JOÃO ANTÓNIO
LOPES RIBEIRO

QUANTIFICAÇÃO MULTIMODAL DA DEPRESSÃO USANDO APRENDIZAGEM AUTOMÁTICA

MULTIMODAL QUANTIFICATION OF DEPRESSION USING MACHINE LEARNING

Universidade de Aveiro
2021

JOÃO ANTÓNIO
LOPES RIBEIRO

QUANTIFICAÇÃO MULTIMODAL DA DEPRESSÃO USANDO APRENDIZAGEM AUTOMÁTICA

MULTIMODAL QUANTIFICATION OF DEPRESSION USING MACHINE LEARNING

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Informática, realizada sob a orientação científica da Doutora Alina Trifan, Investigadora do Instituto de Engenharia Eletrónica e Telemática de Aveiro e do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro, e do Professor Doutor José Luís Oliveira, Professor Catedrático do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro.

o júri / the jury

presidente / president

Prof. Doutora Pétia Georgieva

Professora Associada do Departamento de Eletrónica Telecomunicações e Informática da Universidade de Aveiro

vogais / examiners committee

Prof. Doutor Rui Lopes

Professor Coordenador do Instituto Politécnico de Bragança

Dra. Alina Trifan

Investigadora do Instituto de Engenharia Eletrónica e Telemática de Aveiro e do Departamento de Eletrónica Telecomunicações e Informática da Universidade de Aveiro

agradecimentos /
acknowledgements

I would like to thank my supervisors for all the help throughout the year. I would also like to thank my parents and my close family for the unconditional support during my five academic years. Also, a special mention to Miguel Alves, for giving me early references on psychological studies and books about depression.

Palavras Chave

Previsão de Depressão, Processamento de Linguagem Natural, Processamento de Voz, Aprendizagem Automática, Multimodal.

Resumo

A depressão é uma doença mental que cada vez mais se está a tornar comum na vida das pessoas e que pode ter implicações muito sérias no ser humano. Mais de 264 milhões de pessoas em todo o mundo sofrem com esta doença e a tendência é para estes números aumentarem ao longo dos anos. Tendo isto em conta, é necessário desenvolver métodos de reconhecimento e de previsão da depressão através da análise de linguagem natural, de comportamentos não verbais e de processamento de voz. Esta é uma área de estudo com elevado interesse, pois tanto pode servir os clínicos no apoio ao diagnóstico e ao tratamento de pacientes com esta doença mental, como também pode servir os pacientes ao receberem um diagnóstico robusto e um guia de tratamento adequado para conseguirem superar a doença. Esta dissertação foca-se no desenvolvimento de um método de quantificação e previsão da depressão numa pessoa, baseando-se em estudos e artigos na área, publicados em conferências internacionais. Com acesso a entrevistas e a dados clínicos, foi realizada uma análise da linguagem e uma análise da voz de cada participante, com o intuito de extrair características específicas que pudessem auxiliar a identificação de depressão. Após esta extração, foram desenvolvidas experiências com modelos unimodais e multimodais com o objetivo de conseguir quantificar corretamente a depressão de cada participante. Estes modelos ultrapassaram a base de referência da conferência AVEC, com resultados comparáveis a outros modelos publicados nesta mesma conferência.

Keywords

Depression Prediction, Natural Language Processing, Speech Processing, Multimodal, Machine Learning.

Abstract

Depression is a mental disorder that is increasingly becoming common in people's lives and that can have serious implications on human beings. Over 264 million people worldwide suffer from this disorder, and the trend is for these numbers to increase over the years. With this in mind, it is necessary to develop depression recognition and prediction methods by analysing natural language, non-verbal behaviours and speech processing. This is an area of study with high interest, since it can support clinicians on patients diagnosis and treatments, as well as it can also serve the patients by receiving a robust diagnosis and an adequate treatment guide so they can overcome the disorder. This dissertation focuses on developing a method that can quantify and predict if a person suffers from depression, basing itself on studies and articles in the area, published in international conferences. With access to interviews and clinical data, a language and speech analysis for each participant was performed, with the intent of extracting key characteristics that could assist depression identification. After the extraction, experiments with unimodal and multimodal models were developed with the objective of quantifying depression correctly for each participant. These models outperformed the AVEC conference baseline and presented comparable results with other published models in that same conference.

Contents

List of Figures

List of Tables

Glossary

AI	Artificial Intelligence
AU	Action Units
LP	Linear Prediction
RF	Random Forests
APA	American Psychiatric Association
BDI	Beck Depression Inventory
CNN	Convolutional Neural Network
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
GAN	Generative Adversarial Networks
HDR	Histogram of Displacement Range
HOG	Histogram of Oriented Gradient
LDA	Latent Dirichlet Allocation
LLD	Low Level Descriptors
MAE	Mean Absolute Error
MDD	Major Depressive Disorder
MLP	Multi-Layer Perceptron
NAQ	Normalized Amplitude Quotient
NRC	National Research Council Emotion Lexicon
PCA	Principal Component Analysis
PDM	Point Distribution Model
QOQ	Quasi-Open Quotient
Rbf	Radial basis function
RNN	Recurrent Neural Network
SVM	Support Vector Machine
SVR	Support Vector Regression
USC	University of South California
USE	Universal Sentence Encoder
XGB	eXtreme Gradient Boosting
ANEW	Affective Norms for English Words
AVEC	Audio/Visual Emotion Challenge
BERT	Bidirectional Encoder Representations from Transformers
CLEF	Conference and Labs of the Evaluation Forum
CLNF	Conditional Local Neural Fields
DAIC	Distress Analysis Interview Corpus
DCNN	Deep Convolutional Neural Network
EHIS	European Health Interview Survey
FACS	Facial Action Coding System
GBDT	Gradient Boosted Decision Trees
GRAM	GRaph-based Attention Model
LMHI	Landmark Motion History Images
LIWC	Linguistic Inquiry and Word Count
LSTM	Long Short Term Memory
MFCC	Mel Frequency Cepstral Coefficients
MRMR	Minimum Redundancy Maximum Relevance
NLTK	Natural Language ToolKit
PTSD	Post-Traumatic Stress Disorder
RMSE	Root Mean Square Error
DCGAN	Deep Convolutional Generative Adversarial Networks
NCS-R	US National Comorbidity Survey - Replication
PHQ-8	Patient Health Questionnaire
VADER	Valence Aware Dictionary and sEntiment Reasoner
BiLSTM	Bidirectional Long Short Term Memory
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders 4th Edition
LR-SGD	Linear Regression with Stochastic Gradient Descent

CHAPTER 1

Introduction

Depression, according to Dozois and Bieling [1], based on the American Psychiatric Association (APA) publication *Diagnostic and Statistical Manual of Mental Disorders 4th Edition (DSM-IV)* [2], "is a heterogeneous phenomenon that ranges from a mild and relatively transient negative mood state (dysphoria or despondency), often associated with a sense of loss, disappointment, or hopelessness, to a debilitating cluster of symptoms that impair most aspects of social or occupational functioning. In its clinical state, major depression refers to a constellation of symptoms that is associated with significant cognitive, emotional, behavioural, physiological, and interpersonal impairment". The most common symptom associated with it is sadness or loss of interest/anhedonia. Other symptoms can be changes in appetite/weight, psychomotor retardation, loss of energy or fatigue, feeling of worthlessness, self-blame or excessive guilt, impaired concentration and impaired ability to make decisions. In worst cases, it can also lead to suicidal ideation, recurrent thoughts of death and attempted suicide. In cases of recurrent depression, the diagnosis is Major Depressive Disorder (MDD) where the first episode is termed a major depressive episode [1].

According to the Global Burden Diseases, Injuries, and Risk Factors Study 2017 [3], more than 264 million people across the world are affected with depression, which indicates that depression is among the most common of psychiatric problems. A survey [4] from the US National Comorbidity Survey - Replication (NCS-R) has shown that about 9.5% of participants had a twelve-month prevalence for any mood disorder and 6.7% had a twelve-month prevalence for MDD. The estimates of an individual to suffer any mood disorder and MDD through his lifetime were 21% and 17%, respectively. The period of life with increased risk of a depressive episode is in the mid to late adolescence and early adulthood, where since adolescence, females are consistently two times more likely to experience depression than males. Depression is characterized by relapse and recurrence, with 50% to 85% of depressed patients experiencing multiple subsequent episodes. The risk of future episodes also increases and the time for an episode to recur decreases with each episode. Also, depression co-occurs often with other medical conditions, such as Alzheimer, Parkison, cancer and multiple sclerosis.

