Using Artificial Intelligence to Aid Depression Detection

by

Deângela C. G. Neves

A thesis submitted in partial fulfillment for the degree of Bachelor of Engineering

in the
Technology Center
Department of Computer Engineering and Automation

9 de Julho de 2019
“There’s a snake in my boot!”

Sheriff Woody
Depression is a mental illness that affects a person’s mood, thinking, and behavior. Besides personal distress, depression is also considered a matter of public health. Recent research shows the advantages of using machine learning algorithms to automate and improve the screening for depression. In this thesis, we address the Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) database, comprising clinical interviews and questionnaire assessments of over a hundred individuals. To automate the screening, we investigate a deep learning multimodal model, combining audio, visual, and text features. These features are extracted using VGGish model, OpenFace, and Doc2Vec, respectively, and fed into a multilayer perceptron (MLP) network to classify individuals as depressed or non-depressed. We compare the proposed approach to similar existing approaches from the literature through standard binary classification metrics.

Keywords: Deep Learning, Depression, DAIC-WOZ
A depressão é um transtorno psicológico que causa alterações comportamentais e no humor de uma pessoa, e é considerada um problema de saúde pública. Estudos recentes mostram as vantagens de se utilizar algoritmos de machine learning não só para automatizar, mas também melhorar o processo de triagem para depressão. Neste trabalho, é feita uma análise de dados exploratória no dataset Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ), que consiste em um conjunto de entrevistas clínicas e questionários de mais de 100 indivíduos. Além disso, utiliza-se a mesma base de dados para desenvolver um modelo multimodal de Aprendizagem Profunda, que combina dados de áudio, vídeo e texto para classificar o resultado de triagens para depressão. Os dados utilizados no modelo são extraídos a partir de uma rede pretreinada VGGish, do OpenFace e Doc2Vec, respectivamente. Em seguida, esses dados são utilizados para alimentar uma rede neural perceptron de múltiplas camadas, que os classifica como depressivos ou não-depressivos na triagem. Por fim, o modelo proposto é comparado a outras abordagens existentes na literatura a partir de métricas padrão de classificação binária.

Palavras-chave: Aprendizado Profundo, Depressão, DAIC-WOZ
Acknowledgements

I owe a deep sense of gratitude to my advisor Leonardo Bezerra for his dedication and keen interest above all his overwhelming attitude to help his students, solely and mainly responsible for completing my work.

I am extremely thankful to Hugo Dzin and Maria Augusta for providing me necessary technical suggestions during my research pursuit.
# Contenido

Abstract ii

Resumo iii

Acknowledgements iv

List of Figures vii

List of Tables viii

1 Introduction 1

2 Background and related work 3
   2.1 Theoretical background . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
      2.1.1 Deep learning . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
      2.1.2 Speech processing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5
      2.1.3 Natural language processing (NLP) . . . . . . . . . . . . . . . . . . . 7
      2.1.4 Computer vision (CV) . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
      2.2 Related work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

3 Exploratory Data Analysis 12
   3.1 PHQ-8 Questionnaires . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13
   3.2 Interview Transcripts . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14
   3.3 Audio Recordings . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16
   3.4 Visual Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

4 Proposed Approach 19
   4.1 Network Architecture . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19
   4.2 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20

