimedia feature extraction.

Nevertheless, due to the nature of acoustic features, they tend to encode sociodemographic information as explained previously. That is why there are studies (Wang et al., 2019), which aim to examine whether vocal abnormalities in people with depression only exist in special situations. This study compared the vocal differences between healthy people and patients with unipolar depression in 12 speech scenarios. The study controlled for irrelevant demographic variables and found three acoustic features (loudness, MFCC5, and MFCC7) that were consistently dissimilar between people with and without depression, which could indicate that these acoustic features may be potential indicators for identifying depression via voice analysis.

Still, it is important to remark the gender-bias in these, for example, at (Alghowinem et al., 2012), they perform an study where the participants perform the PHQ-8 test and respond to a "good news" and "bad news" question, in order to encode sentiment inside the experiment. Results show, that with purely acoustic features, it seems easier to recognise depression from females.

Feature	Description
Waveform	The form of the wave caused by speech.
Loudness	Energy or intensity of a sound, it is related to the physical quantity of sound pressure level (SPL) which is measured in decibels(dB)
MEL filters	Derived from the Mel frequency filter bank applied to an audio signal. The Mel scale is a perceptual frequency scale that approximates the non-linear relationship between frequency and pitch perception in human hearing. (Gowdy and Tufekci, 2000)
Cepstral	The result of taking the inverse Fourier transform of the power spectrum of a signal. One of the most common is MEL-frequency cepstral coefficients (MFCCs), which are commonly used in speech recognition and speaker identification tasks.
Pitch	Pitch refers to the perceived fundamental frequency of the speaker's voice, it is measured in Hertz (Hz), and reflects the rate of vocal fold vibration. (Shrivastav, Eddins, and Anand, 2012)
Voice Quality	Jitter or shimmer are acoustic characteristics of voice signals, and they are caused by irregular vocal fold vibration. They are perceived as roughness, breathiness, or hoarseness in a speaker's voice. (Reynolds et al., 2003)

TABLE 1.1: Most common speech features

1.3 State of the art

When it comes to automatic depression assessment, various approaches have been proposed and studies can be grouped based on the type of features and tools employed for the prediction. For instance, some studies rely solely on lexicon features and employ convolutional neural networks (CNNs) to detect signs of depression in social media posts (Trotzek, Koitka, and Friedrich, 2020). On the other hand, certain studies utilize transcriptions of medical interviews to extract inherent information (Mallol-Ragolta et al., 2019), who use Global Vectors (GloVe) to extract low-level word representations. They compare hierarchical local-global attention networks and hierarchical contextual attention networks, achieving an Unweighted Average Recall (UAR) of 0.60. UAR is an evaluation metric used for imbalanced datasets, calculated as the sum of class-wise accuracy (recall) divided by the number of classes.

Other projects (Sardari et al., 2022) introduce an end-to-end depression detection

model that utilizes a CNN Auto-encoder for automatic feature extraction from raw audio signals of clinical interviews. They report a 0.70 F-score without explicitly using audio features, however they use a data augmentation technique in which they use a sliding window to divide the audios of the dataset, then they perform an undersample on the negative class. This strategy gets close to previous approaches, (Ma et al., 2016), which propose a deep model combining CNN and LSTM for a more comprehensive audio representation, yielding only a 0.72 average F-measure. Additionally, other works employ an MFCC-based Recurrent Neural Network (RNN) to detect and evaluate depression severity levels (Rejaibi et al., 2022). By leveraging simple MFCC features and performing random oversampling on the positive class, the process is efficient and achieves a promising accuracy of 0.76, however the F1-score for depressive participants is 0.46. Still, acoustic features remain a promising field for further exploration in depression assessment.

Moreover, some of the most interesting result come when combining both modalities. For instance there is proposals (Hanai, Ghassemi, and Glass, 2018) of a model composed of two branches of bi-LSTMs, one for audio and the other for text, with their outputs merged into a final feedforward network. The branches were constructed with different topologies, optimized according to the characteristics and information content of each modality. This model achieved a 0.77 F1 score on validation, surpassing previous studies reviewed, which could be due to the fact that bi-LSTM can encode temporal information. Moreover, according to the authors, there is a possibility that temporal and discriminative information related to the speech patterns of individuals with depression is present in speech, and extends over a longer time frame than the syntactic and semantic information that can be extracted from textual data. This finding opens up a new avenue of research in which the relationship between the textual and corresponding audio segments can be explored to gain a better understanding of depression assessment using multimodal data.

Expanding on the research mentioned earlier, studies propose a novel methodology for incorporating Transformers (Vaswani et al., 2017) into depression assessment models (Toto, Tlachac, and Rundensteiner, 2021). In their model, they utilize BERT-embeddings and Audio-Embeddings, along with bi-LSTMs for each modality, and fuse the outputs using attention before feeding the output to a dense layer. This approach aims to leverage the strengths of both modalities while, due to the nature of LSTMs, mantain temporal order. However, this approach is susceptible to memory loss, which could affect model performance. To address this issue, they utilize a type of attention that helps the model maintain information while preserving strong relationships between audio and text features. Moreover, they divide the dataset in 10 topic based datasets, depending on the question that is being asked to the participant. The proposed model achieved an average F1-Score of 0.72 and a maximum of 0.92 on one of the datasets, showcasing the potential of integrating Transformers in multimodal models for automatic depression assessment. Nonetheless, for each of the thematic datasets, they trained 10 models with different initializations, and they chose to report the score of the top 3 scoring on the test set, making the results hard to validate.

Moreover, an other approach is a novel multi-modal topic-attentive model, (Guo et al., 2022) that also employs Transformers in the branches for each data type, using a more advanced version of BERT, RoBERTa, and an improved version of audio embedding, Wav2Vec 2.0. Rather than directly concatenating the embeddings, they use a fusion-module that performs several actions, including topic-attentive attention

from the text data. This attention mechanism helps weighting the importance of specific questions in depression assessment, which enhances the model's performance. After the embeddings are fused, they are fed into a fully connected layer. This approach achieves a 0.64 F1 score, which are not superior results compared to other bi-LSTM approaches, however it does evaluate unimodal and multimodal approaches, stating that the latter is more capable and follows classic train/validation/test evaluation.

Overall, these findings highlight the importance of leveraging deep learning techniques, such as Transformers and attention mechanisms, to incorporate both text and audio information in depression assessment models, and the importance of time-based approaches.

1.4 Scope

The current literature lacks a uniform method for reporting results in this field, so our study suggests a structured approach for exploring different methodologies for depression analysis. This approach includes machine learning (ML) models and traditional feature extraction techniques, based on what we have found in literature, which will act as a reference point and benchmark for further experiments in our project. Additionally, we will focus on transformer models, using pre-trained architectures to try to extract meaningful information from our data. We will then evaluate if this data holds important information for our problem and see if this new approach surpasses classic ML models.

To compare models accurately, we will adopt a standardized evaluation method based on hyperparameter tuning using a validation set, with the final evaluation done on a test set. All decisions, including model choices, will be based on results from the validation set to prevent data leakage into our final results.

Moreover, a crucial aspect of this research entails identifying the effective topics and questions for detecting depression. To achieve this we will divide our data in topic datasets, similar to what is done in AudiBERT, and then a comparative assessment will be conducted between audio and text techniques. We will introduce RoBERTa and Wav2Vec2 models, which will play integral roles in this examination, as we will analyze each domain separately.

Weighing up above all, the last step on this study will consist in the development of a multimodal model that combines both audio and text modalities. By leveraging the unique strengths of audio-based analysis using models such as Wav2Vec2 and text-based generalization using models such as RoBERTa, we aim to enhance the overall accuracy and comprehensiveness of the detection approach as well as analyze if this strategy yields better result than singular modalities.

All code produced by this study can be consulted in the Github Repository of the thesis.

1.5 Document structure

This research work is organized into four key sections. In **Chapter 1**, we have provided a comprehensive introduction to the problem under investigation, defining its scope and significance in the current research landscape.

In **Chapter 2**, we delve into the methodology that forms the backbone of our study. This includes a detailed discussion of the technologies employed in our experiments, as well as a thorough description of the dataset used. We explain our data preprocessing techniques, justifying the choices made and their relevance to our research goals.

Moving on to **Chapter 3**, we discuss our experimental setup in detail. This section contains a comprehensive account of each experiment conducted, the results garnered, and a thoughtful discussion of these outcomes. This chapter aims to provide a clear understanding of our findings and their implications in the broader context of the problem studied.

Finally, in **Chapter 4**, we synthesize the research journey into major conclusions drawn from our findings. We provide thoughtful insights into the implications of our results and their potential applications in future research. This chapter also outlines future directions, highlighting the next steps we intend to take based on the current study's outcomes.

Chapter 2

Methodology

2.1 Technology

The field of natural language processing (NLP) has seen significant advancements in recent years with the development of word embeddings and transformer models. These techniques have revolutionized the way we process and understand human language, and have been widely adopted for various applications such as language translation, text classification, and text generation.

In addition to NLP, there has been a growing interest in audio processing, where similar techniques can be applied to represent and process audio signals. These techniques have been used for various applications such as speech recognition, speaker identification, and emotion recognition.

In the chapter, we will present the necessary explanations on the architectures that we will use for our experiment, aiming to provide an overview of the methodology behind the proposed models.

2.1.1 Word embeddings

Word embeddings (WE) are numerical representations of words that have been widely used in NLP applications. They are designed to capture the meaning and context of words in a high-dimensional space and have gained significant attention in recent years due to their ability to capture the semantic and syntactic properties of words.

There are two main techniques used for training word embeddings: count-based methods and predictive methods. Count-based methods analyze the co-occurrence patterns of words in a large corpus of text and use matrix factorization techniques to extract a low-dimensional representation of the words. On the other hand, predictive methods use neural network architectures to predict the context of a given word based on its surrounding words, such as Word2Vec (Mikolov et al., 2013)

One of the most common WE is GloVe (Pennington, Socher, and Manning, 2014) which is based on the idea of representing words as vectors that capture the relationships between words using co-occurence patterns of a corpus, taking the best of the previously explained methodologies. The model can be described as a log-bilinear model that incorporates a weighted least-squares objective. The fundamental concept driving the model is that the ratios of the probabilities of words co-occurring with each other have the potential to encode some type of meaning, creating this way the WE, **Figure 2.1**.

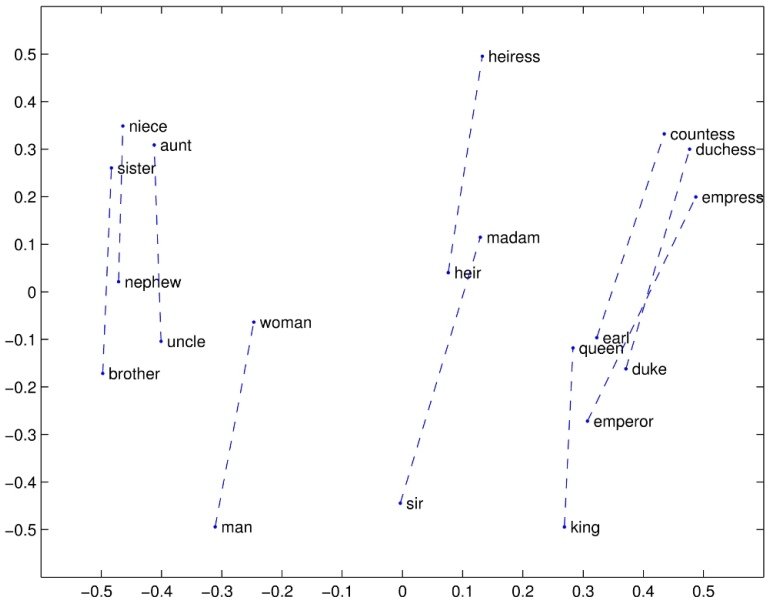


FIGURE 2.1: GloVe embeddings visualization

2.1.2 Transformers

As it has been seen in the previous chapter, recent advances in the field of automatic depression assessment come from the use of complex deep learning models such as Recurrent neural networks (RNN) and Transformers (Vaswani et al., 2017). However, the latter has been widely adopted in different fields such as NLP or CV due to its unique feature of self-attention. Unlike RNNs, this type of attention allows Transformers to process all input at once, by weighing each input of the sequence according to its importance, allowing this architecture (**Figure 2.2**), to overcome classic problems of RNNs such as GPU parallelization or memory loss.

As it can be seen in **Figure 2.2** , in a transformer architecture, the input sequence is first converted into a sequence of tokens using an embedding layer. However, as the model takes the whole data sequence as input, the sense of order is lost. This is why Transformers add a positional encoding vector to each of the tokens resulting from the first embedding layer (**Figure 2.3**), this positional embedding is calculated using *sin* or *cos* functions:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}})$$

where *pos* is the position and *i* is the dimension. With this each of the dimensions of the positional encoding corresponds to a sinusoid, and by controlling the wavelengths of the function, it is possible to encode dimensionality according to the necessity.

These are then passed to the encoder component of the transformer, which generates a new vector per token with the same shape as the input sequence. The multihead attention mechanism in the transformer allows the model to attend to different parts

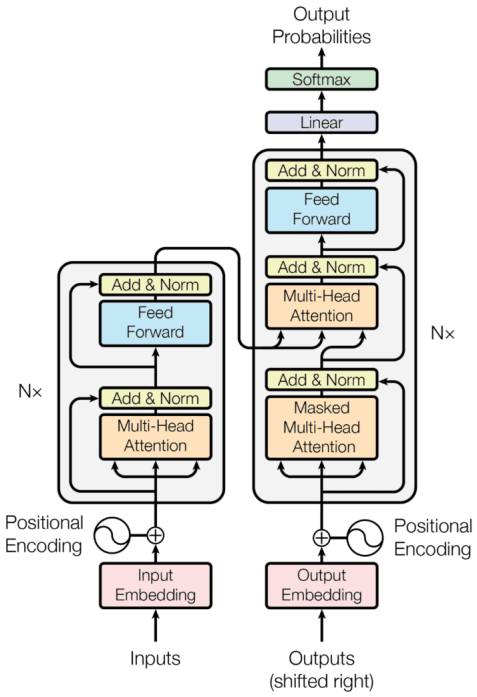


FIGURE 2.2: Transformer Architecture

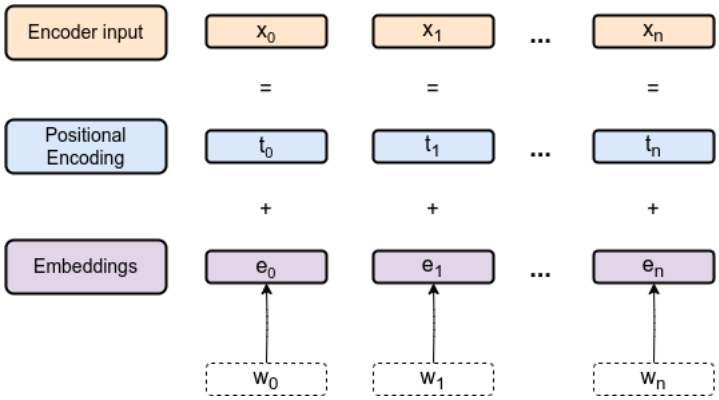


FIGURE 2.3: Positional encoding

of the input sequence, enabling it to capture long-range dependencies and context. This attention mechanism can be leveraged to 4 steps:

1. **Linear Projections**: Each input vector is transformed into three different vectors: a query vector (Q), a key vector (K), and a value vector (V). These transformations are done using separate learned linear projections for each attention head. Formally, for each input vector x_i, we create query, key, and value vectors by multiplying with learned weight matrices W^Q, W^K, and W^V respectively. This is done for each attention head h.

$$Q_i^h = x_i W_h^Q$$

$$K_i^h = x_i W_h^K$$
$$V_i^h = x_i W_h^V$$

2. **Scaled Dot-Product Attention**: For each attention head, the attention scores are computed by taking the dot product of the query vector with the key vector of every other word, followed by a scaling operation (division by the square root of the dimension of the key vectors), and then applying a softmax function. This results in a probability distribution that signifies the importance of each data point in the sequence with respect to the current one.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

3. **Compute Output**: The attention scores are then used to weight the value vectors. The weighted sum of the value vectors forms the output of the attention head.

$$\text{H}_h = \text{Attention}(Q^h, K^h, V^h)$$

4. **Concatenation**: The outputs of all attention heads are concatenated and linearly transformed to result in the final output vectors.

$$\text{MultiHead}(Q, K, V) = [\text{H}_1, \ldots, \text{H}_H]W^O$$

   Where W^O is a learned weight matrix that transforms the concatenated vector back to the original embedding dimension and mixes the information from the different attention heads.