The highest comorbidity rates are depression and anxiety, which surpass 50%, substance abuse, schizophrenia and eating disorders.

According to a study developed by the European Health Interview Survey (EHIS) in 2019 and published by Eurostat [5], there was indeed a higher number of cases of chronic depression in the EU for women than men. 7.2% of EU citizens reported suffering from depression, representing a small increase from 2014 (6.9%). The data reveals that the highest share of the population suffering from chronic depression are most commonly from Western Europe and Central Europe. Figure 1.1 represents a graphic with the data from the survey.

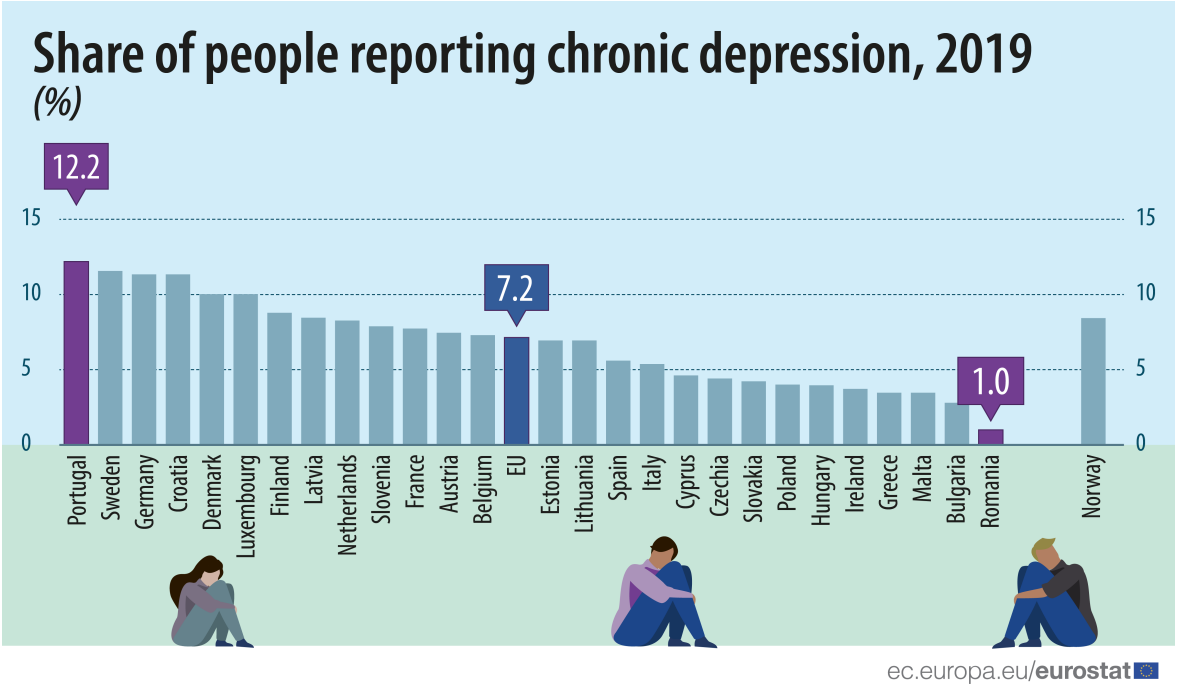

Figure 1.1: Eurostat chronic depression survey results [5]

As stated in the graphic, Portugal had the highest share of the population reporting chronic depression (12.2%), followed by Sweden (11.7%) and Germany (11.6%). The lowest share is from Romania, with only 1%. In addition to that, Portugal recorded the highest share of women reporting chronic depression (16.4%), followed by Croatia and Sweden (13.4%). The highest share of men reporting chronic depression is from Sweden (10.0%), followed closely by Germany (9.9%) then Denmark and Croatia (9.2%).

1.1 Motivation

By looking at these numbers, it is clear to say that it is important to be able to identify and quantify the severity of depression as early as possible in order to avoid major harm to the patient. So, in recent years a lot of studies tackling mental disorders have been published, including tools that can assist psychiatrists on their assessment [6] and machine learning models that can detect mental disturbances either in social media or in medical interviews.[7]

One of these studies [6] was developed by the University of South California (USC), in which the authors developed a virtual interviewer so that patients can be more comfortable and not feel they are being judged. The main idea was to develop a Virtual Human Interviewer that was designed to create interactional situations where it would be possible to automatically assess distress indicators correlated with depression, anxiety or post-traumatic stress disorder Post-Traumatic Stress Disorder (PTSD). These indicators could either be verbal or non-verbal behaviours (e.g. movements). From there, the Distress Analysis Interview Corpus (DAIC) was created, which is the main dataset used in this work. [8]

1.2 Objectives

The main objectives of this work were to explore machine and deep learning methods for analysis of depression in non-structured text and audio. Some of these methods were already published solutions in recent years, referring to the development of multimodal models. Also it was considered relevant if the developed model could achieve the specified baseline from Audio/Visual Emotion Challenge (AVEC) conferences.

This document is organized as follows. Chapter 1 introduces depression as a mental disorder by describing how it affects general population and shows some statistics of this particular disorder. Then the objectives of this dissertation are presented, while explaining the interest in studying this area. Chapter 2 is divided in three sections. The first section will be focused on State of the Art approaches from conferences that used the dataset that was made available for this dissertation (AVEC16 and AVEC17) and in papers from other conferences that studied the measurement of depression through text-only data or multimodal data. The second section details the dataset that this dissertation was developed on and gives an overview on some of the concepts in it. The third scetion presents a summary of the tools used for this work. Chapter 3 focuses on the implementation of this dissertation's solution. It is divided in three parts: pre-processing, feature extraction and selection and the implementation of the machine learning models. Chapter 5 presents and discusses the obtained results from the presented models and compares them with previous published results. Chapter 6 is focused on deducting conclusions from the acquired results and highlighting possible directions of future work.

CHAPTER 2

Background

This chapter is composed by three sections. The first section is an overview of the state-of-the-art. There, the article (SimSenseiKiosk) that first introduced the creation of the DAIC dataset will be detailed. Then, some models that used the DAIC dataset will be presented. This part includes the models developed in the International Workshop on Audio/Visual Emotion Challenge (AVEC). Also, some other models from other conferences like CLEF eRisk will be mentioned, even though they had different objectives or datasets. The second section is an overview of the dataset and some specific audio and video concepts related to it. The last section is a summary of the used tools in this work.

2.1 State-of-the-Art

2.1.1 SimSenseiKiosk

De Vault et al. [6] developed a virtual human interviewer so that the user can feel comfortable with sharing information and reduce stress associated with the user's perception of being judged. Besides, this virtual interviewer (named Ellie) is also able to create and react to interactional situations in order to identify behaviours correlated to psychological distress. The design was split in three development cycles:

- Analysis of face-to-face human interactions to identify potential distress indicators, dialogues and gestures;
- Development and analysis of a Wizard-of-Oz prototype system utilizing two wizard controllers, one for verbal cues and a second for non-verbal cues;
- Development of a fully automatic virtual interviewer able to engage users in 15 to 25 minute interactions.

For the first cycle, the focus was on acquiring and analysing human-human interactions in the same context of clinical assessments. In here, the interviewee behaviours were analysed to identify potential indicators of distress while the interviewer was analysed to identify proper questions and non-verbal behaviours to animate the virtual human. The interviews were

composed by an initial small talk with some neutral questions and then specific questions about possible psychological distress and traumatic events, For the second cycle, a Wizard-of-Oz prototype was created with two human operators to dictate the virtual human's responses and non-verbal behaviours. The virtual interviewer had a limited set of response options and non-verbal behaviours to try to act as a good listener by showing empathy responses and continuation prompts. In Table 2.1, the available option set is summarized.

Option type	**Count**	**Example**
non-verbal behaviours	23	head nod to indicate agreement
interview questions	87	*what are you like when you don't get enough sleep?*
neutral backchannels	24	*uh huh*
positive empathy	11	*that's great*
negative empathy	14	*I'm sorry*
surprise responses	5	*wow!*
continuation prompts	26	*could you tell me more about that?*
miscellaneous	24	*I don't know; thank you*

Table 2.1: Available option set for Wizard-of-Oz

After analysing the interactions, the authors concluded that there were identified significant differences between non-distressed and distressed participants. Also, the finite set provided to the wizards was deemed adequate to conduct interviews that could obtain different responses and behaviours from distressed participants when compared to the non-distressed participants.