5 Conclusions and future work 22
## Lista de Figuras

<table>
<thead>
<tr>
<th>Figura</th>
<th>Descripción</th>
<th>Página</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Simplest MLP architecture, containing only one hidden layer. In this example, the network takes three input features and outputs a single value.</td>
<td>4</td>
</tr>
<tr>
<td>2.2</td>
<td>Example of max pooling filter over a matrix of input features.</td>
<td>5</td>
</tr>
<tr>
<td>2.3</td>
<td>Step-wise summary of MFCC extraction</td>
<td>6</td>
</tr>
<tr>
<td>2.4</td>
<td>Step-wise summary of Log-Mel feature extraction</td>
<td>7</td>
</tr>
<tr>
<td>2.5</td>
<td>Word embeddings.</td>
<td>8</td>
</tr>
<tr>
<td>2.6</td>
<td>Document embeddings.</td>
<td>9</td>
</tr>
<tr>
<td>2.7</td>
<td>CLNF embeddings.</td>
<td>10</td>
</tr>
<tr>
<td>3.1</td>
<td>Gender distribution of depressed participants.</td>
<td>12</td>
</tr>
<tr>
<td>3.2</td>
<td>Gender distribution of non-depressed participants.</td>
<td>13</td>
</tr>
<tr>
<td>3.3</td>
<td>Distribution of questionnaire answers.</td>
<td>13</td>
</tr>
<tr>
<td>3.4</td>
<td>Sentiments detected in the speech of depressed subjects</td>
<td>14</td>
</tr>
<tr>
<td>3.5</td>
<td>Sentiments detected in the speech of non-depressed subjects</td>
<td>14</td>
</tr>
<tr>
<td>3.6</td>
<td>Depressed Syuzhet Plot</td>
<td>15</td>
</tr>
<tr>
<td>3.7</td>
<td>Non-Depressed Syuzhet Plot</td>
<td>16</td>
</tr>
<tr>
<td>3.8</td>
<td>Face landmarks [1].</td>
<td>17</td>
</tr>
<tr>
<td>3.9</td>
<td>Average eye blink rate (EBR) from depressed and non-depressed participants.</td>
<td>17</td>
</tr>
<tr>
<td>4.1</td>
<td>Model architecture</td>
<td>20</td>
</tr>
</tbody>
</table>
Lista de Tabelas

4.1 Depression Detection .................................................. 20
4.2 Results obtained with multimodality and single modality ............ 21
To my dearest uncle and father, Djanilson,
and my beloved great grandmother, Bia,
both of whom have always supported
and inspired me along the way. . . .
Capítulo 1

Introduction

Depression is one of the most common mental disorders worldwide, affecting more than 300 million people [2]. Depressed individuals experience persistent sadness, low mood, and loss of energy, along with other symptoms. Due to these symptoms and the associated social stigma, individuals suffering from depression might be unwilling to seek professional help. Thus, passive options of diagnosis may address this problem and help provide a better screening for depression.

The standard procedure used in depression screening and diagnosis combines questionnaires and clinical interviews to assess the severity of symptoms. The main purpose of the questionnaires is to assist clinicians in quantifying depression symptoms and monitoring symptom changes over time. There are several questionnaires used throughout the world, such as the Hamilton anxiety depression scale (HADS), the Beck depression inventory-II (BDI-II), and the patient health questionnaires (PHQ) [3]. However, it is during the clinical interview that clinicians may examine the mental state and motivation of the patient by history-taking, and establish rapport with them [4]. Interviewing involves simultaneously interpreting visual and audio information provided by the patient, as well as the semantic content of their narrative.

In order to assist the clinical diagnosis of depression, machine learning algorithms have been increasingly used [5–9]. The underlying machine learning problem takes the interview data as input and can be modeled either as a binary classification or as a regression task. In the latter case, the goal is to predict questionnaire scores of individuals. Scores are then compared to known questionnaire thresholds to determine whether a given individual is depressed. Existing approaches vary as to how much of the interview data is used by estimators, with only a handful simultaneously dealing with computer vision, audio, and natural language processing. These multimodal approaches are generally based on deep learning algorithms and have shown promising results [5, 7].
This work investigates a deep learning multimodal estimator, combining audio, visual, and text features. It addresses the Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) database [10], comprising clinical interviews and questionnaire assessments of over a hundred individuals. Interviews comprise audio recordings and textual transcripts, as well as 3D facial scans to represent facial expressions without compromising individual privacy. Features from each domain are extracted using pre-trained deep learning models, namely VGGish for audio processing [11], OpenFace for computer vision [12], and Doc2Vec for natural language processing [13]. Extracted features are fed into a multilayer perceptron (MLP) network to classify individuals as depressed or non-depressed, and we compare our work to the approaches identified in the literature. Source code for this work is available online at https://github.com/deangelacgn/smart-therapist.

The remainder of this thesis is organized as follows. Chapter 2 briefly reviews background concepts required to understand this thesis, and details related work on depression detection. Chapter 3 describes the dataset addressed in this work and presents an exploratory analysis of its characteristics. Chapter 4 introduces and evaluates the multimodal model architecture adopted in this work, comparing it to existing similar approaches from the literature. Finally, Chapter 5 presents our conclusions and discusses future work possibilities.
Depression detection aid through machine learning is a topic for which a number of deep learning approaches can be identified in the literature. In the context of the DAICWOZ dataset considered in this work, estimators can use as input data either questionnaire answers and results in the form of tabular data, or screening interviews unstructured data, for which video, audio, and transcripts are provided. In this chapter, we initially go over background concepts related to the domains of computer vision, and speech and natural language processing. Our focus is on the techniques that relate to this work or to the existing literature on depression detection. Later, the literature on depression detection is reviewed, with a focus on approaches that are based on deep learning.