Moreover the implementation in a Transformer might add residual connections between steps of the multihead attention polishing the process even more. After this, by adding and normalizing the result of this layer, the model enriches the original token representation. Each token is then refined and passed through a feed forward network, which will capture higher-level features of the input sequence, which after another normalization step will return the final new token representation.

The decoder component of the transformer takes the encoder's output to generate the overall output of the model. However, as the output sequence of a transformer is created token by token, hence only the tokens up to that point are available, the first part of this block, uses a different attention. On the next step, there is another attention mechanism that uses the learned encoder embedding to evaluate which tokens of the input are more relevant to the current output token, generating a vector. These vectors are then fed to a feed-forward layer followed by linear transformation and softmax functions to convert the decoder output to predicted next-token probabilities. The decoder output will serve as input of the next decoder until a "end of the sentence", [END] token is found, marking the end of the process.

2.1.3 BERT

Transformers give an opening to complex and context based embeddings, which results in a new WE strategy: transformer based embeddings. One of the most common is BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), and it is considered on of the state-of-the-art language model for NLP.

BERT is a stack of pre-trained Transformer encoders, as language knowledge is included in this layer, this model does not need a decoder stack. By adding dense layers on top of BERT it is possible to fine-tune the architecture to perform different classification tasks taking advantage of its previously learned language understanding, this process is commonly known as transfer learning.

Moreover, BERT uses an special type of embedding layer, this makes the model robust to different input structures, such as long texts with multiple sentences. This layer is conformed of 3 types of embeddings (**Figure 2.4**);

- **Token Embeddings**: The first one, is the common WE, used in NLP and the Transformer original architecture, in this case, BERT uses WordPiece embeddings (Wu et al., 2016), which are a type of subword embeddings that break words into smaller subword units called "pieces". The most frequent pairs of character sequences are iteratively merged to create a vocabulary of subword units. Words are then split into sequences of these subword units, which are mapped to their corresponding WordPiece embeddings. They can handle out-of-vocabulary words better than whole-word embeddings and capture more fine-grained information about the language.
- **Segment Embeddings**: BERT marks the beggining of the input sequence, with a [CLS] token along with the end of each of the sentences with a [SEP] token. Then, each of the input tokens are categorized according to its sentence using this embeddings.
- **Positional Encoding**: They follow the same functionality as in the classic Transformer structure.

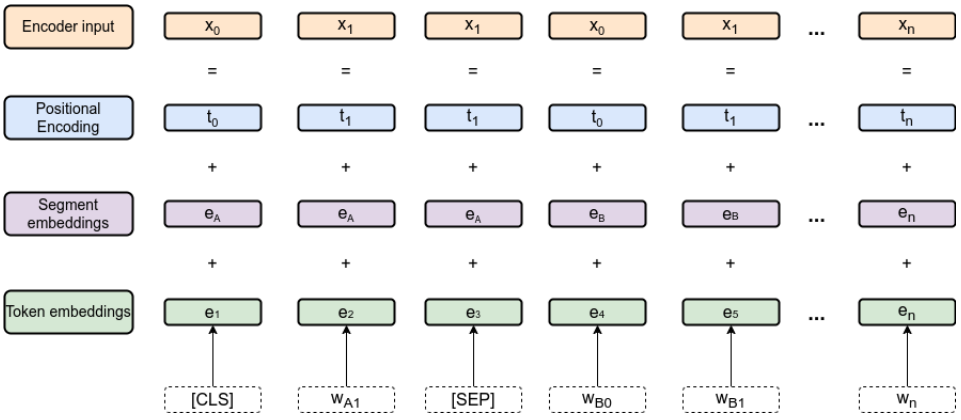


FIGURE 2.4: BERT embedding layer

During the pre-training phase, BERT is trained on a large corpus of text using two objectives:

- **Masked Language Modeling (MLM)**, which randomly masks some of the tokens in the input sequence and trains the model to predict the original token based on the context of the other tokens. This is what gives the Bidirectional name to BERT, as it takes into account previous and future context to make this prediction
- **Next Sentence Prediction (NSP)**, which trains the model to predict whether two input sequences are contiguous or not.

As said before, BERT consists on a stack of encoder layers, as seen in **Figure 2.5**, the output of each one is passed to the next encoder layer as input, and the process is repeated for a fixed number of layers (12 or 24). The final output of the top layer of the stack is used as a representation of the input sequence, and is the resulting BERT-WE.

The use of multiple layers allows this architecture to capture increasingly complex patterns in the input sequence as it processes it. The lower layers of the stack capture local patterns in the input sequence, while the higher layers capture more global patterns that depend on interactions between different parts of the sequence.

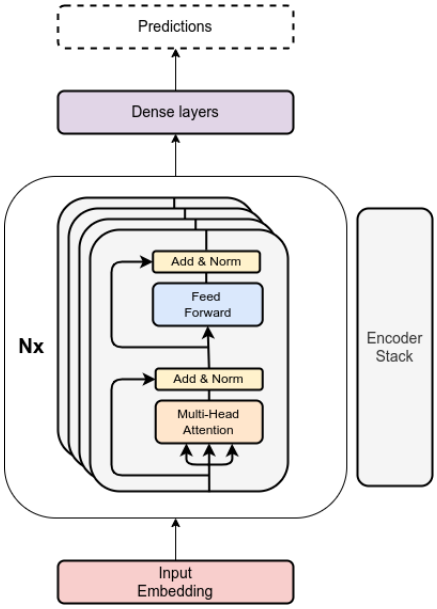


FIGURE 2.5: BERT architecture

2.1.4 RoBERTa

RoBERTa (Robustly Optimizer BERT pretraining Approach) is a language model that seeks to optimize BERT's architecture even further (Liu et al., 2019).

The authors found that BERT was significantly undertrained and could benefit from longer training times with larger batch sizes. They also discovered that the next sentence prediction (NSP) objective, which was originally thought to help the model understand the relationship between sentences, did not contribute significantly to the model's performance.

RoBERTa uses dynamic masking rather than static masking. In BERT, the masking pattern is determined before training and remains the same throughout, but in

RoBERTa, the masking pattern changes for each epoch, which means the model sees different masked versions of the same sentence, leading to better performance while sharing the same architecture.

2.1.5 Wav2Vec 2.0

So far the focus has been on text models and word embeddings. However, these models and training processes can be extrapolated to other grounds. In this case, we focus on how to process speech, in terms of audio, in order to get an embedding from it.

Wav2Vec 2.0, or Wav2Vec2 (Baevski et al., 2020), is an automatic speech recognition system (ASR) that is trained on large amounts of unlabelled audio data to learn speech representations which are possible to fine-tune it to perform other classification tasks. Contrary to its predecessor, Wav2Vec, this new architecture uses a transformer encoder in its core. The idea behind the architecture is similar to BERT, however, due to the nature of audio data it is necessary to make complex transformations to the input.

In order to do so, Wav2Vec introduces a preprocessing layer, the **Feature encoder**, which consists on a CNN layer, called a temporal convolution, followed by layer normalization and a GELU activation function, which can be thought as a smoothed ReLu. The idea behind this is to segment the audio in a vector of dimension Z, where each Z_t represents a 20ms segment so that the raw audio can be fed into the encoder as a vector.

The output of the feature encoder layer is then fed to the **Context network**, which is in essence, a transformer encoder layer stack, which depending on the model size, can contain blocks of 12 or 24 layers. However, it introduces a subtle change due to audio having different time-structures than text. The positional encoding is substituted by grouped convolutions, which help to learn time relations between subgroups of Z_t segments and they receive the name of relative positional embeddings. Once the input has been processed by the stack of layers, it returns a context vector C.

Moreover not only Z is fed to the context network, but also in parallel to the **Quantization layer**. Quantization refers to the process of discretizing values from a continuous space into a finite set of values in a discrete space, and it will help with the model pretraining. Consider a latent speech representation vector, denoted as Z_t, which captures information about two phonemes. Since the number of phonemes in any given language is limited and the number of possible pairs of phonemes is also finite, it is possible to represent them accurately using the same latent speech representation. Moreover, these pairs are finite in number, allowing us to construct a codebook that contains all possible pairs of phonemes. Consequently, the task of quantization boils down to selecting the appropriate code word from this codebook, however the number of distinct sounds in a language can be vast. To facilitate training and practical usage, the authors of Wav2Vec2 devised G codebooks, each comprising V code words. The process of generating a quantized representation involves selecting the best word from each codebook and concatenating these chosen vectors. Subsequently, a linear transformation is applied to the concatenated vectors to obtain the final quantized representation.

In order to choose the best codeword from every codebook, the architecture uses **Gumbel softmax**:

$$p_{g,v} = \frac{exp(l_{g,v} + n_v)/\tau}{\sum_{k=1}^{V} exp(l_{g,k} + n_k)/\tau}$$

where Z_t is mapped like $l \in \mathbb{R}^{GxV}$. Then, $n = -log(-log(u))$ and u are uniform samples from $\mathbb{U}$. and finally τ is a nonegative temperature. Gumbel softmax introduces two variations compared to a normal softmax: randomization and temperature. The inclusion of randomization ensures that the model selects different codewords, preventing it from relying solely on a subset of codebooks. This is particularly important during the initial stages of training. By gradually decreasing the temperature parameter over time, the model can regulate the level of randomization and its impact on the output.

Once the pre-training process is complete, the model begins making predictions. This pre-training strategy follows a similar approach to BERT. Initially, a portion of the latent speech representation, denoted as Z, is masked before being input to the transformer layer. However, in the quantization module, Z remains unchanged. The model is then tasked with solving a contrasting objective. It must identify the correct quantized latent speech representation for a masked time step, selecting it from a set of distractors that are uniformly sampled from other masked time steps within the same utterance. To assess the similarity between the context representations and the quantized latent speech representations, cosine similarity is employed, computed as $sim(a,b) = \frac{a^T b}{|a||b|}$. The model is designed to encourage high similarity with the true positive representation and penalize similarity to the distractors, promoting accurate and discriminative representations. This results in the latent representation of the audio data. A representation of the Wav2vec2 architecture can be seen at **Figure 2.6**, were each of the explained steps are depicted.

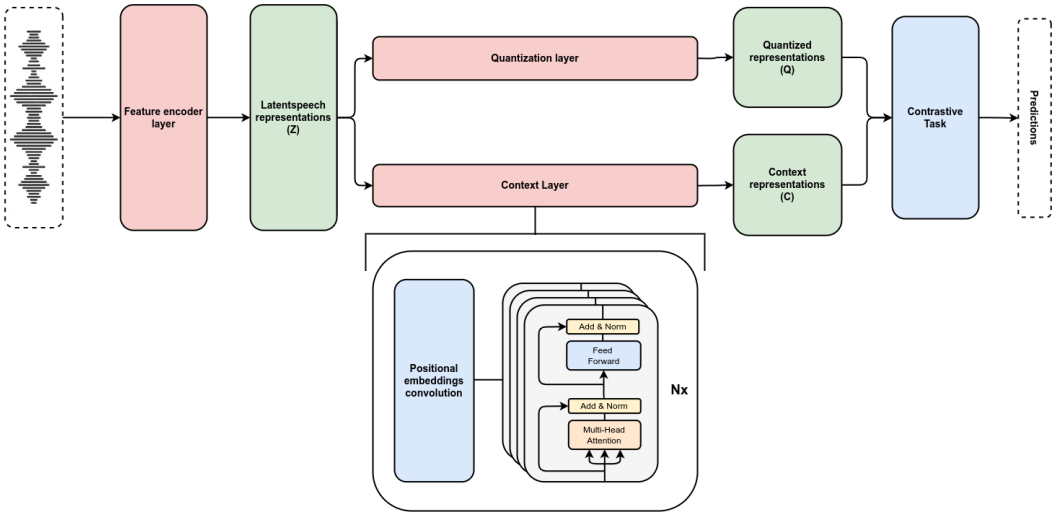


FIGURE 2.6: Wav2Vec2 architecture

2.1.6 Pretrained Architectures

Transfer learning has emerged as a powerful technique in machine learning, facilitating the utilization of pretrained transformers for a wide range of classification tasks, which have shown promise in leveraging their prelearned representations of text and audio.

The transfer learning process begins with the initial pretraining of transformer models on extensive datasets. This phase allows the models to acquire a profound understanding of the underlying structure and semantic relationships in the data. BERT, for instance, undergoes pretraining on text extracted from wikipedia as well as the book corpus (Zhu et al., 2015), while Wav2Vec2 is pretrained on LibriSpeech a corpus of approximately 1000 hours of 16kHz read English speech. As a result, these models capture general language and audio processing knowledge that could be useful for our problem.

Moreover, these pretrained transformer models can be further refined through fine-tuning on task-specific datasets. Fine-tuning involves updating the parameters of the pretrained models using labeled data specific to the target task, in our case depression detection. Instead of training the models from scratch, they retain the learned representations acquired during pretraining. This enables efficient learning on the specific task while mitigating the need for substantial amounts of labeled data. However, if the data is too scarce, it is possible to just freeze the transformer and add dense layers on top of it, which will be trained to use the domain knowledge of the transformer in their classification task.

The application of transfer learning with pretrained transformers could offer several advantages in the context of depression detection. Firstly, the pretrained models have already assimilated a comprehensive understanding of language or audio processing, which is useful for discerning signs of depression from textual or audio data. Leveraging these pretrained representations could allow the models to effectively capture the subtle semantic nuances and emotional cues associated with depression.