For the third part, the virtual interviewer was developed. The core functionalities of the virtual interviewer are:

- audio-visual sensing and non-verbal behaviour understanding;
- natural language understanding and dialogue management;
- non-verbal behaviour generation and behaviour realization.

Examples of non-verbal behaviours for participants that scored positive for depression are different measures of Normalized Amplitude Quotient (NAQ) and peak slope, significantly less head motion variation and expressivity on average and the position of the face of the user and smile intensity. The natural language understanding of the virtual interviewer was composed of four statistically trained utterance classifiers, of which each one captured different aspects, such as generic dialogues (e.g. yes-no questions/answers, wh-questions), identify positive/negative valence to utterances, domain-specific small talk and domain-specific dialogue acts. To generate non-verbal behaviour, an already developed behaviour generation system was used and extended (Cerebella) [9] [10] . An example of dialogue between the virtual interviewer and a participant is shown in Table 2.2.

After the experimental phase of the virtual interviewer, the results were positive. The participants reported that they were willing to disclose and share sensible information with the Wizard-of-Oz and Artificial Intelligence (AI) versions. In terms of rapport, the participants reported similar results between the AI version and face-to-face interviews. Also, the AI version had significantly lower ratings of rapport and system usability when compared to the

Speaker	Utterance
Ellie	*How have you been feeling lately?* (top level question)
Participant	*Not good at all*
Ellie	*Can you tell me about that?* (continuation prompt)
Participant	*Um... yeah I mean... I recently got released from prison not too long ago... uh I'm older now... and a lot more responsibilities and I feel like my life is not where it's supposed to be... I feel like I could've made a lot more better decisions in my life... you know... sometimes feeling sorry for myself... and just things that I've been through and the whole nine yards so yeah*
Ellie	*I'm sorry* (empathy response)

Table 2.2: Example of Ellie's interaction.

wizard version. Participants also mentioned that the AI version was less sensitive to their body language and often produced inappropriate non-verbal behaviours. Regardless, the overall results were satisfactory and the system was already effective.

2.1.2 Articles using DAIC dataset

In this section, an overview of scientific papers using the DAIC dataset will be presented.

Context-Aware Deep Learning for Multi-modal Depression Detection

Genevieve Lam, Huang Dongyan and Weisi Lin [11] developed a multimodal context-aware model where they combined the use of a Transformer for topic modelling and a 1-Dimensional Convolutional Neural Network (CNN). The Transformer was fed with augmented data, by combining relevant topics with the associated transcriptions and audio features like the Mel-frequency spectrogram. The topics were manually extracted instead of using Latent Dirichlet Allocation (LDA), since only a few topics were discussed. The 1D CNN was used to model audio, where it performed convolutions over the time of mel-spectrograms of the participants. The authors chose to use 1-Dimension instead of two in order to have a stronger discriminative ability between audio classes. Then, both these models were fused into a feed-forward model so that it could classify correctly if the participant had depression. For evaluation, the authors only considered categorizing if a participant is depressed or not. They first evaluated the impact of the augmented data on the text and audio models and then the multimodal models. For the text model, where the Transformer was applied, the authors had 3 sets: one where the Transformer gets the full transcript, other with only the important topics found in the topic modelling and the last one where the Transformer gets the topics plus the augmented data to balance the two classes. All of these were fine-tuned with Adam optimizer. For the audio model, the authors also had three sets: one with the mel-spectrogram of the whole interview, the other with the mel-spectrogram of the specific topics and the last one had the same information of the previous set plus the topic-modelling based data augmentation.

These three sets consisted of four 1D convolution layers, with L2 regularization and ReLU activation functions. For the multimodal model, there were also three sets: the first one contained the first set of text and audio model, the second one contained the second model of text and audio model and the last one contained the last text and audio model. The results have shown that the data augmentation sets outperformed the previous published results, especially the text and the multimodal model.

Decision Tree based Depression Classification from Audio Video and Language Information

Le Yang et al. [12] developed a Support Vector Machine (SVM) classifier to estimate the risk of depression, where a fusion of the text features with the speech prosody features was added. It was also used a Histogram of Oriented Gradient (HOG) feature vector with Principal Component Analysis (PCA) in order to reduce the features. The audio and video features were trained in a separate Support Vector Regression (SVR) model with Radial basis function (Rbf) kernel. Then the SVR results with the lowest Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were fused to the multimodal model in order to predict the Patient Health Questionnaire (PHQ-8) score [13]. With the predicted PHQ-8 score, two Decision Trees were built (one for females and other for males), that took into account some relevant aspects of the participant's answers, such as if the person can sleep well or if the person ever got diagnosed with PTSD/Depression. For each of the Decision Trees, the authors modelled them according to the distribution of the answers from the participants (e.g. depressed women responded that they had suffered from PTSD/depression before). With the Decision Trees, then the authors classified if the participant had depression or not. This presented method reported an overfitting problem on the SVR models, which negatively influenced the classification of the decision trees.

Multimodal Measurement of Depression Using Deep Learning Models

Le Yang et al.[14] developed a combination of a Deep Convolutional Neural Network (DCNN) with Deep Neural Network (DNN), which was noted as DCNN-DNN. For the text features, only the participant's answers to specific topics like sleeping disorder or his/her personality were extracted. Then, these answers were the input for a Paragraph Vector model [15], which is an extension of the Word2Vec model that does not need labeled data to learn. For the video features, it was proposed a new global descriptor named Histogram of Displacement Range (HDR), using the 2D landmarks of the participant's face. For the audio features, the authors extracted for each audio segment several Low Level Descriptors (LLD) with the openSMILE tool [16], consisting of spectral and energy features and voicing related dynamic features. For most of the extracted LLD, the first and second order derivatives were also extracted, resulting in a 6092 dimension feature vector. The text, audio and video features were firstly trained in their separate DCNN model, with ReLU as an activation function with Euclidean loss as a loss function. After training the DCNN, the authors connected it with the DNN, also with Euclidean loss. Since the two networks were trained separately, there was no back-propagating loss to the DCNN from the DNN. Then, the outputs from each DCNN-DNN were fused in a DNN model which output the predicted PHQ-8 score. After

some data balancing, two gender-specific models were trained, in order to get the best possible results. The results revealed that the unimodal DCNN-DNN model got a better result than just a single DCNN model for audio and video. For the multimodal model, the results were far better than the baseline.

Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text

Anastasia Pampouchidou et al. [17] presented a fusion model with four different classification approaches:

- gender-based for each individual modality;
- feature fusion model;
- decision fusion model;
- posterior probability classification model.

In terms of video features, much of the DAIC pre-extracted features were used and derived, such as Head Motion, Action Units (AU) detection, and others were computed, like blinking rate or Landmark Motion History Images (LMHI). For the audio features, the pre-extracted features were used as a first set of features, after being computed the low level descriptors. Then, a second set of features consisted of the first ten values of Discrete Cosine Transform (DCT) coefficients, for each of the first set descriptors. A final third set of high level features was computed for the full concatenated time-series. An example of this high-level audio feature is the Mean delay in response to the virtual interviewer's questions. After combining the third set with the first and second set separately and forming two single feature vectors, the first+third feature was of size 494 and second+third feature was of size 1278. For the text features, some of them were derived from Linguistic Inquiry and Word Count (LIWC), others from Affective Norms for English Words (ANEW) and some were personalized, such as ratio of laughter over total number of words. For the feature selection, the top features were obtained empirically. For each of the different classification approaches, a decision tree was applied. For the gender-based, two separate classifiers were built and trained with their respective gender data. The feature level fusion concatenated the unimodal features in a single feature vector. The decision fusion utilized the produced labels from the individual classifiers per modality and combined them to get all the possible combinations. The posterior probability was based on the posterior probabilities from the Decision Trees. It consisted in three layers, where Layer 1 was the Decision Tree of Video Model and Layer 2 was the Decision Tree of Audio + Text Model. The inputs of Layer 3 are the probabilities from Layers 1 and 2 plus the gender label, resulting in a three-item feature vector. The results showed that the gender-based approach outperformed gender-independent in the audio, text model and the audio+text model. The Decision Fusion only performed well on the Development dataset, due to overfitting. When comparing to the baseline, only a model showed better results, which was the gender-based model for the audio model.