2.1 Theoretical background

As previously discussed, to tackle the problem of detecting depression through screening interviews, three different research fields must be explored: (i) speech processing; (ii) natural language processing, and; (iii) computer vision. In common, the most effective techniques currently employed in those domains are based on deep learning approaches. This section briefly revisits the theoretical background related to deep learning and techniques from these three domains that have either been adopted in this work or that relate to existing works we later review.

2.1.1 Deep learning

Deep learning approaches are a refinement of the multilayer perceptron algorithm, where very large networks are considered through specialized architectures. Among the most relevant architectures are convolutional neural networks (CNNs, [14]), long-short term
memory (LSTM) networks [15], and generative adversarial networks (GANs, [16]). Next, we briefly detail MLP and CNN, the most relevant architectures for the context of this thesis.

**Multilayer perceptron (MLP)**

A multilayer perceptron (MLP) neural network consists on a number of artificial neurons grouped in layers such that each neuron in a given layer is connected to all others in the next layer. Neurons represent what the network learns throughout training such that larger weights are set to more relevant information. MLPs may have one or more hidden layers, and the simplest MLP architecture is illustrated in Figure 2.1.

![Simplest MLP architecture](image)

**Figure 2.1:** Simplest MLP architecture, containing only one hidden layer. In this example, the network takes three input features and outputs a single value.

**Convolutional neural networks (CNNs)**

Convolutional neural networks (CNNs) are specialized neural networks originally proposed for the context of computer vision, but that have recently been successfully applied to many other domains [17]. In a convolutional network, different kinds of layers are stacked to obtain the desired behavior. Next, we briefly detail the most common kinds of layers used in convolutional networks.

**Convolutional layers** perform convolutions through filters, effectively performing a feature extraction of the input. As more convolutional layers are stacked on the network, the number of parameters that need to be trained may become to large and the associated computational cost unfeasible.
Pooling layers are generally used after convolutional layers to further reduce the number of parameters of the network. A max pooling layer [18], for instance, keeps only the maximum value parameter in a given filter radius. This is illustrated in Figure 2.2.

Dropout layers are employed to reduce the possibility of overfitting by randomly discarding neurons from the network.

2.1.2 Speech processing

Any sound generated by humans is determined by the shape of their vocal tract. If this shape can be determined correctly, any sound produced can be accurately represented. However, speech signals inherently have nonlinear characteristics, since a given segment of a speech can be generated by an infinite number of vocal tract configurations.

Speech processing involves the analysis and synthesis of speech sounds. Conventional methods of speech processing are generally classified into time and frequency domain. The first domain mostly covers a simpler set of techniques, such as signal energy measurement, level crossing, auto correlation. However, these techniques do not provide enough information when the focus of the analysis is the linguistic content of the speech and how it is being carried by human voice. Such features are better captured by frequency-domain techniques. Besides analytical approaches, CNNs have shown prominent results in audio classification tasks, and have also been successfully applied for feature extraction in the audio domain [11].

This work explores the usage of two analytical frequency-domain techniques to extract audio speech features: Mel frequency cepstral coefficient (MFCC) and Log-Mel features. Moreover, a VGGish pre-trained CNN is also applied for extracting audio features.
Mel frequency cepstral coefficient (MFCC)

MFCCs accurately represent the envelope of the short time power spectrum, which is where the vocal tract manifests itself. In order to precisely grasp what these features consist of, it is essential to understand how they are computed. To achieve that, it is necessary to be familiar with the concepts of Mel frequency and cepstral.

The Mel scale is a scale that relates the actual measured frequency to the perceived frequency of a tone. It scales the frequency in order to match more closely what the human ear can hear, since humans better identify small changes in speech at lower frequencies. Therefore, the Mel frequency is the corresponding value of a given frequency when converted to the Mel scale.

\[
Mel(f) = 1195 \ln \left(1 + \frac{f}{700}\right)
\]  

(2.1)

Cepstral is a tricky word adopted by Bogert et al. [19] to convey the information contained in a spectrum that is neither in the time domain nor in the frequency domain. The word itself is an anagram of spectral with the spec reversed. As shown in Figure 2.3, when computing MFCCs the speech signal is converted to the frequency domain after the application of the Fast Fourier Transform. In the following steps after such conversion, the Discrete Cosine Transform is taken. Since that transform is applied on the frequency spectrum itself, the resulting spectrum is in an unnamed domain and Bogert et al. [19] decided to call it the quefrency domain. Thus, that spectrum of quefrequencies is actually called cepstrum.