2.2 Data Description

The selected dataset for this study is the DAIC-WOZ (Gratch et al., 2014b), which is a component of the broader Distress Analysis Interview Corpus. This corpus encompasses a collection of clinical interviews specifically designed to aid in the diagnosis of various psychological distress conditions, including anxiety, depression, and post-traumatic stress disorder. Within this corpus, the DAIC-WOZ segment focuses on the Wizard-of-Oz interviews, where participants engage with an animated virtual interviewer named Ellie which is controlled by a human interviewer located in a separate room.

The DAIC-WOZ dataset comprises a total of 189 clinical interviews. Each interview is associated with gender information and PHQ-8 scores, which are computed from the answers of the questionnaire, which as explained before, are ranged between 0 and 3. The total score obtained from these eight questions represents the PHQ-8 score, with a higher of 10 score being catalogued as depression.

The duration of the clinical interviews within the DAIC-WOZ dataset ranges from 7 to 33 minutes and is recorded in a sample rate of 16KHz. The specific set of core

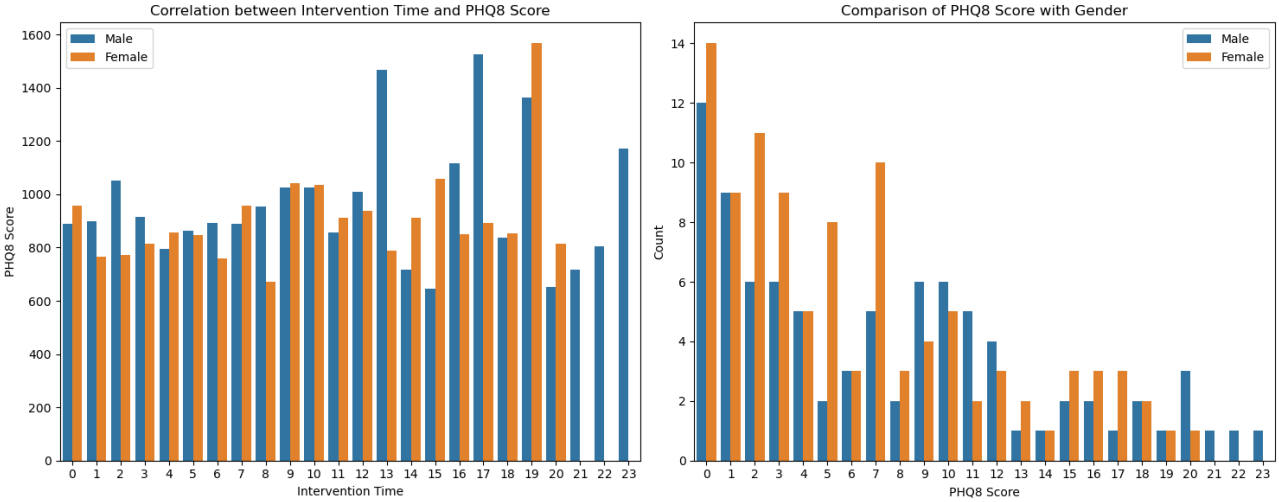


FIGURE 2.7: Distributions of Gender in the PHQ scale and its intervention time

questions and follow-up questions asked during each interview may vary, as does the question format across different participants. Moreover, each of the interviews is transcribed, marking the interventions of Ellie and the patient as well as interventions timestamps time. **Figure 2.7** depicts the distributions segmented by gender and intervention duration, presenting a pattern of longer intervention times coinciding with higher PHQ-Scores. Despite this trend, it's crucial to highlight the presence of a significant imbalance in the dataset with regards to the PHQ Score, the dataset contains a larger proportion of healthy individuals compared to those identified as depressive. Additionally, while there exists this imbalance in the context of mental health conditions, the dataset maintains a balance concerning the gender variable

To better analyze the dataset and emphasize topics crucial in diagnosing depression, we've broken down the data into specific themes. These themes originate from main questions and evolve with follow-up inquiries. To pinpoint these themes, we used a semi-supervised approach to examine every participant's transcriptions.

When looking for topics, we can spot certain patterns in Ellie's interventions that indicate a question is being asked. It's important to know that the original transcriptions don't use question marks, making the task of finding questions more challenging. So, we searched for question words like "who", "when", and "why", sentences that start with verbs like "Have you" or "Do you" and phrases like "tell me about". This search led us to about 110 distinct questions asked by Ellie. Still, there are instances where the question isn't positioned at the beginning of Ellie's intervention. Despite this, after a meticulous review of the list of questions and a thorough comparison with those found in the existing literature, it appears that all questions have had successfully identified.

The following phase of preprocessing consisted on filtering for unique questions, as for each topic, there could be two or three variants to introduce it. To accomplish this, we utilized a WE, known as the Universal Sentence Encoder (USE) (Cer et al., 2018), developed by Google. It has demonstrated considerable efficacy when comparing vectors of sentences with analogous meanings. We computed the embedding vectors

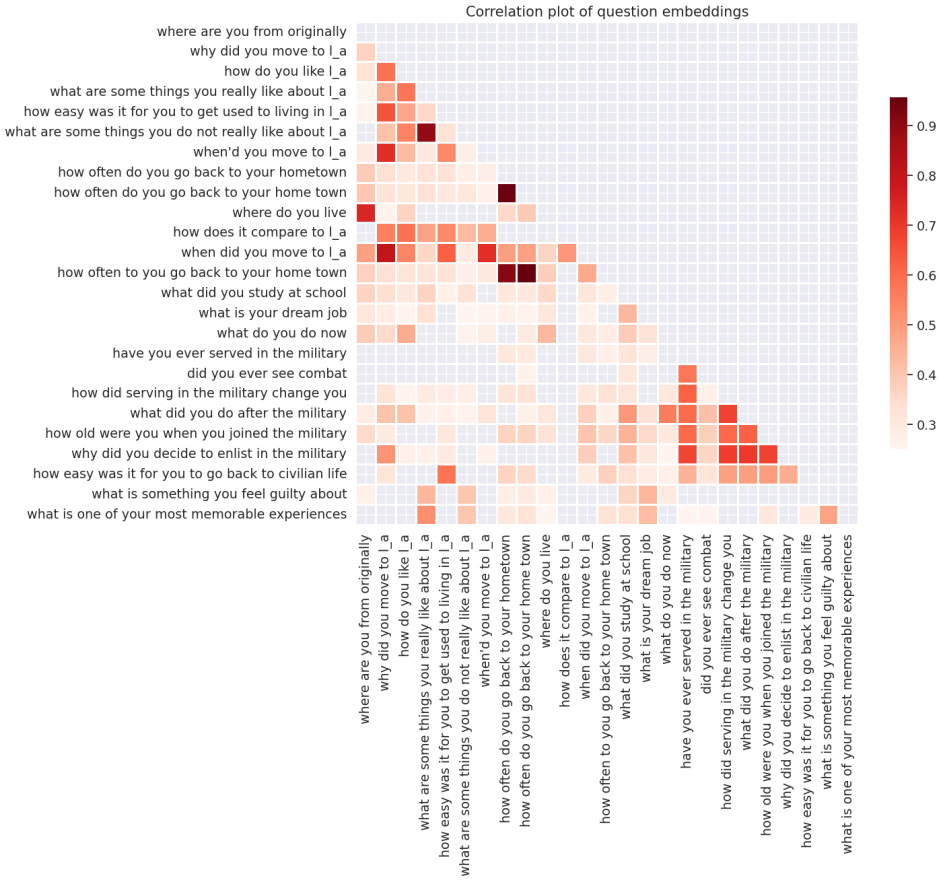


FIGURE 2.8: Correlation map between 25 first questions. Only shows correlation when higher than 0.25.

for all identified questions, and to locate similar questions, we employed the cosine similarity among the embeddings of all of them.

Subsequently, we identified the questions with more than 0.8 similarity to consolidate and group them under a single question, a correlation matrix for the first 25 questions is provided in **Figure 2.8**. This made the question dataset diminish to 85 questions. Nonetheless, USE proved to be sensible to highly similar structures, with minor changes, so a further manual review of the questions was necessary.

Once the preprocessing stage was completed, we quantified the instances of all the questions in each of the dataset's interventions. We then selected those questions that were represented in more than 100 samples and that seemed to be relevant according to literature. This process yielded a list of 13 questions, as depicted in **Table 2.1**. The table also demonstrates that the proportion of subjects diagnosed with depression remains consistent at around 30%. Furthermore, the table displays the average word count an intervention duration for each question. Notably, the first question, "How are you doing today," yields the shortest responses, as the participants generally respond with brief affirmatives like "good" or "fine".

Following the preprocessing steps, as transcriptions are manual, they include markers for sounds such as "laughs", "sniff" or "sigh". However, these markers were enclosed within "<>" which were subsequently filtered out so that the text models could process them correctly. After this cleaning process, we compiled all the

id	Question	Sample	Depressed	Mean words	Mean time
1	how are you doing today	184	29.9%	6.5	2.7s
2	when was the last time you argued with someone and what was it about	182	30.2%	91.2	39.9s
3	how are you at controlling your temper	176	30.7%	36.8	13.9s
4	what are you most proud of in your life	171	30.4%	56.9	27.2s
5	how easy is it for you to get a good night's sleep	172	29.1%	55.7	21.6s
6	have you been diagnosed with depression	167	25.1%	40.4	19.6s
7	have you ever been diagnosed with p_t_s_d	165	28.5%	15.0	7.4s
8	what did you study at school	162	31.5%	57.8	28.5s
9	how would your best friend describe you	163	28.2%	46.8	24.3s
10	how have you been feeling lately	160	30.6%	58.6	25.8s
11	what is your dream job	156	28.8%	58.1	27.8s
12	tell me about the last time you felt really happy	179	30.2%	62.6	29.8s
13	what would you say are some of your best qualities	101	39.6%	53.2	27.9s

TABLE 2.1: Question selection.

Question	**Answer**	**PHQ-8 result**
how are you doing today	mm okay'	Depressed
	huh overwhelmed. i have a funeral to attend tomorrow i found out from my doctor i got some health issues. it is hard. honey i am just putting one foot in front of the other and just trying to get it done. like i said i'm overwhelmed but i can't stop doing what i need to do	Healthy
when was the last time you argued with someone and what was it about	mm my girlfriend and insignificant	Depressed
	oh god when was the last time i really had a argument i don't know it's been awhile um argument it's been a long time that i can't even remember	Healthy
tell me about the last time you felt really happy	sigh ooh that's a good question it's been awhile i'd say year and a half maybe. hmm well there's been a lot going on in my life you know i just lost my parent my dad my last parent just a lot of stuff going on so i'll leave it at that	Depressed
	oh god well there's this guy at church and i really like him and he likes me and he throws kisses at me so	Healthy
what are you most proud of in your life	like i said my kids i'm very proud of 'em	Depressed
	oh my accomplishments acc my accomplishments um overcoming 'em one by one setting a goal and reaching 'em"	Healthy

TABLE 2.2: Example answers for depressed and healthy participants

text provided by the participants in response to the specific question, excluding Ellie's interventions, some of these answers can be seen in **Table 2.2**, and will serve as input to models.

The finalized dataset incorporated the start and end timestamps of each question posed during the intervention. These timestamps served as indicators delineating the specific segments of audio data pertinent to each question that served as the data input for the audio models. For the audio modality, no additional preprocessing was deemed necessary, thereby preserving the integrity of the original audio data.

Chapter 3

Experiments and results

3.1 Metrics

This section will treat the development of experiments that have been performed on the dataset. There will be 3 sets of experiments, related with the modalities of the data an considering both the benchmark and the transformer models. At the end of each modality section we will discuss the according results measured in the following metrics:

- **F1-Score:** The F1-score is computed as the harmonic mean of precision (P) and recall (R), given by the formula:

$$F1 = 2 \times \frac{P \times R}{P + R}$$

  The F1-score considers the trade-off between correctly identifying positive instances and minimizing false positives

- **ROC-AUC Score:** The ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) score represents the area under the receiver operating characteristic curve, which measures the model's ability to distinguish between positive and negative samples.

- **Recall:** Recall, also known as true positive rate or sensitivity, is calculated as the proportion of actual positive samples that are correctly classified by the model, using the formula:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

3.2 Text models

3.2.1 ML models

Literature commonly uses the LIWC library to perform word-class analysis, however the dictionary of words is not open-source. For this, we employed the open-source Python library Empath to perform a similar linguistic exploration of the transcriptions. Using Empath (Fast, Chen, and Bernstein, 2016), we computed groups of words within the transcriptions, as exemplified in **Figure 3.1**, and in fact we can appreciate some differences between groups, starting with the most common word group, which is positive emotion in healthy subjects and negative in depressed ones.

Additionally, following the literature, we calculated several other features, such as the number of utterances, frequency of first-person pronouns, occurrences of first and past tense verbs, and count of adjectives. This process generated a dataset with 199 variables.

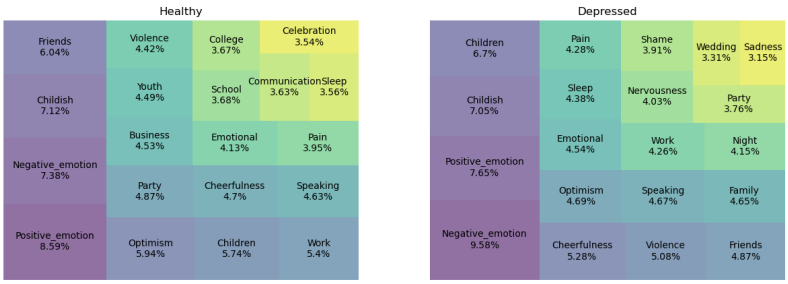


FIGURE 3.1: Top 15 Word groups for depressed and healthy subjects

To further process the data, we employed a pipeline in scikit-learn (Pedregosa et al., 2011) which involved a StandardScaler to normalize the features, followed by Principal Component Analysis (PCA) to retain 90% of the variance in the data of each of the questions

Subsequently, we performed training using three state-of-the-art machine learning models: Support Vector Machine (SVM), Random Forest (RF), and logistic regression (LR). To ensure robust evaluation, we employed a 3-fold grid search cross-validation strategy. Each question-dataset was divided into training and testing sets, with a test set size of 15%. We balanced the distribution of the PHQ-8 label, gender, and intervention length (measured by the number of words) in both the training and testing sets. In order to address class imbalance, we applied random oversampling to the positive class, as recommended in the literature.

3.2.2 Text Transformers

Our research utilized the HuggingFace (Wolf et al., 2020) library to access pre-trained Transformer models. For the study we have considered both BERT and RoBERTa for our experiments.

We used 'roberta-base', a RoBERTa version with 12 transformer encoder layers in the Transformer architecture, the large version has 24. Additionally, we used 'RobertaForSequenceClassification', an expanded version of the base model, stacked with two additional layers on top of the output embeddings, designed for classification tasks. Given our dataset's relatively small size, we opted to freeze the RoBERTa architecture, allowing only the final dense layers to adjust their parameters. This version of RoBERTa operates using the first token of the sequence, equivalent to BERT's [CLS] token, which encapsulates sufficient information for the classification task and yields a 768-dimensional embedding. Then the first layer makes a transformation from 768 to 256 and then the model outputs the prediction.