Graph Attention Model Embedded with Multi-Modal Knowledge for Depression Detection

Wenbo Zheng et al. [18] proposed a multimodal graph attention model embedded approach to classify and predict depression, similar to other solutions applied in different medical fields

(e.g. GRaph-based Attention Model (GRAM) [19]). Also, a multimodal attention mechanism was presented to assist achieving a better performance. Two models were presented, in which the first used a dilated temporal convolutional network and the second used a causal temporal convolutional network. The models receives an embedding knowledge matrix from the cross-modality shared attention mechanism, which was made of the inter-modality-level attention vector and the intra-modality-level attention vector. Figure 2.1 represents the proposed graph attention model with multimodal knowledge.

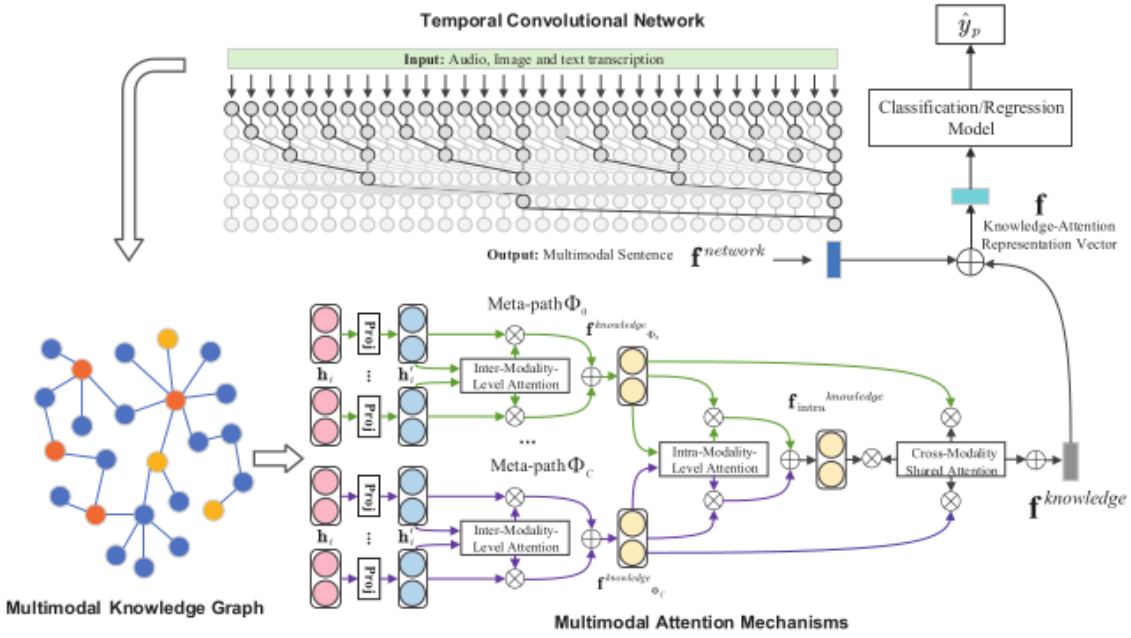

Figure 2.1: Graph attention model embedded with multimodal knowledge [18].

The results showed that this approach outperformed the state-of-the-art approaches in terms of classification and regression.

Multi-Modal Depression Detection and Estimation

In this article, Le Yang [20] had the objective of developing a model that can efficiently improve depression estimation/detection and utilize Generative Adversarial Networks (GAN) to generate depression data, as a way to tackle the lack of it. It was presented a two-level Deep Convolutional Generative Adversarial Networks (DCGAN) to generate audio features from the DAIC dataset, which were used to augment the training data. When compared with previous results, the data augmentation did effectively improve the performance of depression estimation, by reducing the Root Mean Squared Error on the regressor. Also, in order to get better results from Action Unit detection, it was proposed a 3-dimensional CNN integrated with a 2-dimensional CNN. The 3D-CNN receives the video segment, which obtains a global spatiotemporal information of that segment. Meanwhile, the 2D-CNN receives each frame and learns spatial information. Then, in a fully connected layer, the 3D-CNN output and 2D-CNN output is concatenated to achieve multi-label AU detection of the current frame. This model outperformed 2D-CNN and 2D-CNN-Long Short Term Memory (LSTM) approaches.

Multi-level Attention Network using Text, Audio and Video for Depression Prediction

In this article, Anupama Ray et al. [21] focused on an attention-based fusion network that used a slightly different version of the DAIC dataset. For each modality, the attention layer teaches the network the most important features within itself to create the context feature. The context features of each modality passed through feedforward networks that are fused with Bidirectional Long Short Term Memory (BiLSTM). The output of each feedforward contains the most important features per modality, which are concatenated and applied to another attention layer. Then, the output of this layer is fused with the output of the stacked BiLSTM output and passed through the regressor, giving the predicted PHQ-8 questionnaire result. The loss of the regressor is back-propagated to train the weights, in order to ensure end-to-end training. For the text features, the Universal Sentence Encoder (USE) [22] was used to get sentence embeddings. Then, these embeddings were used as input to two layers of stacked BiLSTM to train the model. For the audio, the Mel Frequency Cepstral Coefficients (MFCC) and the extended GeMAPS [23] features are some examples of the extracted features. Also, a high dimensional representation of the audio sample was extracted by passing it through a Deep Spectrum and a VGG network, to which was denominated as deep densenet feature. As a unimodal model, it also had a two layer stacked BiLSTM, with the last output layer to pass it to a multi-layer perceptron, with ReLU as activation function. For the video, the Pose, Gaze and Action Units of the participants were used as features to input in a BiLSTM model, followed by a maxpooling and then a regressor. For the multimodal model, several approaches were taken in order to validate the multi-layer attention network. Some combinations were video-text,audio-text and video-audio-text. The results showed that even though the models were only applied to the development subset, since there were no annotations on the test subset, they still were better than the baselines and in some situations (e.g. text), they were better than the state-of-the-art.

Detecting Depression with Audio/Text Sequence Modeling of Interviews

In this article, Alhanai, Ghassemi and Glass [24] had the objective to detect depression by modelling audio and text sequences of the interaction between the virtual agent and the participant. So three models were developed:

- logistic regression without conditioning on the type of questions asked;
- logistic regression with conditioning on the type of questions asked;
- LSTM using the sequences of responses and without knowledge of the type of questions that prompted the response.

For the first one, a logistic regression model was applied with L1 regularization and a feature vector containing 279 audio features and 100 text features was obtained after feature selection. For the second one, the logistic regression model applied in the first case was weighted with probabilities based on the question made by the virtual agent. The weights were based on the performance of k informative queries on the training set, that were above a specific threshold. For the third one, a bi-directional LSTM was used to model sequential data, with a hyperbolic tangent (tanh) activation function. The optimal hyperparameters were obtained empirically.

In this model, the audio model and the text model had their own BiLSTM with different hyperparameters, and then their outputs were merged into a feedforward network. The results revealed that for the first model, text features performed better than audio features in terms of classification. For the second model, audio features performed better, with excellent precision rates but with poor recall rates. Even so, when comparing the F1 metric of the audio features between the first and the second model, there was an improvement in performance by adding the question weights. The third model outperformed, especially the fusion model. The sequence model in general outperformed or had similar results with the first two models, with the best results being presented by the multimodal fusion model.

Depression-level assessment from multi-lingual conversational speech data using acoustic and text features

In this article, Cenk Demiroglu et al. [25] had the objective to develop a fusion model with text and audio features by using a multilingual feature selection strategy with datasets from three languages (English, German and Turkish). The English dataset is the DAIC dataset. The German dataset was distributed as part of an earlier edition of AVEC (AVEC'14). It consisted in recordings of 84 participants and their corresponding Beck Depression Inventory (BDI) questionnaires [26]. The Turkish dataset was collected at a hospital in Istambul, consisting of interviews and BDI questionnaires from 70 participants. Three feature selection models were proposed based in the Minimum Redundancy Maximum Relevance (MRMR) method:

- Multi-lingual computation of relevance (ml-MRMR);
- Clustering approach;
- Robust computation of redundancy (RCR).