![Figure 2.3: Step-wise summary of MFCC extraction](image)

Log-Mel

Similar to MFCCs, Log-Mel features are also related to the concepts of Mel scale and vocal tract shape. Their extraction procedures are nearly the same, as shown in Figure 2.4. However, instead of computing the Discrete Cosine Transform as the final
step, the output of Mel-Filterbank energy logs itself is used as feature.

**VGGish CNN**

The VGGish convolutional neural network [11] is a pretrained model released by Google for audio event detection and generating audio features. The model was trained with AudioSet, a large-scale database containing more than 2 million YouTube video soundtracks [20].

**2.1.3 Natural language processing (NLP)**

Human language varies from country to country, and carries a variety of subjective aspects, e.g., nuances, idioms, slangs, etc. Additionally, languages present a dynamic nature, constantly changing over the course of time. **Natural language processing (NLP)** is a research field focused on processing and analyzing the interactions between human language and computers. It has a variety of applications, such as machine translation, text classification, and text summarization.

Universal embeddings have become an important tool for machine learning algorithms that tackle NLP problems [13]. Next, we review two of the most commonly adopted embeddings for NLP, namely Word2Vec and Doc2Vec.

**Word2Vec**

Word2Vec is a family of models used to generate word embeddings, that is, convert each word from a set of sentences into a feature vector. To obtain word embeddings of size N, a neural network with a single hidden layer of size N is trained, and the weight matrix that maps the input layer to the hidden layer of the network is used as the word embeddings. This network can be designed in two possible ways:
Figura 2.5: Word embeddings.

- Continuous skip-gram: given a word, predict the words likely to appear around it (the context). This is illustrated in Figure 2.5.

- Continuous bag-of-words (CBOW): given a set of words, predict the word that will most likely appear in that context.

Each method has better results in different situations: while CBOW is not as good as skip-gram at representing rare words, its embeddings have better accuracy for frequent words. CBOW is also several times faster to train than the skip-gram method. On the other hand, skip-gram works well even with a small amount of data, and produces better embeddings than CBOW for rare words.

The main benefit of word embeddings is that words that appear in similar contexts tend to be close together in the vector space. Moreover, simple arithmetic operations can be used to understand the relation between words. For example, $\text{WordVector("king") - WordVector("man")] = \text{WordVector("queen") - WordVector("woman")}$.

**Doc2Vec**

Doc2Vec works similarly to Word2Vec, but instead of generating a feature vector for each word, one vector is generated for each document (or sentence, or paragraph.) Like word2vec, the neural network that generates the document embeddings can be designed in two ways, analogous to CBOW and skip-gram:
- Distributed bag-of-words (PV-DBOW): given a document, predict the words likely to appear in it;
- Distributed memory (PV-DM): given a document and a set of words, predict the word that most likely will appear near those words, within that document.

The distributed bag-of-words approach is illustrated in Figure 2.6.

2.1.4 Computer vision (CV)

*Computer vision* (CV) is a broad research field that seeks to reproduce human perception and accurate understanding of visual information in computers. Common CV applications include face recognition, image captioning, and autonomous vehicles. Deep learning techniques have achieved remarkable results in most of these CV tasks, significantly outperforming state-of-art results from other approaches [17]. In general, this has been accomplished through *visual embeddings*, a feature extraction approach that reduces input images to their aspects that are more relevant to task one needs to address.

For instance, the DAICWOZ dataset provides video for the screening interviews under the *conditional local neural fields* (CLNF) representation [1], illustrated in Figure 2.7. These embeddings have been computed using the OpenFace toolkit [12], developed by researchers of the University of Cambridge, capable of facial landmark detection, facial action unit recognition, and head pose and eye-gaze estimation.

The main characteristics of CLNF embeddings are that (i) they have been devised for the purpose of facial landmark detection and tracking, and hence can be adopted for facial expression analysis, and; (ii) they conceal the identity of the individuals. Together,
Figura 2.7: CLNF embeddings.

these characteristics motivate the adoption of CLNF by the curators of DAICWOZ. As a drawback, the unavailability of raw images means that alternative visual embeddings cannot be considered in the context of this work.