We upheld balance in our dataset by accounting for variables like the PHQ8 label, gender, and word count. We allocated 15% of the dataset for testing, and from the remaining 85%, we reserved another 15% for validation. Similar to our machine

learning strategy, we applied oversampling to the training set to ensure balance for the positive class.

To minimize the influence of randomness on our results, we trained three models, each initialized with different weights, on the same dataset. This strategy enabled us to report the mean score for each question across the models, providing a more robust indicator of model performance. It is important to note, however, that we did not execute cross-validation in this study due to hardware constraints.

Text model underwent training using a learning rate of 2e-3 and a batch size of 64 for a maximum of 50 epochs. We employed the Adam optimizer with weight decay, setting ϵ at $1x10^{-8}$ to improve the numerical stability of the optimization process.

To mitigate the risk of overfitting, we implemented an early stopping mechanism, monitoring validation loss with a patience threshold of five iterations. Upon detecting an increase in validation loss, the mechanism activates a patience counter. Concurrently, the training process also monitors the validation F1 score, ensuring the best performing model, in terms of F1 score, is retained at all times. This strategy allows for the optimal model to be retrieved upon the conclusion of the training process.

RoBERTa's output embeddings, which we mapped to two dimensions using a PCA with two components, are shown in **Figure 3.2**. Notably, there is no clear embedding from which we could anticipate highly accurate results and it is plausible that our models will primarily learn to minimize errors. For example, question 7 appears seemingly random, there is no cluster of data that can differentiate between classes, which is expected given the question is "What did you study at school". Conversely, question 5, "Have you ever been diagnosed with depression", suggests that a high recall score may be attainable, as positive classes are grouped at the right, albeit with potentially lower specificity due to the distribution of data points.

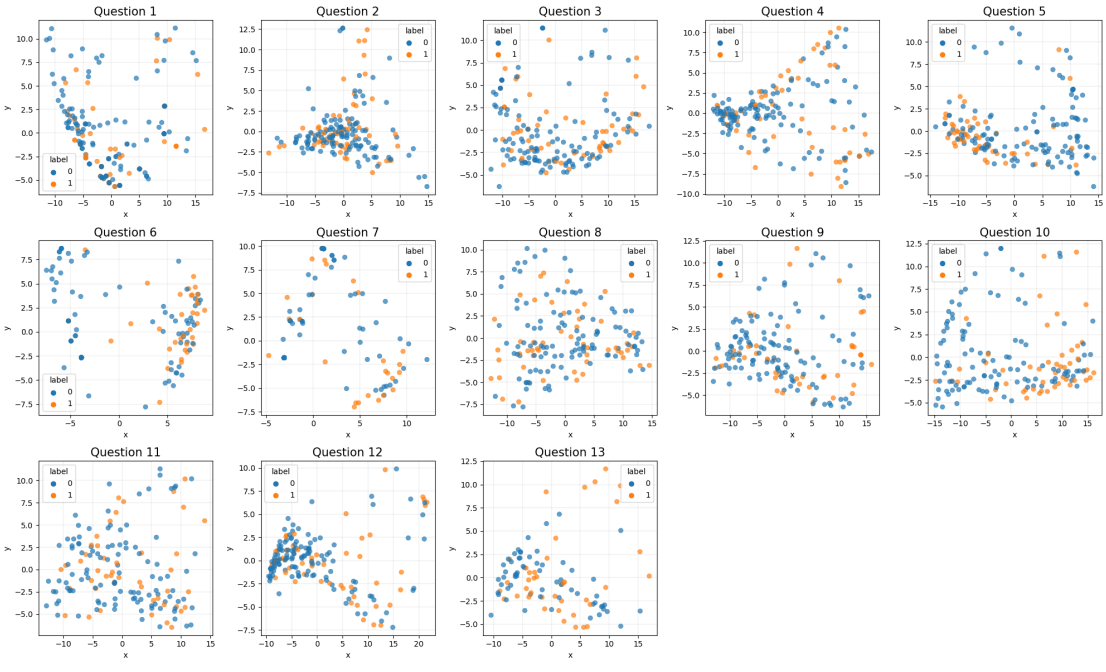


FIGURE 3.2: Roberta Embeddings

3.2.3 Results

Question	Model	Validation			Test		
		Recall	AUC	F1	Recall	AUC	F1
1. How are you doing today	Classic ML	0.65	0.61	0.48	0.63	0.55	0.41
	BERT	0.76	0.72	0.61	0.75	0.78	0.58
	RoBERTa	0.70	0.76	0.62	0.75	0.78	0.52
2. When was the last time you argued with someone.	Classic ML	0.21	0.51	0.26	0.25	0.48	0.25
	BERT	1.00	0.47	0.48	0.16	0.58	0.16
	RoBERTa	0.70	0.41	0.53	0.41	0.58	0.40
3. How are you at controlling your temper	Classic ML	0.82	0.67	0.56	0.56	0.58	0.48
	BERT	0.92	0.47	0.45	0.63	0.55	0.44
	RoBERTa	1.00	0.53	0.46	0.14	0.59	0.18
4. What are you most proud of in your life	Classic ML	0.56	0.61	0.50	0.75	0.68	0.54
	BERT	0.44	0.54	0.55	0.58	0.66	0.51
	RoBERTa	0.85	0.73	0.48	0.38	0.53	0.34
5. How easy is it for you to get a good night's sleep	Classic ML	0.56	0.66	0.49	0.86	0.76	0.60
	BERT	0.74	0.68	0.60	0.81	0.717	0.64
	RoBERTa	0.68	0.73	0.57	0.81	0.68	0.60
6. Have you been diagnosed with depression	Classic ML	0.25	0.48	0.32	0.33	0.64	0.44
	BERT	0.62	0.76	0.59	0.83	0.77	0.65
	RoBERTa	0.75	0.67	0.61	0.66	0.77	0.54
7. Have you ever been diagnosed with PTSD	Classic ML	0.22	0.57	0.32	0.29	0.61	0.40
	BERT	1.0	0.59	0.48	0.28	0.45	0.33
	RoBERTa	1.0	0.57	0.48	0.19	0.58	0.24
8. What did you study at school	Classic ML	0.70	0.57	0.50	0.75	0.56	0.46
	BERT	0.82	0.70	0.55	0.70	0.54	0.49
	RoBERTa	0.81	0..60	0.52	0.54	0.63	0.47
9. How would your best friend describe you	Classic ML	0.61	0.59	0.48	0.43	0.58	0.40
	BERT	0.87	0.42	0.43	0.23	0.37	0.21
	RoBERTa	0.95	0.49	0.46	0.33	0.54	0.37
10. How have you been feeling lately	Classic ML	0.45	0.53	0.38	0.57	0.53	0.40
	BERT	0.79	0.70	0.56	0.68	0.69	0.54
	RoBERTa	0.70	0.82	0.69	0.33	0.55	0.28
11. What is your dream job	Classic ML	1.00	0.46	0.44	1.00	0.50	0.45
	BERT	1.00	0.54	0.49	0.29	0.52	0.18
	RoBERTa	0.91	0.56	0.50	0.29	0.58	0.7
12. Tell me about the last time you felt really happy	Classic ML	0.67	0.39	0.31	1.00	0.50	0.44
	BERT	0.90	0.58	0.5	0.25	0.49	0.22
	RoBERTa	0.83	0.76	0.66	0.50	0.67	0.49
13. What would you say are some of your best qualities	Classic ML	0.30	0.42	0.31	0.16	0.45	0.20
	BERT	0.75	0.64	0.68	0.22	0.64	0.18
	RoBERTa	1.00	0.52	0.62	0.28	0.57	0.20

TABLE 3.1: Results for Text models in validation and test sets

Table 3.1 outlines the results for both our machine learning benchmark models and the Transformer models, BERT and RoBERTa. Contrary to initial expectations, Transformer models do not consistently outperform their traditional ML counterparts. There are numerous possible explanations for this unexpected phenomenon.

One key consideration is the size of the training datasets. Traditional ML models are known to yield better results when trained on smaller datasets compared to complex models, potentially explaining their superior performance in our case. Furthermore, the limited size of our datasets restricts our ability to fine-tune the Transformer models effectively.

If the original dataset distribution used for training the Transformer model varies significantly from our DAIC-WOZ dataset, the model may not generalize well to

our specific task, this discrepancy in data distribution could be a contributing factor to the underperformance of the Transformer models. Moreover, this theory is further supported by the embeddings distribution that was mentioned in previous RoBERTa section.

Nonetheless, the visualization does suggest certain anticipated behaviors. For instance, Question 4, "how easy is it for you to get a good night's sleep," presents a test F1 score of 0.64 as per BERT's evaluation. Similarly, Question 5, "Have you ever been diagnosed with depression" procures relatively robust test results, with a 0.65 F1 score and a 0.77 AUC score according to BERT. However this last question will undoubtedly yield results, given the direct correlation between the question and the label to predict.

Furthermore, we notice that traditional ML models generally display higher resistance to overfitting compared to Transformers, which could be attributed to their respective training strategies. Traditional ML models employ grid-search cross-validation, ensuring that the datapoints in each fold differ and thereby allowing for variations in distributions. Transformers, however, train consistently with the same data distributions. Consequently, if the validation data happens to be "easier" to classify than the test data, the Transformer models are likely to struggle in accurately classifying the data.

3.3 Speech models

3.3.1 ML models

For processing speech variables, we used Librosa (McFee et al., 2015), an open-source Python library renowned for broad options on speech recognition functions. Utilizing this tool, we computed 40 Mel-Frequency Cepstral Coefficients (MFCCs), Root Mean Square (RMS) energy—an approximation of loudness—and pitch. This computation produced a total of 50 variables, which were subsequently incorporated into another Scikit-learn pipeline, primarily comprised of a StandardScaler for feature normalization.

Given that the dimensionality was not overwhelmingly high in the context of these speech variables, we did not utilize PCA for this phase. We replicated the same cross-validation strategy, model selection, train/test split proportions, and oversampling techniques as we employed for the text models.

3.3.2 Wav2Vec2

For the implementation of Wav2Vec2, we once again relied on the HuggingFace library. However, despite Wav2Vec2 offering a sequence classification model akin to RoBERTa, we opted for the Wav2vec2Model class, which solely returns the embedding matrix. This decision was primarily motivated by the high dimensionality of both the model and the audio features.

The average duration of the audio for each intervention is approximately 25 seconds. Consequently, when calculating the features, we account for the number of seconds multiplied by the sampling rate—16 kHz in the case of the DAIC-WOZ dataset. This process, in essence, means that our model is processing vectors of roughly 16000*25 in length. This significantly increases the computational load and memory usage.

To mitigate this issue, we adopted a strategy where the maximum length for each question audio intervention was based on the median length of its population, avoiding this way outliers. Moreover, in case a question is long in its nature, we capped the interventions at 25 seconds. Then, to optimize the training phase, we precomputed these embeddings and stored them, this way, when necessary they could be loaded from storage. This two-step process, comprising of the precomputation of the embeddings followed by the training of the dense layers responsible for classification, helped streamline the computational workload and effectively manage the memory resources.

The architecture adopted for Wav2Vec2 mirrors that used for RoBERTa, with one key distinction— we averaged the embedding vector along the time axis. The resulting data was subsequently fed into a linear layer, which transformed the 768-dimension Wav2Vec2 output embedding into a 256-dimension vector. This vector was then directed towards the prediction stage.

For the optimization process, we again utilized AdamW, retaining the same epsilon value as employed in the text transformer models. The learning rate was set at 10^{-3}, and we used a batch size of 32.

The EarlyStopping strategy previously described was also implemented here to monitor the validation loss and F1 score, and prevent overfitting. Similarly, the training process followed the structure established during the text transformer phase.

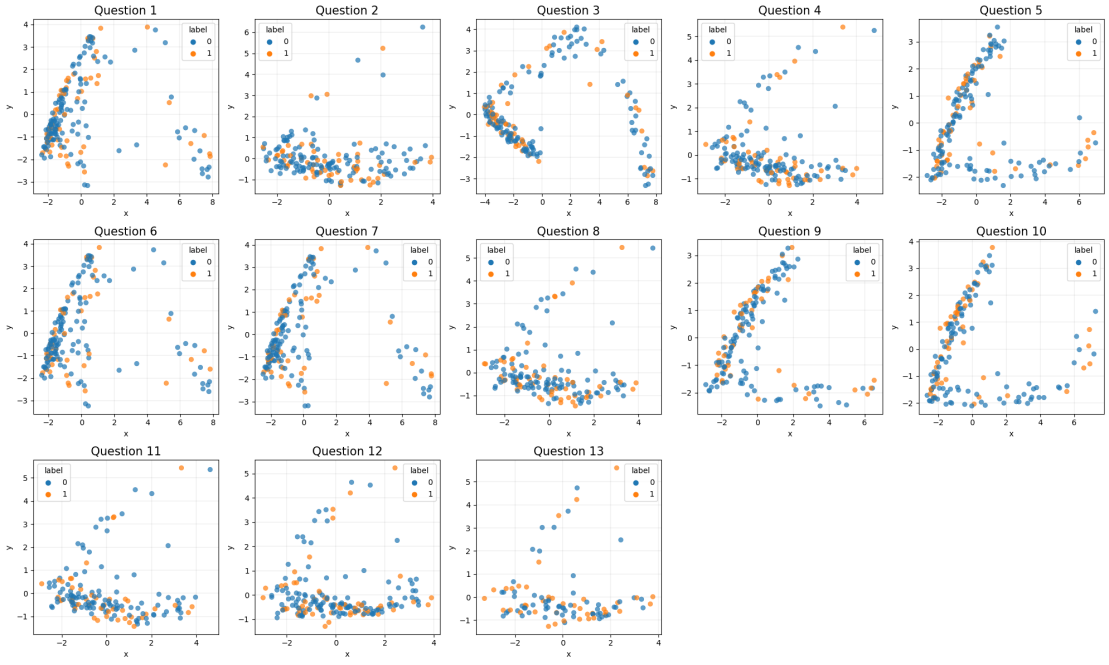


FIGURE 3.3: Wav2Vec2 Embeddings

In this particular scenario, we dismissed the use of 'wav2vec2-base', due to its underperformance. Instead, we pivoted towards a model trained specifically for an emotion recognition task, anticipating that it could yield superior results given the nature of our study.

The embeddings derived from our emotion-focused model are depicted in **Figure 3.3**. These visualizations display distinct differences when compared with the embeddings from the RoBERTa model. A noticeable feature is that the positive and negative classes do not appear to diverge clearly; rather, they tend to cluster within the same space.

This phenomenon could potentially stem from the same underlying issue we observed with text data. Given the complexity of audio signals, a larger dataset might be necessary to effectively train the model and obtain appropriate embeddings. Based on these visualizations and our preliminary analysis, we anticipate that the results from this model may not meet our initial expectations.