The ml-MRMR method was proposed since there was limited training data in the depression detection problem, that usually leads to unreliable feature correlations. So, to increase the number of samples for each class, the authors assembled the samples from other datasets and populated them. Even so, there might be classes with low samples. So, samples from neighbouring classes in a different language are used if there is a small distance between the target class and the neighbour class. With this, some adjustments are made to the F-statistic in order to get better results. In the clustering approach, the depression classes are clustered and the number of classes in the MRMR training is reduced in order to improve the feature selection performance by increasing the data available for each class. Firstly the data is split uniformly into N classes and then the centroids are calculated. The centroid must be a non-empty class. When the centroids are calculated, each class is assigned to the nearest centroid, allowing for a more uniform distribution of samples per class. For the RCR approach, since class labels are not required to compute redundancy (Pearson's correlation), it is possible to use unlabeled speech data to compute redundancy for feature selection. With this approach, Pearson's coefficient distribution had shown lower variance for the English dataset which improved the performance of the feature selection. For the fusion model, a feature extraction was needed to be performed. For the text feature extraction, there were 15 extracted features, mostly related to the rate of speech, sentiment associated with the question-answer pair and

average length of the utterances. For the audio features, the datasets already provided them. An example of these features are the extended GeMAPS features and the MFCC. To improve the reliability of acoustic-based features, the approach taken was to split the data into two classes: one with high levels of depression score (BDI $>$ 30) and the other with low levels of depression (BDI $<$ 18). If the audio model generates an in-between estimate, the text model is taken into account. If the text model also generates an estimate between 18 and 30, then the audio model estimate is used. Else (e.g. audio model predicts high results and text model predicts low result), the final estimate is computed by fine-tuning the audio model prediction by getting it closer to the opposite class. Also, a baseline system applied with a MRMR feature selection and SVR for regression problem and SVM for classification problem was taken into consideration and evaluation. In this paper there were two sets of experiments: Compare the proposed feature selection algorithms with baseline MRMR for all datasets and test for regression and classification tasks; Fusion of text and audio features for the Turkish dataset and compare with the baseline. In terms of results, the ml-MRMR outperformed the baseline MRMR in some regression tasks. The RCR approach was only effective when applied with ml-MRMR. The clustering approach had also good results, but the ml-MRMR approach had better ones. Also, the fusion algorithm improved the performance of the estimation when compared to the baseline, by reducing the spread of the prediction errors. On top of that, the best results are obtained when using the ml-MRMR approach with the fusion algorithm.

2.1.3 CLEF articles

The following articles were presented on the Conference and Labs of the Evaluation Forum (CLEF) eRisk 2020 [27], with models for the second sub-task, whose objective was to measure the severity of the signs of depression through a historic analysis of social media posts and BDI questionnaires [28] [1].

Deep learning architectures and strategies for early detection of self-harm and depression level prediction

Ana-Sabina Uban and Paolo Rosso [29] extracted numerical features from social media posts using the LIWC and the emotional state of the user with LIWC and National Research Council Emotion Lexicon (NRC) [30]. Then, the authors calculated the average and standard deviation of these features per post, in order to obtain their feature vector. So, three models were presented. One was a logistic regression model with the aforementioned features as numerical vectors, the second was a SVM with Rbf kernel, using the same features as the first one, and the last was a SVM with Rbf kernel with USE features for each of the posts in a user's history.

Early Risk Detection of Self-Harm and Depression Severity using BERT-based Transformers

In this paper, Rodrigo Martínez-Castaño et al. [31] focused on exploring BERT-based models to classify depression. The authors presented three models, two of these used RoBERTa, which

[1] https://psychologicalprofessional.com/wp-content/uploads/2017/07/Becks-Depression-Inventory-BDI-II.pdf

is a replication study of Bidirectional Encoder Representations from Transformers (BERT), developed by Facebook AI. The other model is XLM-RoBERTa, also developed by Facebook AI, in an attempt to improve multilingual language models. The main difference between these presented models is the base language model and the pre-processing of the training data.

Early Mental Health Risk Assessment through Writing Styles, Topics and Neural Models

In this paper, Diego Maupomé et al.[32] took two different approaches. One was user-based, which tries to relate users to each other, and the other was answer-based, which tries to relate a user to the text of a filled depression questionnaire. Then, LDA was applied to create topic vectors and measure the distance between them for both approaches. Also, a Contextualizer encoder was applied. For the user-based case, the documents were encoded in parallel and the similarity between them was calculated with the angular similarity. For the answer-based ones they were encoded simultaneously, outputting the probability of the author being the same. A fifth model based on the writing style of a document (or stylometry) was also used in order to try to characterize its author. For this, several linguistic features were used, such as frequency of words and characters and length of sentences and words. The models with better performances were the user-based instead of the answer-based, since they did not require annotated data (users with known depression questionnaire scores).

Deep learning models to measure the Severity of the Signs of Depression using Reddit Posts

Amina Madani et al. [33] used the Skip-gram model to obtain a vector space which was used to predict the context word for a given target word, in order to find word representations that would be useful to predict surrounding words in a post. Then, two Deep Learning models were presented. One consisted in using a CNN with a ReLU activation function and the other was a combination of a Recurrent Neural Network (RNN) with Long Bi-Long Short Term Memory(BiLSTM). In the BiLSTM model, the first LSTM analyses from the beginning of the sentence to the end and the other analyses from the end to the beginning. Both models calculated for a question the frequency of each generated answer to choose the most recurrent answer.

Use of psycholinguistics features and machine learning for the classification and quantification of mental diseases

Alina Trifan, Pedro Salgado and José Luís Oliveira [7], combined a Tf-idf weighting scheme for Bag of Words features with a rule-based estimator, taking into account several psycho-linguistic features that characterize depressed users. Some of these psycho-linguistic features are e.g. use of self-related words (me, myself, I), use of absolutist words (never, all, always, nothing, etc.) and mentions of words related to mental disorders (depression, bipolar, psychotic, OCD). The model using Tf-idf with a Passive Aggressive classifier with a batch training presented the best results.

2.1.4 Other articles

The following articles are unrelated to the DAIC dataset nor with AVEC and CLEF eRisk conferences. However, since the authors also studied and developed models to quantify and predict depression, the articles were deemed relevant for this dissertation.

Text-based Detection of the Risk of Depression

In this study performed by Jana M. Havigerová et al. [34], the main objective was to discover a relationship between linguistic characteristics of a written text and the level of the emotional state of depression of its author. So, it was decided to separate and analyse both genders and multiple genres of texts, that split into two categories: formal and informal letters. The study was conducted to Czech speakers, while using the Depression, Anxiety and Stress Scale (DASS-21) as a measure of depression. The participants had to write four fictional letters and were recommended to write about 180 to 200 words for each one. The letter scenarios was:

- Cover letter (formal, positive sentiment);
- Letter from holidays (informal, positive sentiment);
- Complaint (formal, negative sentiment);
- Letter of apology (informal, negative sentiment).

After collecting all the texts, a linguistic analysis was performed, with a total of 24 linguistic features extracted. Some examples are e.g. words per sentence, ratio of verbs per number of sentences, coherence index

$$Coh = \frac{particles + conjunctions + prepositions}{3 * numberofsentences} \tag{2.1}$$

and aggressiveness index.

$$aggressiveness = \frac{numberofverbs}{wordcount} \tag{2.2}$$

After performing some data analysis and feature selection, the feature vector was introduced in a regression model defined by the Nagelkerke coefficient ($r^2 > 0.4$). For each letter scenario and gender, a regression model was created (8 models total). Although there were some limitations on the results due to a small sample and a high percentage of depressed men in the sample, it was still able to conclude that the best letter to predict depression is with a text describing a holiday for both men and women. Also, the best linguistic features differ from male and female.