2.2 Related work

As previously discussed, recent research in the domain of depression detection has shown prominent results with deep neural networks (DNNs). These networks are able to extract high-level features from raw data after being trained over a large amount of data, and generally outperform the results obtained by traditional handcrafted features when fed into machine learning models.

In general, depression detection approaches based on deep neural networks have targeted the DAIC-WOZ dataset, same as this thesis. However, many tackle only a pair of fields of applications among computer vision, speech, and natural language processing:

Al Hanai et al. [9] exploit the potential of long-short term memory (LSTM) networks to model the communication process during clinical interview. The proposed automated model learns audio and text features from sequences of questions and answers, making explicit content modeling techniques unnecessary.

Yang et al. [5] use deep convolutional neural networks (DCNN) to extract audio and visual features from the clinical interviews. The extracted features are then fed into another multilayer perceptron network (MLP) to predict the PHQ-8 score.

Tzirakis et al. [6] propose a deep residual network (ResNet) for extracting visual features, while extracting audio features with a DCNN. In order to track the context of the information being processed, both sets of features are then concatenated and fed into an LSTM.

A handful of works attempt to grasp more faithfully the clinician-patient communication interactions by combining video, audio, and text features.
Williamson et al. [7] compute analytical features for speech recognition and process transcripts using word embeddings and feature extraction. A Gaussian staircase model is built from these features alongside the visual features provided by the dataset.

Haque et al. [8] follow a similar approach to the work of Williamson et al. [7], but compute features on a sentence level. Processed features are fed to a temporal convolutional network [21], which is shown to outperform an LSTM. Both binary classification and regression tasks are investigated, but no information is provided on the classifier and regressor adopted.

The approach proposed in this thesis is similar in essence to the works described above, to which it is compared. In the next section, we discuss the DAIC-WOZ dataset and conduct an exploratory data analysis stage to understand its main characteristics.
The Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) dataset [10] was used to conduct the experiments. It contains 189 clinical interviews and PHQ-8 questionnaire answers of depressed and non-depressed patients. The interviews were conducted by a computer-animated avatar called Ellie, which was remotely controlled by a clinician. For each patient, the audio recording and transcripts of the interview are provided, along with 3D facial scans extracted from the video recordings.

The dataset is split into training, development and test set. The first one contains 107 samples, 29 out of them being of depressed participants. The development set is composed of 35 samples, 12 out of them being of depressed subjects. Finally, the test set contains 47 samples, nine out of them being of depressed participants. The gender distribution of samples is well-balanced, as shown in Figure ?? and ??, preventing gender bias in deep learning algorithms. However, the number of non-depressed subjects is about four times larger than that of depressed ones, which can introduce bias towards non-depressed classification.

![Gender distribution of depressed participants.](chart.png)
3.1 PHQ-8 Questionnaires

The Eight-Item Personal Health Questionnaire Depression Scale (PHQ-8) is in the spectrum of diagnostic tools used for measuring the severity of depressive disorders in large clinical studies\cite{22}. As its name suggests, it is composed of eight multiple choice questions regarding patient’s level of distress, eating habits, and concentration. Answers are divided into four categories scoring from 0 to 3, respectively: *not at all*; *several days*; *more than half days*, and; *nearly every day*. The higher the value, the more likely the patient has the condition, and participants who achieve a PHQ-8 score equal to or greater than 10 are diagnosed as depressed.

The distribution of all questionnaire answers is shown in Figure 3.3. Answers provide a broad scope of emotional and behavioural changes in the routine of patients. In the DAIC-WOZ dataset, interview participants have experienced more changes in sleep and tiredness. However, the variability in the rates of the eight items listed among four different categories is subtle, making it difficult to retrieve more in-depth information of the state of patients’ condition. In this thesis, the final PHQ-8 score is used to identify when an individual is depressed or not.
3.2 Interview Transcripts

The interview transcripts are organized in CSV files, one per participant. Each file has four columns that indicate, respectively:

1. The starting time of each speech;
2. The ending time of each speech;
3. Who’s the speaker: Ellie or the interviewee;
4. The content of each speech.

In order to better understand what kind of features describe and distinguish interviews of depressed participants from non-depressed ones, a lexicon-based sentiment analysis was performed. The NRC Emotion Lexicon [23] was adopted for the analysis and consists of a list of 14,182 English words, which are associated with the emotion more likely to be carried by them. Eight basic emotions were manually annotated across the words: disgust, anger, trust, fear, anticipation, surprise, sadness, and joy.