3.3.3 Results

		Validation			Test		
Question	**Model**	**Recall**	**AUC**	**F1**	**Recall**	**AUC**	**F1**
1. How are you doing today	Classic ML	0.53	0.57	0.45	0.63	0.60	0.45
	Wav2Vec2	0.73	0.73	0.65	0.17	0.55	0.22
2. When was the last time you argued with someone.	Classic ML	0.45	0.53	0.39	0.75	0.60	0.48
	Wav2Vec2	0.80	0.43	0.45	0.41	0.56	0.40
3. How are you at controlling your temper	Classic ML	0.47	0.53	0.40	0.66	0.67	0.57
	Wav2Vec2	0.70	0.60	0.46	0.33	0.40	0.29
4. What are you most proud of in your life	Classic ML	0.45	0.53	0.46	0.50	0.60	0.44
	Wav2Vec2	0.77	0.64	0.51	0.25	0.49	0.23
5. How easy is it for you to get a good night's sleep	Classic ML	0.59	0.62	0.48	0.71	0.67	0.50
	Wav2Vec2	0.70	0.51	0.46	0.19	0.71	0.23
6. Have you been diagnosed with depression	Classic ML	0.66	0.62	0.45	0.66	0.76	0.55
	Wav2Vec2	0.71	0.34	0.37	0.44	0.60	0.42
7. Have you ever been diagnosed with PTSD	Classic ML	0.60	0.62	0.38	0.42	0.53	0.35
	Wav2Vec2	0.68	0.63	0.55	0.76	0.78	0.57
8. What did you study at school	Classic ML	0.41	0.49	0.38	0.50	0.62	0.47
	Wav2Vec2	0.68	0.31	0.36	0.46	0.46	0.31
9. How would your best friend describe you	Classic ML	0.41	0.55	0.36	0.43	0.58	0.40
	Wav2Vec2	0.68	0.56	0.45	0.47	0.40	0.30
10. How have you been feeling lately	Classic ML	0.54	0.60	0.44	0.71	0.69	0.55
	Wav2Vec2	0.68	0.57	0.50	0.29	0.58	0.33
11. What is your dream job	Classic ML	1.00	0.47	0.44	1.00	0.50	0.45
	Wav2Vec2	0.91	0.45	0.47	0.29	0.38	0.29
12. Tell me about the last time you felt really happy	Classic ML	0.38	0.44	0.32	0.63	0.66	0.52
	Wav2Vec2	0.93	0.46	0.49	0.0	0.32	0.00
13. What would you say are some of your best qualities	Classic ML	0.59	0.57	0.52	0.33	0.48	0.33
	Wav2Vec2	0.75	0.296	0.50	0.29	0.42	0.28

TABLE 3.2: Results for speech models in validation and test sets

Indeed, the results obtained from this model, as seen **Table 3.2** confirm our anticipations—it significantly underperformed on the data. Surprisingly, the traditional ML models, despite being less explored in our study, demonstrated more promising results than the transformer model.

Literature frequently cites fine-tuning as a crucial step for improving performance with Wav2Vec2 models. Given this, it is plausible that our Wav2Vec2 model could benefit from a fine-tuning process, as the dataset it is pretrained is not conversational

as ours. However, we must highlight that the volume of data available is not sufficient for training a single modality effectively, and, in order to do so, we would lose the topic datasets, which ultimately could affect negatively as some questions and answers are noise that prevents the model from accurate predictions.

The nature of our dataset and the problem at hand makes it challenging to apply oversampling to generate synthetic samples without creating false expectations for the model, and, in fact, we hypothesize that the observed overfitting in the validation dataset could be attributed to random oversampling. The model might be prioritizing the classification of positive samples, thereby leading to an influx of false positives. This inference is supported by the low F1 scores in validation, compared to the recall scores, indicating a probable decrease in precision. As embeddings behave randomly, the model is just choosing between one class or another and not learning in most cases.

3.4 Multimodal models

3.4.1 Attention Model

For our multimodal model, we harvested the embeddings from both RoBERTa and Wav2Vec2 models to investigate whether the combination of both could enhance the former's performance. As singular performance was poor, we decided to implement a multihead attention mechanism with four heads for each data modality, following the extraction of the initial RoBERTa token and time-axis averaging of Wav2Vec2. We then conducted an element-wise multiplication with the original embedding to emphasize the significant features of each modality.

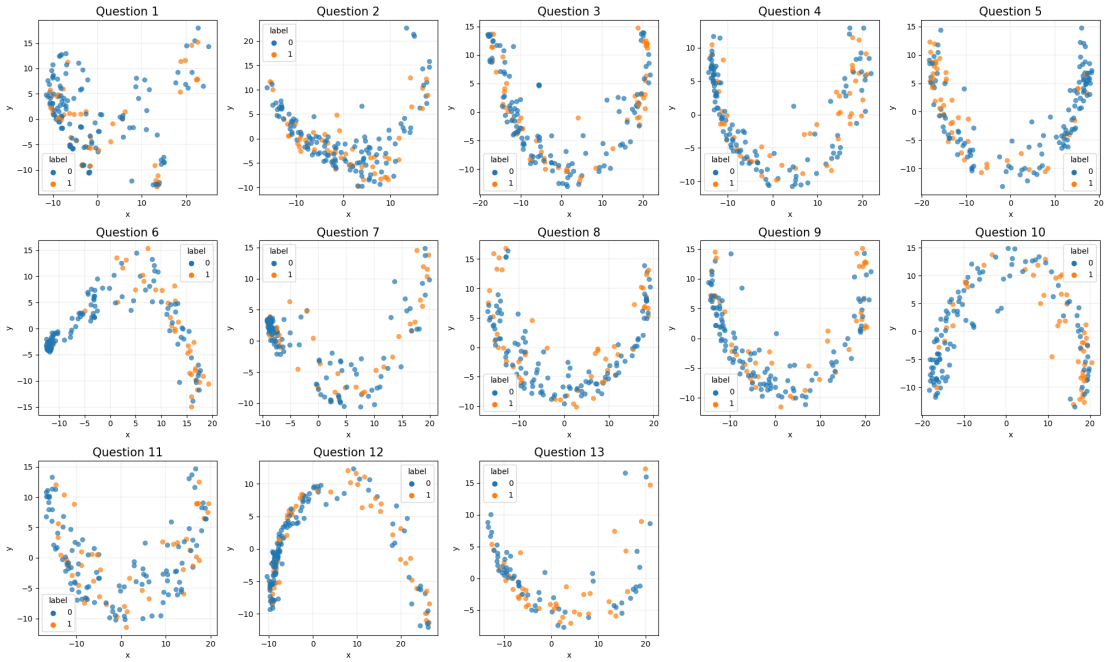


FIGURE 3.4: Multimodal Embeddings

Afterwards, akin to the transformer model's process, we employed an add-and-normalize operation on the data. The normalized data was then funneled into a projection layer, transforming the multimodal 768-dimensional embedding into a 256-dimensional representation, in line with the dimensions produced by the Wav2Vec2 and RoBERTa models. To mitigate overfitting, a dropout layer was inserted following the add-and-normalize stage along as the same Early Stopping strategy as in the other methodologies.

This multimodal model was trained using a learning rate of 10^{-4}, a batch size of 64, and the same AdamW optimizer as utilized in previous models. Similar strategies were adopted for balancing the training, validation, and test distributions, as well as for the overall training process.

In **Figure 3.4**, we visualize the model's embeddings after the add-and-normalize operation. The data distribution remains complex, though we discern instances where the positive class demonstrates a higher degree of clustering, notably in questions 4, 5, 9, and 11. Thus, we anticipate some predictive power from these models. However, akin to the challenges encountered with RoBERTa, if the test set incorporates examples that deviate significantly from the general cluster, the model might struggle to generate accurate predictions.

3.4.2 Results

The multimodal results are shown in **Table 3.3**. This approach exhibits superior predictive capabilities on the validation dataset, with improved F1 scores in 8 out of 13 questions compared to the standalone transformer models. This indicates that the new embeddings produced by combining two modalities can more effectively distinguish between classes than the individual Wav2Vec2 and RoBERTa models. However, it is possible that Wav2Vec2 introduces more distractors than beneficial data, so a decline in F1 scores from training to testing can be observed. This is equally attributable to both the challenges posed by the transformers and the increased difficulty of the test set. Indeed, the maintenance or even improvement of AUC scores in the multimodal model, compared to the standalone models, indicates a balanced performance in classifying each class. The AUC is a performance metric for binary classification problems and it represents the model's ability to correctly classify positive and negative examples at varying thresholds. A higher AUC suggests that the model has a good measure of separability, and is capable of distinguishing between the classes effectively. So, despite some shortcomings in F1 scores, the model's robust AUC scores provide assurance of its overall predictive accuracy.

Interestingly, it seems that the multimodal approach has particularly boosted the performance on certain questions. For instance, significant improvements were observed in responses to the questions "What would you say are some of your best qualities?" and "How have you been feeling lately?" Furthermore, "How are you doing today?" achieved the highest F1 score amongst all questions, despite being the one that typically receives the shortest responses. This highlights the potential of multimodal models in interpreting even succinct responses more effectively.

		Validation			Test		
Question	**Model**	**Recall**	**AUC**	**F1**	**Recall**	**AUC**	**F1**
1. How are you doing today	Wav2vec2	0.73	0.73	0.65	0.17	0.55	0.22
	RoBERTa	0.76	0.72	0.61	0.75	0.78	0.58
	Multimodal	0.79	0.75	0.61	0.64	0.77	0.65
2. When was the last time you argued with someone.	Wav2Vec2	0.80	0.43	0.45	0.41	0.56	0.40
	RoBERTa	1.00	0.47	0.48	0.16	0.58	0.16
	Multimodal	0.75	0.39	0.43	0.47	0.49	0.35
3. How are you at controlling your temper	Wav2Vec2	0.70	0.60	0.46	0.33	0.40	0.29
	RoBERTa	0.92	0.47	0.45	0.63	0.55	0.44
	Multimodal	0.63	0.64	0.60	0.31	0.42	0.29
4. What are you most proud of in your life	Wav2Vec2	0.77	0.64	0.51	0.25	0.49	0.23
	RoBERTa	0.44	0.54	0.55	0.58	0.66	0.51
	Multimodal	0.83	0.63	0.60	0.49	0.51	0.36
5. How easy is it for you to get a good night's sleep	Wav2Vec2	0.70	0.51	0.46	0.19	0.71	0.23
	RoBERTa	0.74	0.68	0.60	0.81	0.72	0.64
	Multimodal	1.00	0.93	0.84	0.61	0.82	0.56
6. Have you been diagnosed with depression	Wav2Vec2	0.71	0.34	0.37	0.44	0.60	0.42
	RoBERTa	0.62	0.76	0.59	0.83	0.77	0.65
	Multimodal	0.79	0.76	0.62	0.64	0.63	0.49
7. Have you ever been diagnosed with PTSD	Wav2Vec2	0.68	0.63	0.55	0.76	0.78	0.57
	RoBERTa	1.0	0.59	0.48	0.28	0.45	0.33
	Multimodal	0.48	0.61	0.38	0.50	0.59	0.39
8. What did you study at school	Wav2Vec2	0.68	0.31	0.36	0.46	0.46	0.31
	RoBERTa	0.82	0.70	0.55	0.70	0.54	0.49
	Multimodal	0.46	0.63	0.46	0.44	0.53	0.39
9. How would your best friend describe you	Wav2Vec2	0.68	0.56	0.45	0.47	0.40	0.30
	RoBERTa	0.87	0.42	0.43	0.23	0.37	0.21
	Multimodal	0.76	0.68	0.59	0.39	0.52	0.34
10. How have you been feeling lately	Wav2Vec2	0.68	0.57	0.50	0.29	0.58	0.33
	RoBERTa	0.79	0.70	0.56	0.68	0.69	0.54
	Multimodal	0.71	0.84	0.83	0.83	0.75	0.56
11. What is your dream job	Wav2Vec2	0.91	0.45	0.47	0.29	0.38	0.29
	RoBERTa	1.00	0.54	0.49	0.29	0.52	0.18
	Multimodal	0.71	0.46	0.52	0.24	0.52	0.27
12. Tell me about the last time you felt really happy	Wav2Vec2	0.93	0.46	0.49	0.0	0.32	0.00
	BERT	0.90	0.58	0.5	0.25	0.49	0.22
	Multimodal	0.67	0.69	0.59	0.33	0.50	0.32
13. What would you say are some of your best qualities	Wav2Vec2	0.75	0.30	0.50	0.29	0.42	0.28
	RoBERTa	0.75	0.64	0.68	0.22	0.64	0.18
	Multimodal	0.72	0.65	0.61	0.52	0.75	0.58

TABLE 3.3: Results for multimodal model in validation and test sets

3.5 Multimodal for other classification tasks

3.5.1 Mild Depression inclusion

Upon thorough examination of the results obtained from various Transformer models, it becomes evident that the Multimodal model exhibits more robust performance on the validation set. As mentioned earlier, all consequential decisions will be based on the performance in this particular set. Therefore, we were impelled to explore different configurations of the PHQ label. The DAIC-WOZ dataset offers the participants' scoring data from the PHQ8 questionnaire, where a score equal to or greater than 10 serves as the cut-off for identifying depression.

However, it is essential to consider that according to the PHQ authors, scores ranging from 5 to 10 could suggest the presence of mild depression. In light of this, we

opted to plot the RoBERTa embeddings - which have shown more promise than their Wav2Vec2 counterparts - to analyze the distribution of varying degrees of depression severity. The results are illustrated in **Figure 3.5**.