Acoustic and language analysis of speech for suicide ideation among US veterans

This study performed by Anas Belouali et al. [35] had the objective of developing a model that can detect suicide ideation among US veterans, by extracting and analysing audio recordings. Since suicide ideation is one of the most dangerous signs of depression, it was considered suitable. In this study, the veterans used a smartphone app to record their answers to the psychiatric questionnaires and were prompted to answer open-ended questions regarding their health in the recent weeks. With the audio recordings, then it was possible to extract audio

features and transcribe them, so that it was also possible to extract linguistic features. For the acoustic features, some of the extracted ones are the chroma vector and its deviation, MFCC, and some prosodic features as well, like the fundamental frequency (F0). For the linguistic features, the transcriptions were obtained by using the Google speech-to-text API[2]. Then, the linguistic features were extracted. Some examples are sentiment analysis, by assessing the general polarity of the recordings, some LIWC features and word frequencies of absolutist words. For the feature selection, an ensemble approach was taken to select the top performing features. Since there was a class imbalance on the dataset, the SMOTE technique was applied to oversample the minority class in the training sets. This was done in order to avoid information leakage. Then, six algorithms were tested, such as, Logistic Regression, Random Forests (RF), SVM, eXtreme Gradient Boosting (XGB), K-nearest neighbours and Deep Neural Networks. For model evaluation, a 5-fold nested cross-validation was performed. Then, three different models were built to assess the performance of audio and text features separately and also combined. In terms of results, the recordings from "suicidal" veterans were majorly different from non-suicidal in terms of energy. The XGB classifier performed best on acoustic features, while Random Forests performed better in linguistic features. Overall, tree-based models did outperform the remaining models.

2.2 Dataset and Concepts

The DAIC dataset [8] contains clinical interviews to support the diagnosis of psychological distress conditions, like anxiety, depression and post-traumatic stress disorder. It is composed of 188 interviews and PHQ-8 questionnaire answers. The interview data includes:

- Full interview transcripts, with verbal and non-verbal features (e.g. laughs, sighs);
- Full audio recording of the interviews;
- Extracted audio features from the interviews;
- Extracted video features from the interviews.

The dataset is split into three parts: the training subset, which contains interview data from 107 participants; the development (dev) subset that contains 35 of them and the test subset that has the remaining 46. The transcripts contain both interviewer and interviewee lines, where each line has an annotation of when it started and when it finished. Example:

107.57	108.594	Ellie	why did you move to la
110.02	114.87	Participant	because i wanted to pursue my acting career

The transcripts are in tab-separated values format (tsv), where most words are in lower case except e.g. locations. Incomplete words are completed and have an annotation on how the word was pronounced. Example: people <peop>. Also, unrecognizable words are marked as 'xxx'.

[2]https://cloud.google.com/speech-to-text/

The audio recordings were recorded with a head mounted microphone at 16kHz. The audio features are sampled at 100Hz every 10ms. These features were extracted with COVAREP [36] (v1.3.2) [3] and were split into two comma-separated values (csv) files, where one of them contains the first five formants and the other contains the remainder of the features. These features will be detailed in 2.2.1. In the extracted audio features, there is also a flag that mentions if there is voice activity from the participant in a specific segment, which was denominated VUV (Voiced/UnVoiced). In this dissertation, unvoiced segments were discarded. This was the approach since the vocal folds were detected to not be vibrating, which would bring misleading values in most of the audio features.

The video features are separated into several files. The first contains 68 2D points of the participant's face for each timestamp. The second contains the detection of AU. The third contains 68 3D points of the participant's face for each timestamp. The fourth contains the gaze direction of both eyes. The fifth is a binary file containing Felzenswalb's HOG [37] of the participant's face. The last one contains the pose of the participant, including the rotation of the participant's head.

The PHQ-8 Questionnaire [4] is a multiple-choice self-report inventory that is used to screen patients for the presence and severity of depression. The questions cover specific topics related to depression symptoms such as eating problems, sleeping problems and loss of interest. Each question has four options that range from "0" (Not at all) to "3" (Nearly every day). After filling the questionnaire, the clinician sums all of the responses and categorizes the severity of depression. In Table 2.3, there's a representation of the categories of PHQ-8 score and their corresponding severity.

Score	**Severity**
0-4	None/Minimal
5-9	Mild
10-14	Moderate
15-19	Moderately Severe
20-24	Severe

Table 2.3: PHQ-8 scores and corresponding severities.

In general, a score equal or above 10 represents that the patient has depression.

2.2.1 Audio Concepts

In Figure 2.2, there is an overview of COVAREP and its methods. COVAREP (Collaborative Voice Analysis Repository for speech technologies) is an open-source tool for speech processing algorithms and techniques. These techniques particularly focus on:

- Periodicity and Synchronisation;
- Sinusoidal Modeling;
- Spectral Envelope Estimation and Formant Tracking;
- Glottal Analysis;

[3] https://github.com/covarep/covarep
[4] https://www.selfmanagementresource.com/docs/pdfs/English_-_phq.pdf

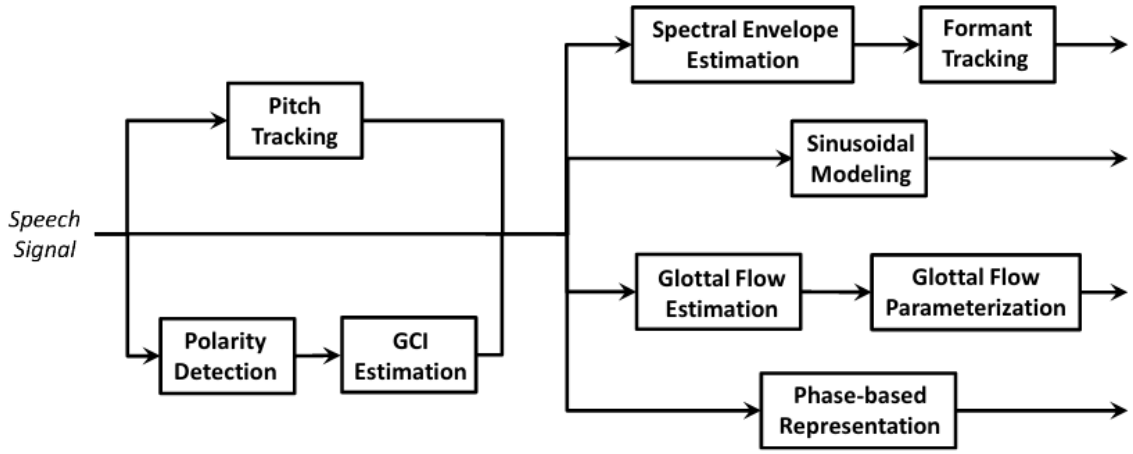

Figure 2.2: Overview of COVAREP [36].

- Phase Processing.

The Periodicity and Synchronisation methods are mostly related to pitch tracking, speech polarity detection and detection of the glottal closure instant. An example of the pitch tracking part is the estimation of the Fundamental frequency, which is used to represent the periodicity of the speech signal. Speech polarity detection is particularly relevant on the performance of multiple analysis techniques (e.g. Glottal Closure Instant estimation or glottal analysis), that is based on the skewness of the Linear Prediction (LP) residual signal. The GCIs, being the pseudo-periodic instants of significant excitation of the voice, requires a great detection algorithm that is accurate and robust so that it is possible to perform pitch-synchronous analysis procedures and non-modal phonation, which is present in this tool. It is possible to get Sinusoidal Models such as the Harmonic Model or the quasi-harmonic Model by extracting the amplitude peaks of the Discrete Fourier Transform (DFT) of short time windows of voiced speech signals. This is possible because the periodicity caused by the glottal excitation translates to a harmonic structure in the speech spectrum. So, by exploring sinusoidal and harmonic model parameters, it is possible to get other models, such as spectral envelopes or glottal flow parameterisation. For formant tracking, it is possible to use temporally weighted LP methods or Discrete All-Pole just as it is used for spectral envelope estimation or by processing the negative derivative of the argument of the chirp-z transform (differential phase or group-delay spectrum). The glottal analysis can be split in two parts: the glottal flow estimation and the glottal flow parameterisation. Glottal flow (GF) estimation or source-filter separation is the process of vocal-tract and glottal flow components estimation from a speech signal. It is important to separate them since it enables their distinct characterisation and modeling. The parameterisation of the glottal flow is extremely relevant since it provides useful applications for speech processing, particularly for this work. Parameters such as NAQ or Quasi-Open Quotient (QOQ) are some of the features provided in the DAIC dataset. As for the phase processing, the Relative Phase Shift and its frequency derivative (Phase Distortion) are made available to extract meaningful phase information. The first one is focused on the

perceptual importance of phase information in speech and speaker verification while the latter is more oriented to emotional valence detection and also glottal parameter estimation.