The sentiments extracted from the speech of depressed subjects, when compared to ones from non-depressed participants are almost identical, as shown in Figures 3.4 and 3.5. One might expect that the rate positive words used by non-depressed subjects, when compared to the depressed participants would be far superior. Based on those results, it is possible to notice that a lexicon-base based approach is not enough to spot the signs of depression conveyed through language.
To perform further investigation, the Syuzhet package [24] was used to analyze how the subjects’ narratives are organized according to the sentiments detected in each part of their speech using a variety of sentiment dictionaries. Therefore, Syuzhet and Macro Shape plot were obtained from both depressed and non-depressed participants’ speech. Both graphs evaluate the scales of sentiments across a narrative. Sentiments range between -1 and 1, the closer to one the more positive it is. The narrative time is measured in number of sentences. An example of such graphs is shown in Figure 3.6 and 3.7. They were extracted from a depressed and non-depressed participant in the database, respectively.

The Syuzhet plot given in those figures is composed of three line graphs, each of them representing a smoothing method applied on the data. The red line indicates the Discrete Cosine Transformation (DCT) of the narrative, which is used to extract sentiment and sentiment-derived plot arcs from text. Therefore, it is possible to evaluate the overall emotional trends through narrative arcs. The gray dashed line shows the results of applying a moving-average filter on the narrative in order to smooth the short-term fluctuations of sentiments detected in the text sentences, highlighting the extended emotional sense of the stories being told by each participant. Finally, the blue dashed line graph consists of another smooth pattern of the sentiments created using Loess regression.

The simplified macro shape plot, as its name suggests, is a flatten representation of how the three curves mentioned above represent the data. It is computed with a DCT that
Figura 3.7: Non-Depressed Syuzhet Plot

retains fewer low frequency components, minimizing the noise. Also, it uses the reverse transform process to normalize the x-axis to 100 units.

When comparing graphs in Figure 3.6 and Figure 3.7, it is possible to note the prevalence of arcs associated with negative sentiments in the narrative of the depressed participant, while in the narrative of the other participant positive and negative arcs are more balanced.

3.3 Audio Recordings

The raw audio recordings are provided in WAV format, one per participant. They last approximately 16 minutes on average, and mostly contain the speech of the interviewee only. Their sampling rate is 16 kHz.

3.4 Visual Data

The raw interview video recordings are not provided as a matter of privacy. Instead, a set of face features extracted by timestamp with OpenFace toolkit [12] are provided:

- 3D facial landmarks;
- Eye gaze estimation;
• Head pose tracking;

• Facial action unit (AU) features;

From this set of visual features, this section mainly addresses facial landmarks (Figure 3.8) and eye gaze estimations. Some works in health sciences propose the existence of a correlation relationship between eye blink rate (EBR) and depressive symptoms, since EBR has been investigated in many studies as a marker of dopamine levels [25]. Thus, the eye gaze features can be used to examine the presence of such differences in EBR between depressed and non-depressive subjects.

Figure 3.9 depicts boxplots of EBR values computed for non-depressed and depressed participants. The differences in distributions are, however, very subtle, likely not enough for models to differentiate between participant groups.

Figure 3.8: Face landmarks [1].

Figure 3.9: Average eye blink rate (EBR) from depressed and non-depressed participants.
In this chapter, we have conducted an exploratory analysis on the DAIC-WOZ dataset to better understand its characteristics. In the next chapter, we proposed and evaluate a multi-modelo deep learning approach to address the depression detection problem.
Proposed Approach

To address the problem of detecting depression in screening interviews, a multimodal model using audio, visual and text features is proposed. The architecture of the model is shown in Figure 4.1.

Firstly, the three sets of data are fed into deep learning models for feature extraction. Then, the feature embeddings generated by each of them are concatenated into a single embedding. Finally, this embedding is passed as input to a multilayer perceptron network (MLP) to be classified.

4.1 Network Architecture

The Face Feature Extractor represents the OpenFace model for extracting face landmarks. The extracted features consists of 68 feature vectors that represent face landmarks. Since these landmarks are extracted by timestamp throughout the interview, an average facial scan is computed for each participant, generating a unique vector of 68 features. This average face feature representation is computed by taking the element-wise mean of all feature vectors of a given interview.