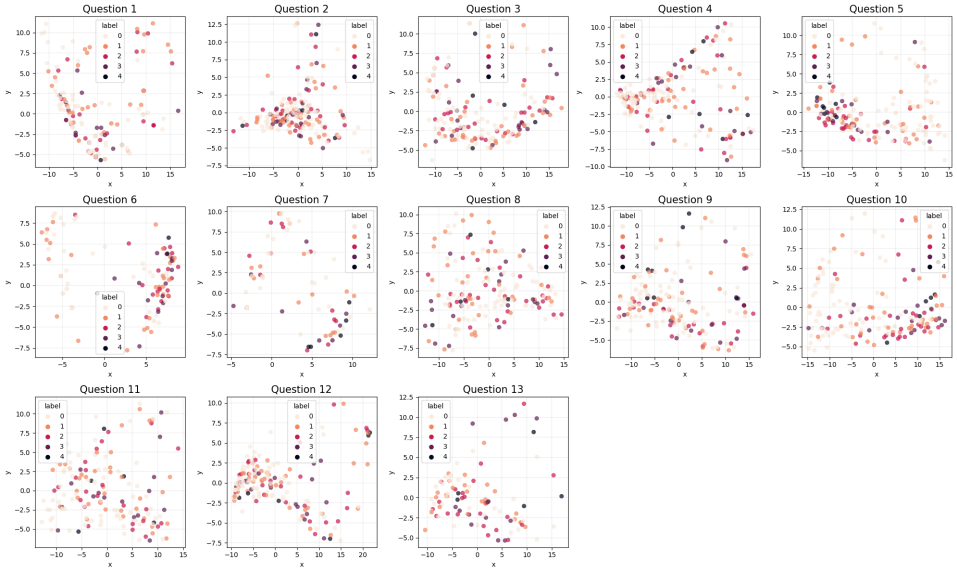


FIGURE 3.5: RoBERTa Severity Embeddings

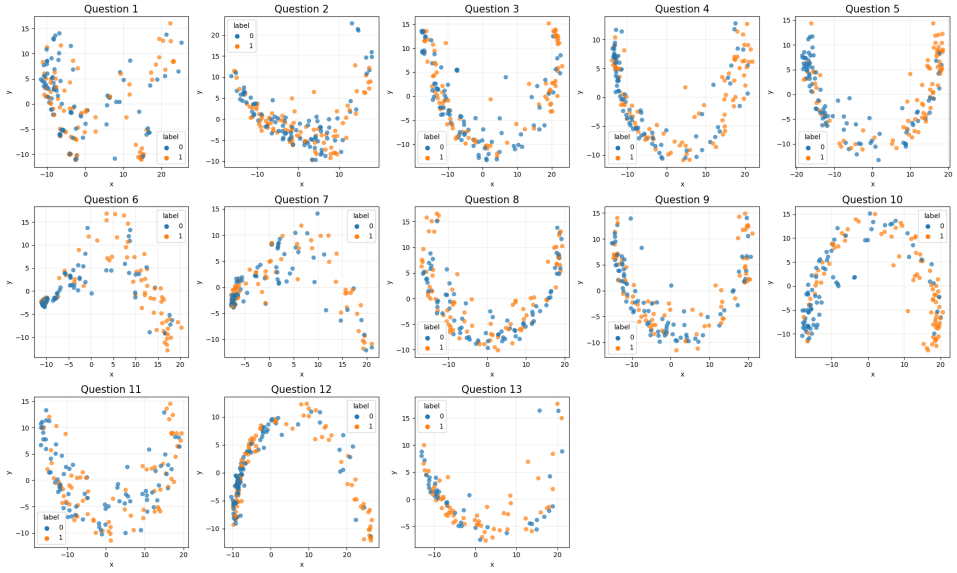


FIGURE 3.6: Multimodal embeddings including mild depression cases

In the depicted figure, the labels signify the severity of depression, where 0 indicates a healthy individual, 1 indicates mild depression, and the remaining signify varying levels of severity, the darker the more severe. The visual representation suggests similar clustering between the mild and severe groups. This observation implies that the points of confusion in the original binary classification might originate from participants experiencing mild depression. Consequently, we were motivated to train a multimodal model with the capacity to not only detect severe depression but to

	Validation			Test		
Question	**Recall**	**AUC**	**F1**	**Recall**	**AUC**	**F1**
1. How are you doing today	0.82	0.68	0.76	0.73	0.78	0.74
2. When was the last time you argued with someone.	0.98	0.57	0.76	0.70	0.43	0.57
3. How are you at controlling your temper	0.81	0.72	0.72	0.61	0.60	0.61
4. What are you most proud of in your life	0.89	0.76	0.80	0.75	0.63	0.67
5. How easy is it for you to get a good night's sleep	0.71	0.75	0.75	0.70	0.77	0.71
6. Have you been diagnosed with depression	0.72	0.70	0.72	0.62	0.61	0.65
7. Have you ever been diagnosed with PTSD	1.00	0.50	0.69	0.78	0.52	0.68
8. What did you study at school	0.92	0.59	0.71	0.50	0.49	0.52
9. How would your best friend describe you	0.56	0.33	0.48	0.47	0.56	0.53
10. How have you been feeling lately	0.88	0.79	0.82	0.71	0.76	0.74
11. What is your dream job	0.77	0.46	0.61	0.68	0.74	0.67
12. Tell me about the last time you felt really happy	0.84	0.70	0.77	0.73	0.67	0.68
13. What would you say are some of your best qualities	1.00	0.53	0.81	0.96	0.65	0.79

TABLE 3.4: Results for Multimodal model including mild depression

incorporate detection of mild cases as well. In this pursuit, we merely adjusted the batch size, reducing it to 32.

Altering the labels in this manner also effectively eliminated the issue of imbalance, resulting in a more equitably distributed sets. In **Figure 3.6**, we display the trained embeddings derived from the attention mechanism. The depiction illustrates a marked enhancement compared to the severe multimodal model. In questions that are inherently discriminative in severe cases, we can observe significant improvements. For instance, question 10 exhibits a clear clustering on the right for class 1, thus showing the progress made by this adjustment, we can expect then, improved results.

3.5.2 Results

The outcomes indeed corroborate our hypothesis, showcasing noticeable enhancements when compared to models solely detecting severe depression, as evident in **Table 3.4**. This suggests that mild depression is more akin to severe depression than it is to non-depressed individuals. Consequently, we infer that the true challenge lies

in assessing depression severity rather than simply identifying its presence. Furthermore, our experiment indicates that certain questions may be more sensitive to variations in severity than to the existence of depression itself. For instance, question 13 yielded the highest F1 scores in this model, yet its importance was diminished when only severe depression was considered. Conversely, question 1 maintained its relevance across models, underscoring its utility in assessing this mental health condition.

It is reasonable to surmise that these results could be further improved with an increase in data volume and fine-tuning of the models, as previously suggested. Furthermore, as the proportion of depressed subjects increased, it appears that for certain questions where the model lacked sufficient predictive power (for example, question 8), the model defaulted to classifying all responses as 1. This trend is substantiated when comparing recall rates between validation and test sets.

Chapter 4

Conclusions and future work

4.1 Conclusions

Our exploration into the application of Transformer models, classic machine learning methods, and multimodal approaches for depression classification in the DAIC-WOZ dataset provided meaningful insights.

First, the usage of pre-trained Transformer models, particularly RoBERTa and BERT, did not necessarily yield superior results compared to traditional ML models. This may have been due to the disparity between the original training data for these Transformer models and the DAIC-WOZ dataset. The lower performance of Transformers could also be linked to the constraints imposed by our dataset size, which limited the extent of fine-tuning we could perform on these models.

The visualizations of RoBERTa, and more importantly, Wav2Vec2 embeddings pointed to a potential lack of differentiation between positive and negative classes in the data representation. This suggested that even though Transformers could extract high-level features from the data, these features might not always be discriminative enough for specific tasks.

Our classic ML models demonstrated robustness against overfitting and presented competitive performance, despite their simplicity compared to Transformers, and it remarks the idea that it is not always necessary to use overly complex models. It's important to highlight the role of grid-search cross-validation here, which likely contributed to this robustness by ensuring the models were evaluated on diverse data distributions, avoiding the case of hard test samples.

The application of Wav2Vec2 to speech data didn't yield expected results. While we utilized a model trained for emotion recognition, hoping it would better handle our dataset, the results remained subpar. This underperformance could have been due to the complexity of audio data and the limitations posed by our data size.

Our multimodal approach attempted to blend text and audio data to improve results. While the combined model offered improved validation scores, the improvement did not consistently translate to the test set. However, promisingly, this approach maintained or even improved AUC scores, indicating a balanced classification of each class.

Taken together, our findings suggest that while Transformer models hold promise for depression analysis tasks, the specific characteristics of the dataset and task at hand greatly impact their performance. Classic machine learning models should not

be discounted, as they can provide competitive results and demonstrate robustness in diverse settings.

Furthermore, our results underscore the potential and challenges of multimodal learning. The combination of different modalities can indeed boost model performance, as evidenced by certain questions that showed improved F1 scores. However, ensuring this improvement is generalized across diverse test instances remains a challenge.

Additionally, we have determined that the primary challenge of this study lies in assessing the severity of depression. By considering depression as a broad category, we were able to achieve improved results when considering both severe and mild cases. This suggests a potential to reframe this issue as a multilabel problem, which could offer a more nuanced understanding of the distinct degrees of depression. Interestingly, our research has also suggested that RoBERTA demonstrated sensitivity towards the different severity levels, as evidenced by the variation in the default embeddings corresponding to the diverse depression degrees.

Additionally, we could take into consideration different topics that demonstrate higher sensibility for detecting depression from spontaneous speech, which are important assets to take into consideration. It is also important to mention that question 4, about sleeping habits, is present also on the PHQ8 test, which might be the reason for its acceptable results.

Besides this, we can also say that we have found similar results as in topic-based state-of-the-art depression detection, though exact comparison is hard due to the variability of result reporting. However, we were able to asses differences and similarities between classic ML models and Transformers by setting similar training and evaluation frameworks.

4.2 Future work

This is study will continue its course at AcceXible, which is the company that has given support to this project. In this matter, we will continue investigating on multiple study pathways and how to improve these results. These are some of the possible ways of improving the current methodology:

- **Fine-tune text and audio models**. We are using only 13 questions out of all the dataset. The discarded questions could be used to fine tune the transformer models in order to better fit the data coming from DAIC-WOZ dataset. This could even expand more in the case of audio, due to the fact that each audio could be broken into several pieces using a sliding window, similar to what it is done in Sardari et al., 2022.
- **Consider multiclass classification**. The problem itself might be too complex in nature, but taking into account the different severity degrees presented in the PHQ8 paper we might be able to break down the problem into several labels and evaluate the differences between them.
- **Explore classic feature extraction**. If transformers do not result in acceptable performances, the natural step would be to explore in depth ML models. We

used basic computations as benchmarks, but it is possible that with more sophisticated methodologies we could improve the results. Moreover, it is possible that taking in consideration the mild severity of the participant, the results on ML models might improve even more.

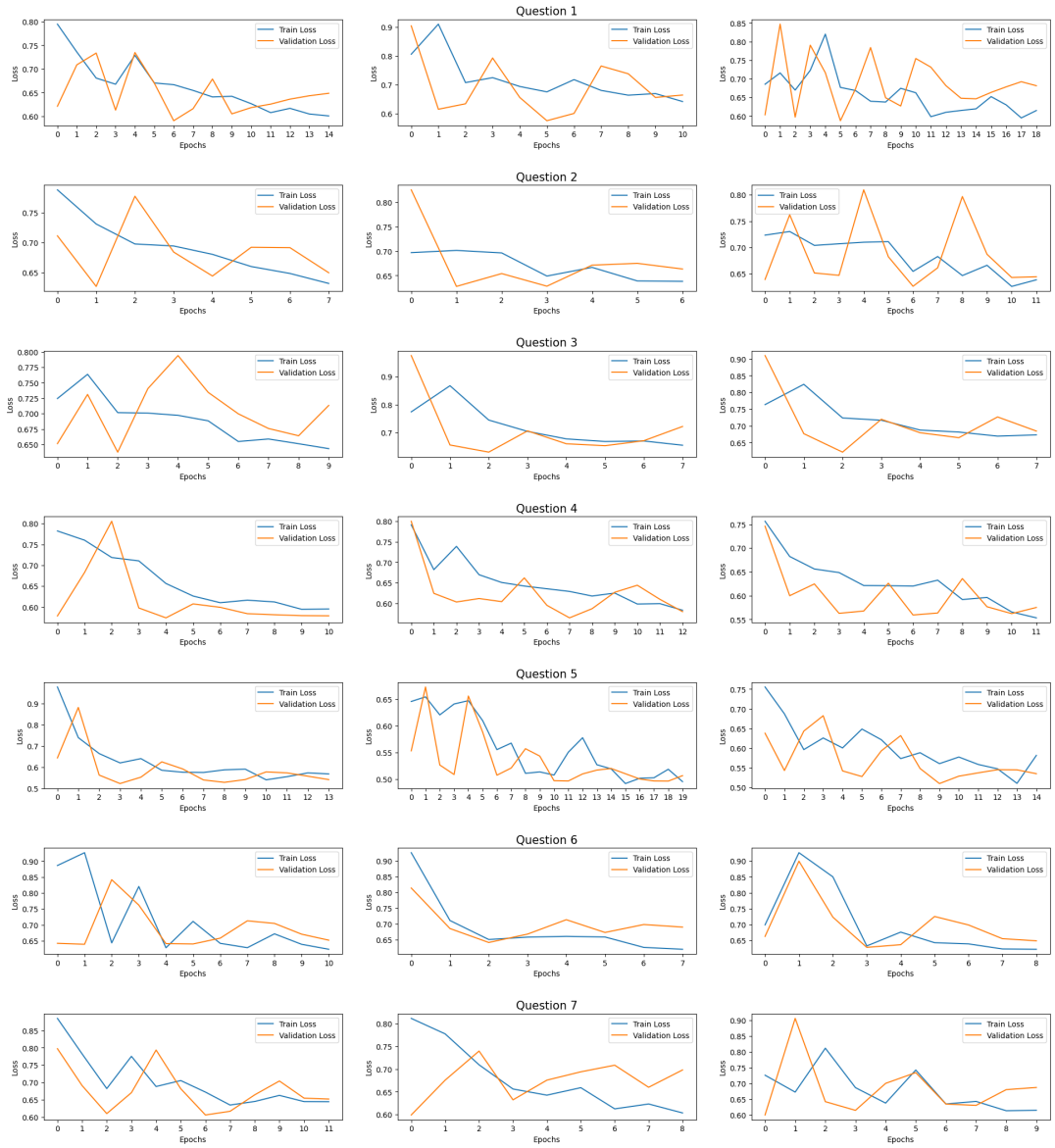
- **Further data cleansing**. It is possible that some questions are not perfectly collected, as we have used a semi-supervised strategy to retrieve them. It could be interesting to, now that we have labeled a big enough group of questions, train a basic model that learns to recognise them in order to further curate the data that is fed into the models.
- **Explore LSTMs** The scope of this project did not take into consideration LSTMs, however, given the nature of our data, it could be of interest to take into account this family of models. We are working with short sequences, so memory loss would not suppose a problem and moreover, they are not as dependant on data quantity as Transformers.
- **Explore other Transformer architectures** Here we used the state of the art on audio and text to perform our analysis. However, there are novel architectures such as Whisper, (Radford et al., 2022), which might yield more promising features.
- **Explainabilty**. Probably the most important topic to take into consideration considering we are treating with health data science. It is compulsory to know and understand the behaviour of the transformer models, as with generalization, comprensibility is also lost. It could start by checking which tokens RoBERTa is taking more in consideration, or what piece of each audio is the most important for Wav2Vec2. The first step towards this goal could be an analysis on how gender, the only sociodemographic variable DAIC-WOZ provides, affects our classification.

In conclusion, each potential direction for future research is driven by a common goal: enhancing the accuracy of depression detection models. There is considerable potential for further investigation in fine-tuning models, exploring different feature extraction methods, and experimenting with other Transformer architectures, amongst others.