A general overview of the pre-extracted features in the dataset with COVAREP is presented below:

- F0- Fundamental Frequency, or the first harmonic. It's the primary acoustic correlated with pitch, which is affected by the frequency of vocal fold vibration at the glottis;
- NAQ- Normalized Amplitude Quotient, method to parameterise the glottal closing phase;
- QOQ- Quasi-Open Quotient, method correlated to Open Quotient, parameterises the glottal open phase;
- H1H2- Also known as h1-h2 ratio, refers to the ratio from the first harmonic and the second harmonic;
- PSP- Parabolic Spectral Parameter, for the quantification of the glottal volume velocity waveform;
- Rd- shape parameter of the Liljencrants-Fant model (LF) of the glottal pulse dynamics;
- MDQ- Maxima Dispersion Quotient, glottal measure to differentiate breathy from tense voice;
- peakSlope- spectral tilt/slope of wavelet responses from the glottis;
- MCEP- Mel Cepstrum Envelope, represents the vocal tract and the source excitation of the speech frame. It is also the Fourier transform of a spectral envelope of the Mel log spectrum;
- HMPD: Harmonic Model and Phase Distortion
  - HMPDM- Number of log-Harmonic coefficients (24) (Phase Distortion Mean);
  - HMPDD- MFCC-like phase variance (12) (Phase Distortion Deviation);
- Formants- vocal tract resonance frequencies;
- MFCC- Mel Frequency Cepstrum Coefficient, represents the short-term power spectrum of a sound, usually used as features for speech recognition problems.

Also, the dataset provides an extra acoustic feature related to the phrasal creak, which was discarded since it is a zero-value column.

2.2.2 Video Concepts

The video features presented earlier were extracted from Openface [38], an open source framework that implements facial behaviour algorithms, such as Conditional Local Neural Fields (CLNF) [39] for facial detection and tracking. The CLNF model extracts specifically the facial landmark and tracking, an estimation of the head pose and an estimation of the eye gaze. This model is based on a Point Distribution Model (PDM) that captures landmark shape variations and patch experts to capture local variations of each landmark. The Action Units, as stated by the Facial Action Coding System (FACS), is a standard to categorize human facial movements as a physical expression of emotions. So, this framework presents a module that can detect the presence and the intensity of Action Units. In Table 2.4, a list of AUs that this framework can detect is presented. The I represents Intensity and the P

represents Presence. Also, a representation of the aligned face of a participant is extracted and saved in HOG. The HOG is particularly used for human detection or in this case facial movement detection. It counts the occurrences of gradient orientation in specific blocks of an image.

AU	Full name	Prediction
AU1	Inner brow raiser	I
AU2	Outer brow raiser	I
AU4	Brow lowerer	I
AU5	Upper lid raiser	I
AU6	Cheek raiser	I
AU7	Lid tightener	P
AU9	Nose wrinkler	I
AU10	Upper lip raiser	I
AU12	Lip corner puller	I
AU14	Dimpler	I
AU15	Lip corner depressor	I
AU17	Chin raiser	I
AU20	Lip stretched	I
AU23	Lip tightener	P
AU25	Lips part	I
AU26	Jaw drop	I
AU28	Lip suck	P
AU45	Blink	P

Table 2.4: List of AUs in OpenFace framework.

2.3 Tools

This work was developed with the aid of these following tools:

Python

Python [5] is an interpreted high-level programming language with high readability. It is one of the most popular programming languages in general and in machine learning applications. It was the selected programming language for this work.

NumPy

NumPy [6] [40] is an open source Python library that offers extra functionalities for matrices and arrays operations. Its use in this work is crucial, since most tools depend on this library.

spaCy

spaCy [7] [41] is an open source library designed for Natural Language Processing. In this work, it was used especially in the extraction of text features from the interviews transcriptions.

[5] https://www.python.org/
[6] https://numpy.org/
[7] https://spacy.io/

NLTK

Natural Language ToolKit (NLTK) [8] [42] is a Python library used in Natural Language Processing tasks. Here, it was used in the pre-processing part as a parser and as a tokenizer.

Librosa

Librosa [9] [43] is an audio and music processing library in Python. It was used to process the interviews and extract some audio features, like the MFCC.

VADER

Valence Aware Dictionary and sEntiment Reasoner (VADER) [10] [44] is a Python tool to perform sentiment analysis. It was used for sentiment analysis in the text model.

gensim

gensim [11] [45] is an open source library for unsupervised topic modelling. In this work, gensim was used to perform topic modelling in order to extract the most recurrent topics from the participants sentences.

scikit-learn

scikit-learn [12] [46] is a Python machine learning library. Some of the available models by this library were used in this work.

COVAREP

COVAREP [13] is an open source project of speech processing algorithms. An overall explanation of this tool is given in 2.2.1. The audio features in the dataset were computed with this tool.

Transformers

Transformers [14] [47] is a library with multiple available pre-trained models. In this work, it was used for sentiment analysis.

2.4 Summary

This chapter focused primarily on the research made for estimating or identifying depression, either by analysing clinical data or by analysing social media history postings. Also, the DAIC dataset was explained, while reviewing some concepts related to the audio and video features. The final section introduced the tools used for this dissertation.

[8] https://www.nltk.org/
[9] https://librosa.org/
[10] https://github.com/cjhutto/vaderSentiment
[11] https://radimrehurek.com/gensim/
[12] https://scikit-learn.org/stable/index.html
[13] https://github.com/covarep/covarep
[14] https://huggingface.co/transformers/

CHAPTER 3

Implementation

In this chapter, the work developed in this dissertation will be detailed. In this work, the developed experiments focused on the text and audio modality and ultimately on a fusion model with both modalities. The first sub-chapter is focused on pre-processing, the second one on feature extraction and selection and the last one on the machine learning model implementations, which includes a topic modelling experiment.

3.1 Pre-processing

The pre-processing had two steps: the first one was the processing of the transcripts and the other one was the processing of the interviews recordings. Since the transcripts provide the full interviews of the participants in a TSV file, it was possible to extract the participants' answers. With these, two text pre-processing procedures were implemented: one for the text model itself and the other for the topic modelling. For the text model, the NLTK regex tokenizer was used. Then, every word with less than three characters, scrubbed entries (marked as 'xxx') and stopwords were discarded. Stopwords are the most common words in a specific language (e.g. 'the', 'a', 'an', 'I', etc.). In this case, the stopword list that was used was the NLTK english stopwords. For the topic modelling, the NLTK default word tokenizer was used. Then, spaCy's Part-of-Speech tagging was used to identify nouns and verbs. After identifying them, the remaining tokens were discarded. The purpose of doing this way was to try to find more specific topics, since the topics with adverbs or adjectives may not provide better results. For the audio part, since the dataset had a set of audio features already extracted, the first 40 MFCC of each interview was processed and stored in a text file. Also, unvoiced segments were discarded, as mentioned in Section 2.2.

3.2 Feature extraction and selection

3.2.1 Text feature extraction

For the text features, some of them were derived from LIWC [48] and others from [34]. The extracted features are either simple metrics such as the number of sentences or more specific

ones like non-verbal expression ratio. The extracted text features are:

- number of sentences (1);
- average word count per sentence (2);
- reflexive pronoun ratio (3);
- adjective ratio (4);
- adverb ratio (5);
- non-verbal expression ratio (sighs,laughters,sniffles and uhms) (6-9);
- pronominalisation index (Equation 3.1) (10);
- readiness to action (Equation 3.2) (11);
- sentence complexity (Equation 3.3) (12);
- positive sentiment and negative sentiment ratio computed by the transformers sentiment analysis model (13-14);
- positive sentiment and negative sentiment ratio computed by VADER (15-16).

In total, there were 16 text features extracted. The main difference between the sentiment analysis performed by the transformers and the one performed by VADER is that the first one is a pre-trained BERT model that can only return a positive or a negative response according to the requested sentence. The latter only assesses the polarity of the sentence, by summing the valence scores of each word and then normalizing them. It can return positive, neutral or negative responses.

$$Pronominalisation - index = \frac{totalpronouns}{totalnouns} \tag{3.1}$$

$$Readiness - to - action = \frac{totalverbs}{totalnouns} \tag{3.2}$$

$$Sentence - complexity = \frac{totalverbs}{totalsentences} \tag{3.3}$$

3.2.2 Audio feature extraction

For the audio features, a similar approach was taken like most of the articles presented in Section 2.1.2. Therefore, all the extracted features mentioned in 2.2.1, including the first five formants, and the MFCC previously extracted in Section 3.1 were included. For each of the MFCCs, the computed descriptors were:

- arithmetic mean;
- standard deviation;
- skew;
- kurtosis.