The Speech Feature Extractor consists of a pretrained VGGish model that extracts 128-dimensional embeddings of every 10ms of the audio recording. Using a similar approach adopted for post-processing face features, the sequence of audio embeddings generated for each audio recording is average mixed into a single array of size 128 by the computation of element-wise mean of elements.

The Text Feature Extractor used was Doc2vec, in the DBOW mode, with input text data being only the content of the answers of participants to interview questions. Each document embedding contains 300 features.
Finally, a MLP is used as an standard classifier to give the screening result for depression. The network is composed of 4 hidden layers with 70 neurons each. Its input consists of a 496-dimensional embedding and the output layer consists of a single neuron that indicates the binary screening result, with 1 being positive for depression.

![Model Architecture](image)

**Figure 4.1:** Model architecture

### 4.2 Results

Since it is a binary classification model, the F1 score is used for evaluation as it better deals with class imbalances. The results obtained with the multimodal network architecture were compared to the F1 scores of other multimodal architectures from the literature, as shown in Table 4.1.

<table>
<thead>
<tr>
<th>Method</th>
<th>F1 Score</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gong et al.</td>
<td>70.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Williamson et al.</td>
<td>81.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Haque et al.</td>
<td>76.9</td>
<td>71.4</td>
<td>83.3</td>
</tr>
<tr>
<td>MLP</td>
<td>47.41</td>
<td>44.79</td>
<td>51.42</td>
</tr>
</tbody>
</table>

At a first glance, there is a noticeable difference between the results of the approach taken in this thesis and the results from current literature. This divergence may be possibly caused by the level of granularity adopted for processing the data. Some works adopt a more fine-grained, sentence-level strategy, where one feature vector is generated for each sentence, meanwhile this thesis uses a document-level strategy.

Since each interview was treated as a single document, the same procedure had to be done for audio and visual features, reducing each of them to a single embedding per document. This approach cannot extract the same amount of information that more fine-grained approach is able to and that may explain the discrepancy in accuracy when compared to other techniques.
The model was also compared to unimodal strategies, in each of the three fields: textual, visual and audio features. The results obtained with a single set of features feeding the MLP are shown in Table 4.2.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>F1 Score</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>52.11</td>
<td>43.18</td>
<td>65.71</td>
</tr>
<tr>
<td>Log-Mel</td>
<td>54.75</td>
<td>54.55</td>
<td>62.85</td>
</tr>
<tr>
<td>VGGish</td>
<td>66.67</td>
<td>60.71</td>
<td>73.91</td>
</tr>
<tr>
<td>3D Face</td>
<td>11.76</td>
<td>20</td>
<td>8.3</td>
</tr>
<tr>
<td>Doc2Vec</td>
<td>37.5</td>
<td>75</td>
<td>25</td>
</tr>
<tr>
<td>Multimodal features</td>
<td>47.41</td>
<td>44.79</td>
<td>51.42</td>
</tr>
</tbody>
</table>

For audio features, no mixing or condensing was required since all extracted features were flattened, as to more accurately preserve the whole content of the interview. For facial features, the model was trained using the mean facial expression (each document was associated with a single mean face). To obtain embeddings for textual features, a doc2vec model was trained using as input a document containing all of the interviewee’s sentences, without the interviewer’s questions.
Conclusions and future work

Even though the results of this thesis did not outperform state-of-the-art research on depression detection, the techniques presented are relatively simple compared to the complex models used to achieve the best accuracy in the current literature. Moreover, this thesis also presents an approach that is not well explored in previous works, and that is using the pre-trained VGGish network for speech embedding for the use case of detecting depression.

One weakness of the approach proposed in this thesis is that, for each interview, the three features (transcript, audio and facial landmarks) are all separately converted into a single embedding each, that are then used as input for the classification model. Because of this, it’s not possible for the model to learn the correlation between features for separate moments of the interview, but instead only the correlation between features for the interview as a whole. For instance, if the interviewee has a happy facial expression for most of the interview, but becomes sad at a specific moment, the embeddings for the facial landmark will only present the overall facial expression during the interview, and the model won’t have information about which moment of the speech and transcript triggered this change.

On future works, a better approach may be splitting each interview in multiple chunks, generating embeddings for each chunk individually, and then using those embeddings to train a temporal-convolutional network for classification. This way, more detailed information may be extracted from each interview, and the temporal relationship between the three kinds of features may be better presented to the model, hopefully increasing the overall accuracy of detecting depression.
Bibliografia