Appendix A

Training and validation loss graphics

A.1 RoBERTa



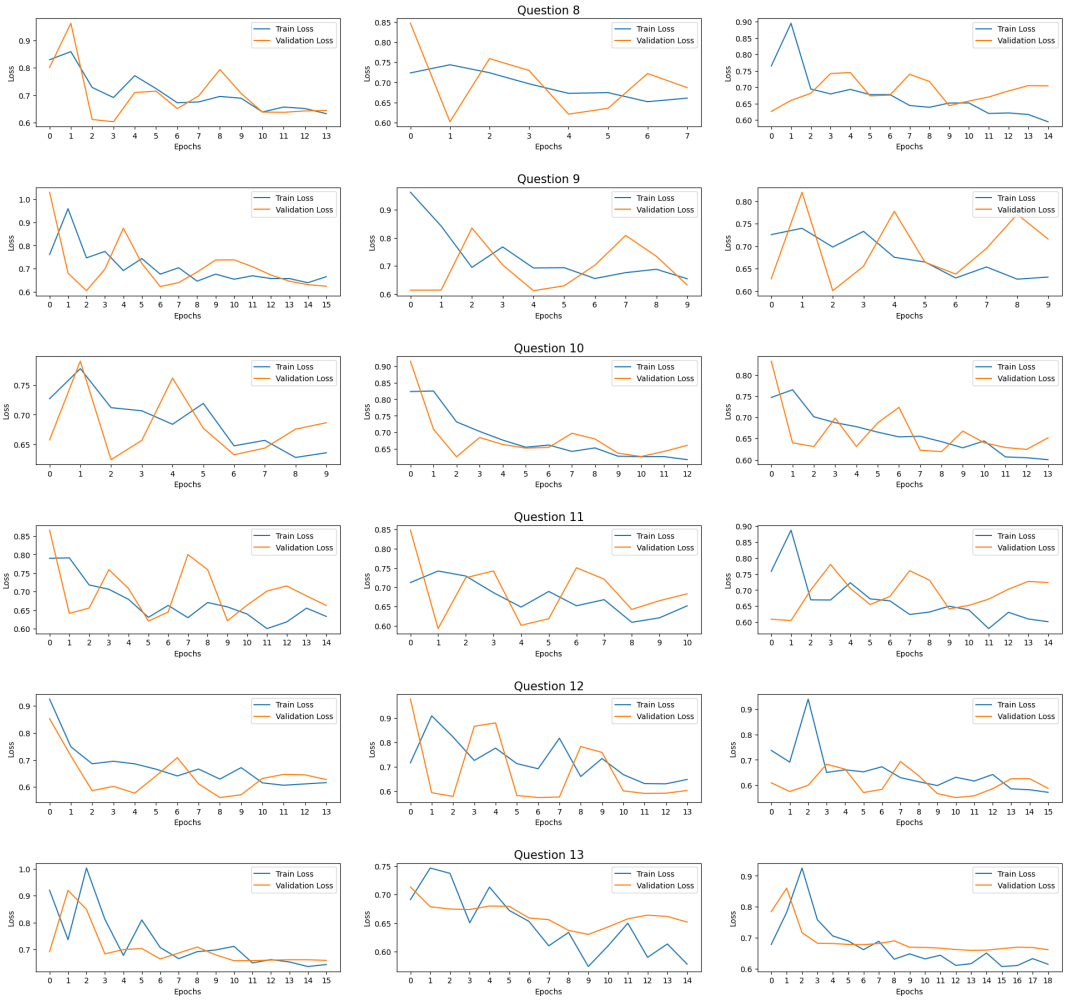
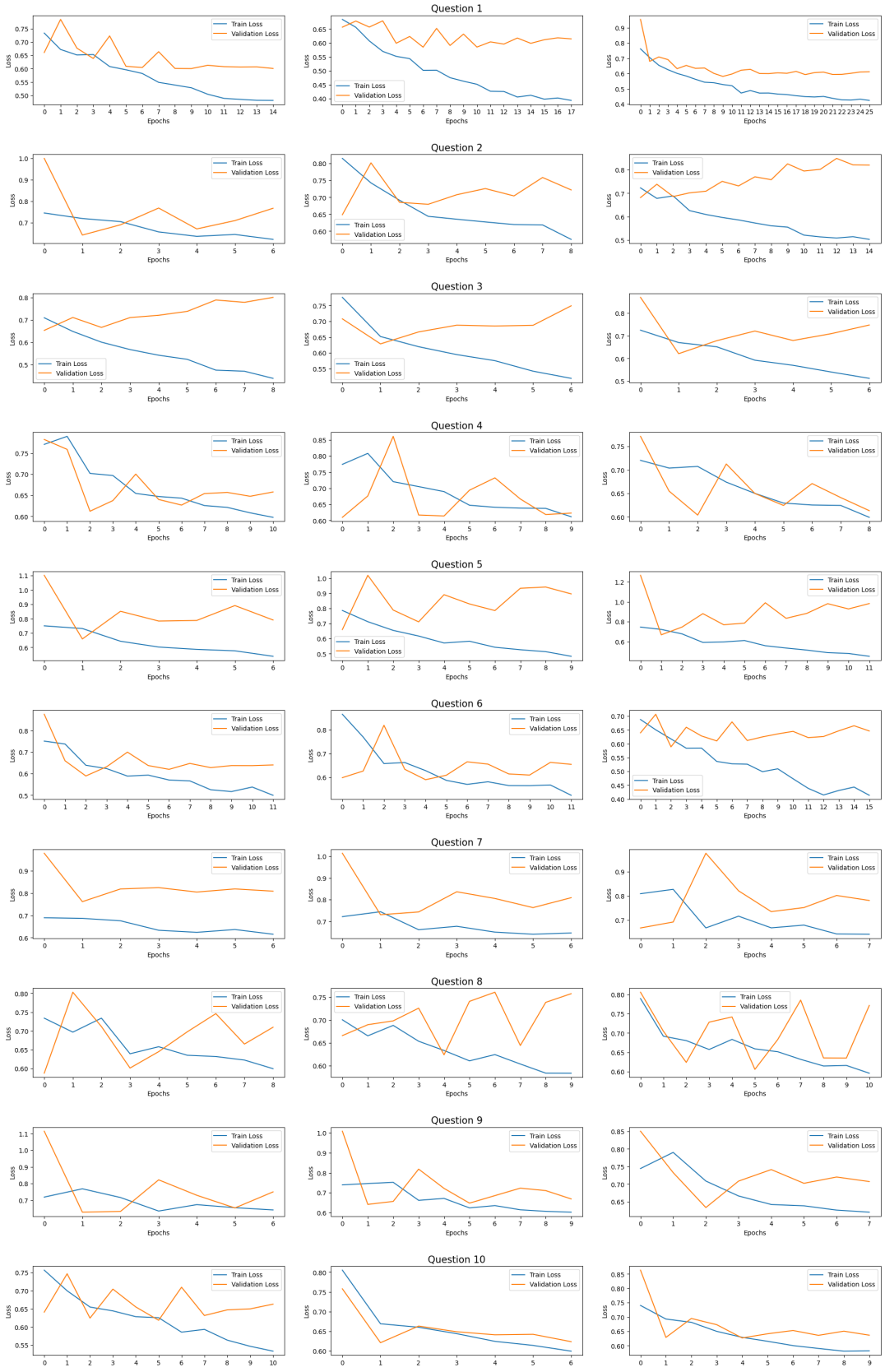


FIGURE A.1: RoBERTa loss

A.2 Wav2vec2


Question 1
Question 2
Question 3
Question 4
Question 5
Question 6
Question 7
Question 8
Question 9
Question 10
Train Loss
Validation Loss
Loss
Epochs

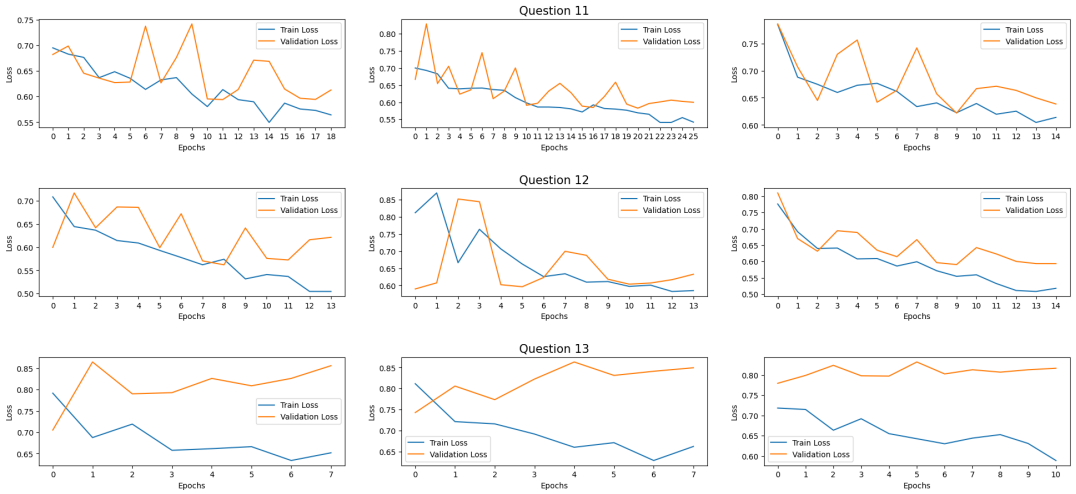
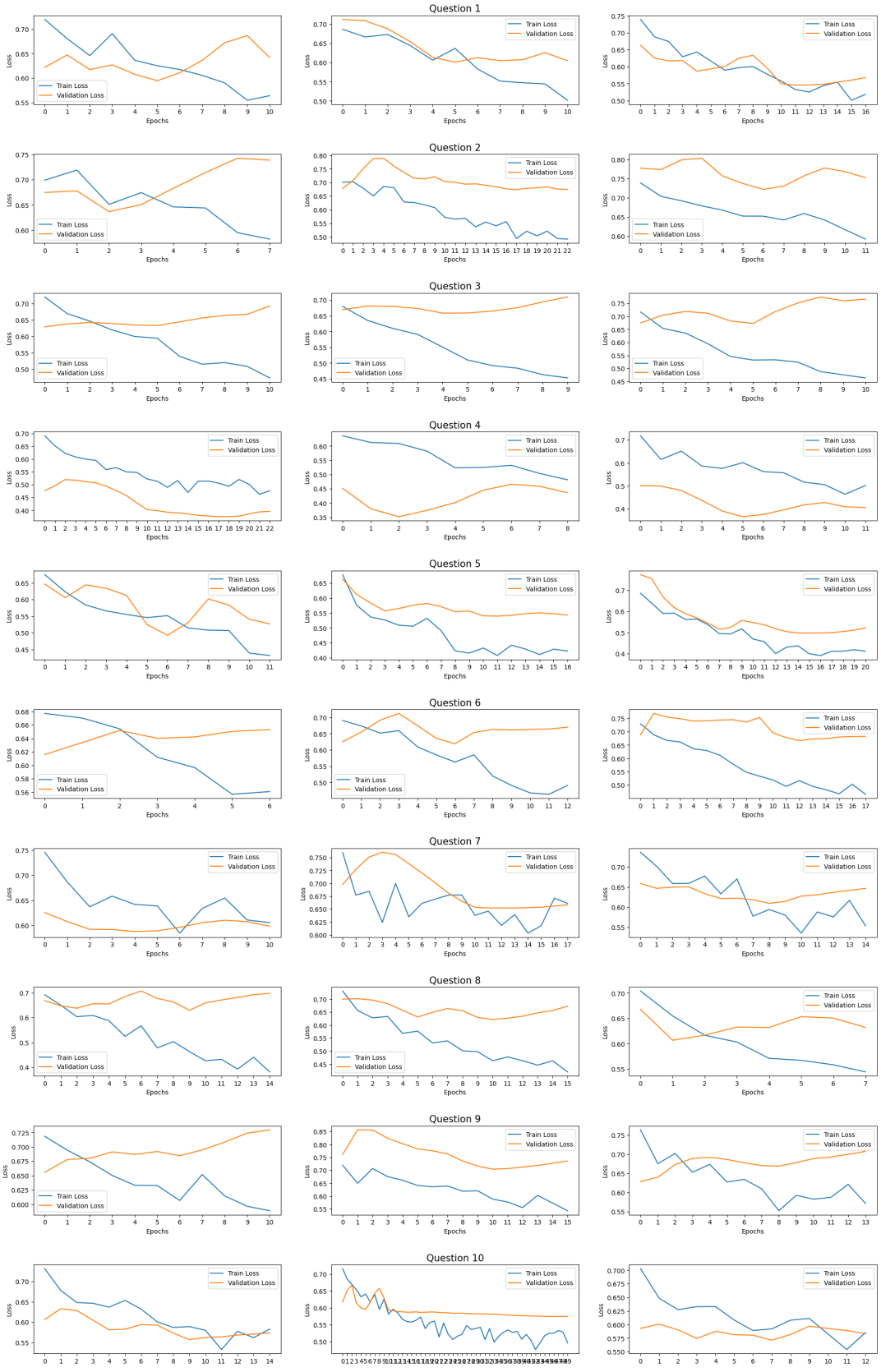


FIGURE A.2: Wav2Vec2 loss

A.3 Multimodal



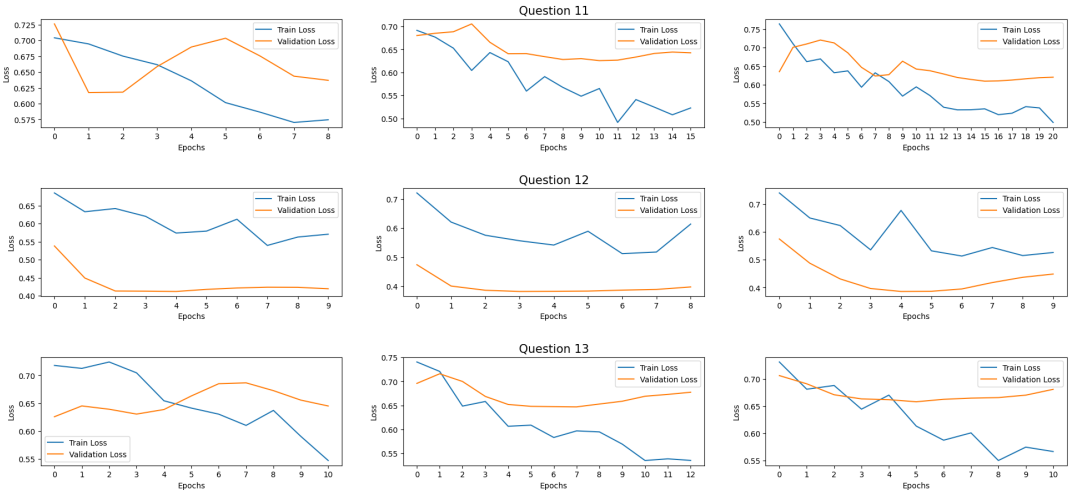
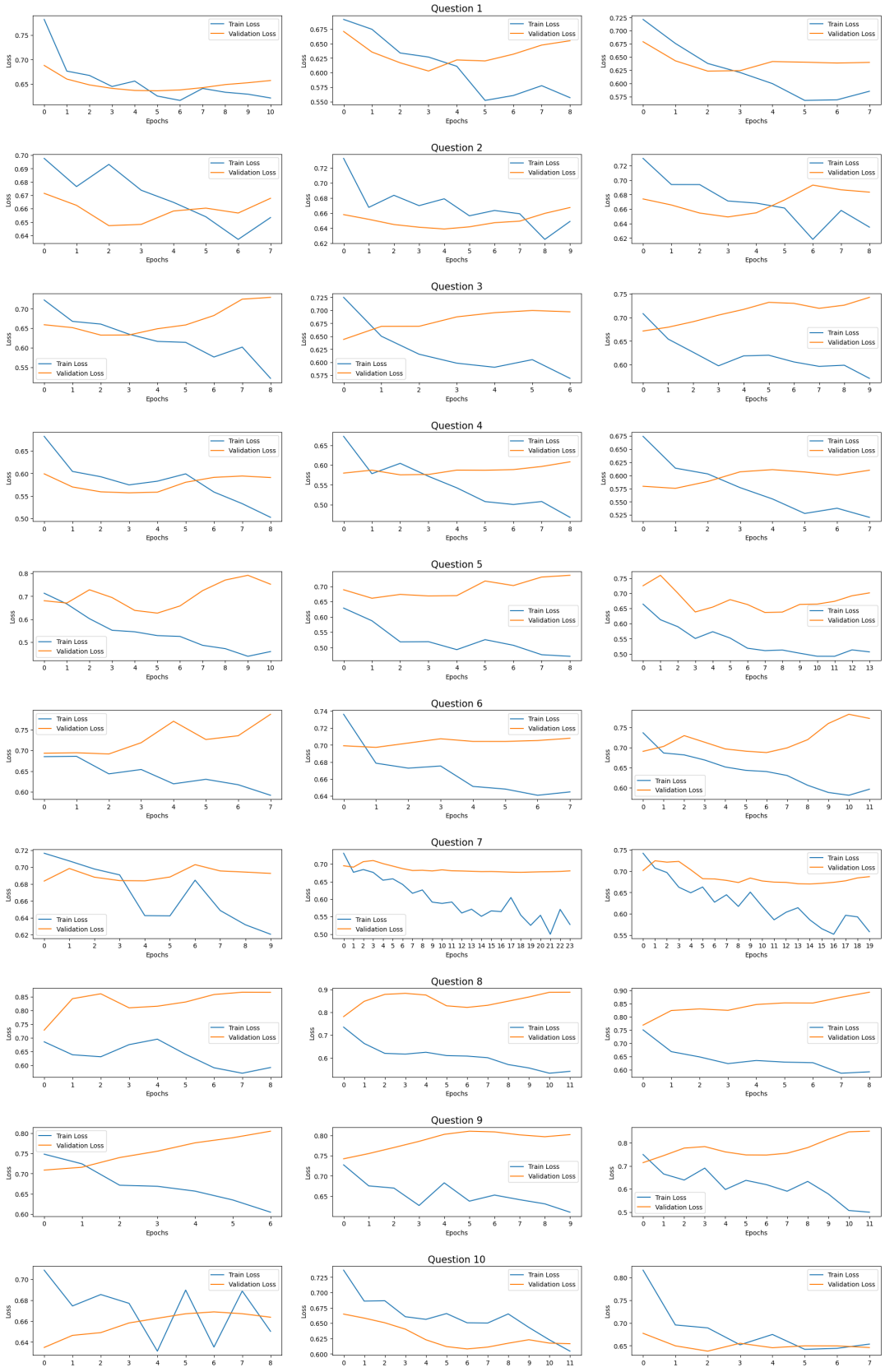


FIGURE A.3: Multimodal loss

A.4 Mild Multimodal



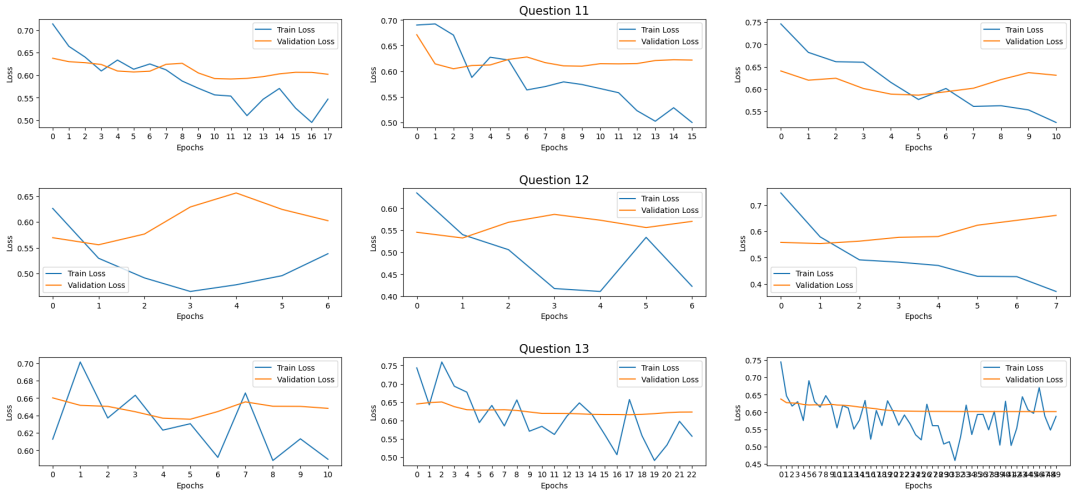


FIGURE A.4: Mild Multimodal loss

Bibliography

Alghowinem, Sharifa et al. (Jan. 2012). "From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech." In.

— (Oct. 2013). "Detecting Depression: A Comparison between Spontaneous and Read Speech". In: DOI: 10.1109/ICASSP.2013.6639130.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders*. 5th ed. Washington, D.C.: American Psychiatric Association.

Baevski, Alexei et al. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. arXiv: 2006.11477 [cs.CL].

Brown, Steven R and Walter Weintraub (Mar. 1984). "Verbal Behavior: Adaptation and Psychopathology". In: *Political Psychology* 5, p. 107. DOI: 10.2307/3790837.

Cer, Daniel et al. (2018). *Universal Sentence Encoder*. arXiv: 1803.11175 [cs.CL].

Chaieb, Leila, Christian Hoppe, and Juergen Fell (2022). "Mind wandering and depression: A status report". In: *Neuroscience & Biobehavioral Reviews* 133, p. 104505. ISSN: 0149-7634. DOI: https://doi.org/10.1016/j.neubiorev.2021.12.028. URL: https://www.sciencedirect.com/science/article/pii/S0149763421005765.

Cohen, Alex S., Yunjung Kim, and Gina M. Najolia (2013). "Psychiatric symptom versus neurocognitive correlates of diminished expressivity in schizophrenia and mood disorders". In: *Schizophrenia Research* 146.1, pp. 249–253. ISSN: 0920-9964. DOI: https://doi.org/10.1016/j.schres.2013.02.002. URL: https://www.sciencedirect.com/science/article/pii/S0920996413000777.

Darby, John K., Nina Simmons, and Philip A. Berger (1984). "Speech and voice parameters of depression: A pilot study". In: *Journal of Communication Disorders* 17.2, pp. 75–85. ISSN: 0021-9924. DOI: https://doi.org/10.1016/0021-9924(84)90013-3. URL: https://www.sciencedirect.com/science/article/pii/0021992484900133.

DeVault, David et al. (Jan. 2014). "SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support". In: vol. 2, pp. 1061–1068. ISBN: 9781450327381.

Devlin, Jacob et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 [cs.CL].

Edwards, To'Meisha and Nicholas S. Holtzman (2017). "A meta-analysis of correlations between depression and first person singular pronoun use". In: *Journal of Research in Personality* 68, pp. 63–68. ISSN: 0092-6566. DOI: https://doi.org/10.1016/j.jrp.2017.02.005. URL: https://www.sciencedirect.com/science/article/pii/S0092656616302884.

Epstein, Ronald et al. (Sept. 2010). ""I Didn't Know What Was Wrong:" How People With Undiagnosed Depression Recognize, Name and Explain Their Distress". In: *Journal of general internal medicine* 25, pp. 954–61. DOI: 10.1007/s11606-010-1367-0.

Eyben, Florian et al. (Oct. 2013). "Recent developments in openSMILE, the Munich open-source multimedia feature extractor". In: pp. 835–838. DOI: 10.1145/2502081.2502224.

Falicov, Celia (Oct. 2003). "Culture, society and gender in depression". In: *Journal of Family Therapy* 25, pp. 371 –387. DOI: 10.1111/1467-6427.00256.

Fast, Ethan, Binbin Chen, and Michael Bernstein (Feb. 2016). "Empath: Understanding Topic Signals in Large-Scale Text". In: DOI: 10.1145/2858036.2858535.

Gowdy, J.N. and Z. Tufekci (2000). "Mel-scaled discrete wavelet coefficients for speech recognition". In: *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*. Vol. 3, 1351–1354 vol.3. DOI: 10.1109/ICASSP.2000.861829.

Gratch, Jonathan et al. (Feb. 2013). "User-State Sensing for Virtual Health Agents and TeleHealth Applications." In: *Studies in health technology and informatics* 184, pp. 151–7.

Gratch, Jonathan et al. (Jan. 2014a). "The Distress Analysis Interview Corpus of human and computer interviews." In.

Gratch, Jonathan et al. (May 2014b). "The Distress Analysis Interview Corpus of human and computer interviews". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 3123–3128. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf.

Guo, Yanrong et al. (2022). *A Topic-Attentive Transformer-based Model For Multimodal Depression Detection*. arXiv: 2206.13256 [cs.MM].

Hanai, Tuka, Mohammad Ghassemi, and James Glass (Sept. 2018). "Detecting Depression with Audio/Text Sequence Modeling of Interviews". In: pp. 1716–1720. DOI: 10.21437/Interspeech.2018-2522.

Hebbar, Rajat, Krishna Somandepalli, and Shrikanth Narayanan (2018). "Improving Gender Identification in Movie Audio Using Cross-Domain Data". In: *Proc. Interspeech 2018*, pp. 282–286. DOI: 10.21437/Interspeech.2018-1462.

Kroenke, Kurt et al. (2009). "The PHQ-8 as a measure of current depression in the general population". In: *Journal of Affective Disorders* 114.1, pp. 163–173. ISSN: 0165-0327. DOI: https://doi.org/10.1016/j.jad.2008.06.026. URL: https://www.sciencedirect.com/science/article/pii/S0165032708002826.

Levis, Brooke et al. (2018). "Probability of major depression diagnostic classification using semi-structured versus fully structured diagnostic interviews". In: *The British Journal of Psychiatry* 212.6, 377–385. DOI: 10.1192/bjp.2018.54.

Liu, Yinhan et al. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv: 1907.11692 [cs.CL].

Lord, J. R. (1921). "Manic-depressive Insanity and Paranoia. By Prof. Emil Kraepelin; translated by R. Mary Barclay, M.A., M.B.; edited by George M. Robertson, M.D., F.R.C.P.Edin. Edinburgh: E. &; S. Livingstone, 1921. Demy 8vo. Pp. 280. Forty-nine illustrations, eighteen in colour. Price 12s. 6d". In: *Journal of Mental Science* 67.278, 342–346. DOI: 10.1192/bjp.67.278.342.

Ma, Xingchen et al. (Oct. 2016). "DepAudioNet: An Efficient Deep Model for Audio based Depression Classification". In: pp. 35–42. DOI: 10.1145/2988257.2988267.

Mallol-Ragolta, Adria et al. (Sept. 2019). "A Hierarchical Attention Network-Based Approach for Depression Detection from Transcribed Clinical Interviews". In: pp. 221–225. DOI: 10.21437/Interspeech.2019-2036.

McFee, Brian et al. (2015). "librosa: Audio and music signal analysis in python". In: *Proceedings of the 14th python in science conference*. Vol. 8.

Mikolov, Tomas et al. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv: 1301.3781 [cs.CL].

Moore II, Elliot et al. (2008). "Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech". In: *IEEE Transactions on Biomedical Engineering* 55.1, pp. 96–107. DOI: 10.1109/TBME.2007.900562.

Ozdas, A. et al. (2000). "Analysis of fundamental frequency for near term suicidal risk assessment". In: *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics. 'cybernetics evolving to systems, humans, organizations, and their complex interactions' (cat. no.0*. Vol. 3, 1853–1858 vol.3. DOI: 10.1109/ICSMC.2000.886379.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pennebaker, James, Martha Francis, and Roger Booth (Jan. 1999). "Linguistic inquiry and word count (LIWC)". In.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (Jan. 2014). "Glove: Global Vectors for Word Representation". In: vol. 14, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

Radford, Alec et al. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*. arXiv: 2212.04356 [eess.AS].

Reed, Lawrence, Michael Sayette, and Jeffrey Cohn (Nov. 2007). "Impact of depression on response to comedy: A dynamic facial coding analysis". In: *Journal of abnormal psychology* 116, pp. 804–9. DOI: 10.1037/0021-843X.116.4.804.

Rejaibi, Emna et al. (2022). "MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech". In: *Biomedical Signal Processing and Control* 71, p. 103107. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2021.103107. URL: https://www.sciencedirect.com/science/article/pii/S1746809421007047.

Reynolds, Douglas et al. (May 2003). "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition". In: vol. 4, pp. IV –784. DOI: 10.1109/ICASSP.2003.1202760.

Sardari, Sara et al. (2022). "Audio based depression detection using Convolutional Autoencoder". In: *Expert Systems with Applications* 189, p. 116076. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.116076. URL: https://www.sciencedirect.com/science/article/pii/S0957417421014147.

Shrivastav, Rahul, David Eddins, and Supraja Anand (Mar. 2012). "Pitch strength of normal and dysphonic voices". In: *The Journal of the Acoustical Society of America* 131, pp. 2261–9. DOI: 10.1121/1.3681937.

Stasak, Brian et al. (Sept. 2016). "An Investigation of Emotional Speech in Depression Classification". In: pp. 485–489. DOI: 10.21437/Interspeech.2016-867.

Toto, Ermal, ML Tlachac, and Elke Rundensteiner (Oct. 2021). "AudiBERT: A Deep Transfer Learning Multimodal Classification Framework for Depression Screening". In: pp. 4145–4154. DOI: 10.1145/3459637.3481895.

Trevino, Andrea, Thomas F. Quatieri, and Nicolas Malyska (2011). "Phonologically-based biomarkers for major depressive disorder". In: *EURASIP Journal on Advances in Signal Processing* 2011, pp. 1–18.

Trotzek, Marcel, Sven Koitka, and Christoph M. Friedrich (2020). "Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences". In: *IEEE Transactions on Knowledge and Data Engineering* 32.3, pp. 588–601. DOI: 10.1109/TKDE.2018.2885515.

Vaswani, Ashish et al. (2017). *Attention Is All You Need*. arXiv: 1706.03762 [cs.CL].

Wang, Jingying et al. (Oct. 2019). "Acoustic differences between healthy and depressed people: a cross-situation study". In: *BMC Psychiatry* 19. DOI: 10.1186/s12888-019-2300-7.

WHO (2017). *Depression and other common mental disorders: global health estimates*. Technical documents, 24 p.

Wolf, Thomas et al. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. arXiv: 1910.03771 [cs.CL].

Wu, Yonghui et al. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. arXiv: 1609.08144 [cs.CL].

Zhu, Yukun et al. (Dec. 2015). "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books". In: *The IEEE International Conference on Computer Vision (ICCV)*.