For the COVAREP and the formants, the extracted descriptors were:

- arithmetic mean;
- standard deviation;
- skew;
- kurtosis.

- max value;
- min value;
- variance;
- median.

Also, the number of voiced segments by the participant was included as a feature. In total, there were extracted 777 audio features.

3.2.3 Feature Selection

To select the best combination of features, a couple of measures were taken. Firstly, the features with low variance were discarded. This is a regular approach because low variance features are not informative enough, which leads to worse results. Then, for the audio features, the highest scoring percentage of features on a univariate statistical test were only taken into account. The test used was f_regression, which is derived from Pearson's correlation coefficient (or r_regression) and returns a F-score and its corresponding p-values.

3.3 Implementation of Machine Learning Models

For the implementation of the machine learning models, the experiments were separated in two parts: isolated implementations for the Text model and for the Audio model and a fusion model of Text+Audio features. The considered machine learning models were:

- Linear Regression with Stochastic Gradient Descent (LR-SGD);
- SVR;
- RF;
- AdaBoost regressor;
- Gradient Boosted Decision Trees (GBDT);
- Multi-Layer Perceptron (MLP).

The LR-SGD is a linear approach to fit a model with a linear predictor function. The SGD is an iterative method used to optimize the objective function in a smooth way via learning rate, where the gradient loss is estimated at each sample and the model is consequently updated. The SVR is an adapted version of SVM to suit regression problems. It is based on statistical learning frameworks and only depends on a subset of the training data, since the cost function ignores any training samples that are close to their target. The RF is an ensemble algorithm based on randomized decision trees. This algorithm uses perturb-and-combine techniques designed for trees, which creates a diverse set of classifiers by introducing randomness. The prediction of the ensemble is based on the average prediction of individual classifiers. This algorithm can achieve a reduced variance by combining multiple trees, at the cost of an increase of bias, although it often provides an overall better result. The scikit-learn implementation combines classifiers by averaging their probabilistic prediction, instead of choosing the class by majority vote. Adaboost, short for Adaptive Boosting, is a popular boosting algorithm. It is based on fitting a sequence of weak learners, such as small decision trees, on repeatedly modified versions of the data. The predictions from these weak learners are combined into a

weighted sum (or majority vote) to produce the final prediction. The algorithm is considered adaptive since weak learners are corrected through several iterations when the prior ones misclassified the prediction. The scikit-learn AdaBoost implementation is AdaBoost.R2 [49]. The GBDT is a generalization of boosting that allows optimization of arbitrary differentiable loss functions. Like other boosting methods, it builds an additive model in a forward stage-wise fashion. It is an accurate and effective method used in regression problems. The MLP is a feed-forward neural network that trains using backpropagation with the identity function as activation function in the output layer. It has three layers (input, hidden and output) which each one of them contains perceptrons. The scikit-learn uses the square error as the loss function.

For all the models, the best combination of hyper-parameters was empirically discovered. From Table 3.1 to Table 3.6, the hyperparameters from each model are presented. The LR-SGD model (Table 3.1) has seven different hyperparameters. The loss function measures the fitting of the model. The penalty is a regularization term that penalizes the model complexity. Alpha is a non-negative constant that controls the regularization (or penalty) strength. If the learning rate is set as *optimal*, then alpha also influences it. Epsilon is a regulator for the epsilon-insensitive and squared-epsilon-insensitive loss functions, which are equivalent to a SVR loss function. For epsilon-insensitive, epsilon is used as a threshold to evaluate the difference between the current prediction and the correct value. If the difference is lower than the threshold, it is ignored. The learning rate is used to control the step-size in the parameter space. If it is set to *invscaling*, then the learning rate is calculated with eta0 and power_t parameters. For this model, the scikit-learn Standard Scaler was used, which standardizes the features by removing the mean and scaling them to the unit variance.

LR-SGD	Loss function	Penalty	Alpha	Epsilon	Learning Rate	eta0	power_t
Text	epsilon-insensitive	L1	0.0001	0.1	Optimal	NA	NA
Audio	squared-error	L1	0.01	NA	Invscaling	0.1	0.5
Fusion	squared-epsilon-insensitive	L1	0.1	0.1	Invscaling	0.1	0.5

Table 3.1: LR-SGD hyperparameters.

The SVR (Table 3.2) has three hyperparameters. The kernel takes the data as input and transforms it, according to the kernel's specific mathematical function. The C is a non-negative regularization parameter. The strength of the regularization is inversely proportional to C. Epsilon, as stated earlier, it is used as a threshold to evaluate the difference from the actual value and the predicted value. The Standard Scaler from scikit-learn was also implemented.

The RF (Table 3.3) has only two parameters. Number of estimators refers to the number of trees in the forest and the criterion is the function used to measure the split's quality.

Adaboost's learning rate refers to the weight applied to each weak learner at each boosting

SVR	Kernel	C	Epsilon
Text	RBF	0.1	0.01
Audio	RBF	1	0.5
Fusion	RBF	1	0.5

Table 3.2: SVR hyperparameters.

RF	Number of estimators	Criterion
Text	50	MSE
Audio	500	MSE
Fusion	500	MSE

Table 3.3: RF hyperparameters.

iteration. A higher learning rate can increase the contribution of each learner. The loss function for Adaboost and GBDT is specifically used when updating the weights after a boosting iteration. The learning rate for the GBDT act differently from Adaboost, since in GBDT it shrinks the contribution of each tree. Table 3.4 and Table 3.5 show the hyperparameters for Adaboost and GBDT, respectively.

AdaBoost	Number of estimators	Learning rate	Loss function
Text	50	2	Exponential
Audio	50	5	Exponential
Fusion	50	0.1	Linear

Table 3.4: AdaBoost hyperparameters.

GBDT	Loss function	Learning rate	Number of estimators
Text	Squared error	0.01	100
Audio	Absolute error	0.01	1000
Fusion	Absolute error	0.01	1000

Table 3.5: GBDT hyperparameters.

The MLP has five different parameters. The activation function parameter is for the hidden layer only, because the activation function in the output layer is the identity function by default, as mentioned earlier. The solver is used for weight optimization. The alpha is the regularization term for L2 penalty, which is the default.

MLP	Hidden layer size	Activation function	Solver	Alpha	Learning rate
Text	100	Logistic	Adam	0.1	Constant
Audio	100	Identity	Adam	0.1	Constant
Fusion	100	Identity	Adam	0.1	Constant

Table 3.6: MLP hyperparameters.

The metrics used to evaluate the performance of each of the models are as follows:

- RMSE (3.4);

- MAE (3.5);
- F1-Score (3.6).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(P_i - O_i)^2}{n}} \tag{3.4}$$

$$MAE = \sum_{i=1}^{n} \frac{|O_i - P_i|}{n} \tag{3.5}$$

P_i - Predicted value for participant i

O_i - Observed value for participant i

$$F1_{score} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.6}$$

The RMSE and the MAE are specific regression metrics used by the AVEC conference to evaluate the performance of the models. These metrics are computed taking the golden values and the predicted values into account. The F1-Score is a classification metric used to evaluate the capability of the model to categorize if a participant has depression. After calculating the predicted PHQ-8 for each participant, each value will then be checked if it represents a participant with depression (PHQ-8 >= 10), as mentioned in Section 2.2.

3.3.1 Topic Modelling

As an experiment, topic modelling with LDA was applied to the dataset. LDA is an unsupervised learning method that is used to discover common "topics" in a series of documents. In this case, the objective was to find if there were different topics from depressed participants and non-depressed participants. The fine-tuning of LDA was based in the Coherence metric (c_v), which is typically used for topic modelling evaluation. LDA has four parameters, which are the alpha, the eta, the number of passes and the number of topics. Alpha handles the number of topics per document and eta handles the number of words per topic. The number of passes represent the number of times the corpus is iterated during training. The number of topics represent the number of topics extracted from training. The best parameters were with alpha as *asymmetric* (Equation 3.7), eta as *symmetric* ($\frac{1.0}{n_{topics}}$), with two passes and 10 topics, with a coherence of 0.782.

$$alpha = \frac{1.0}{topic_{index} + \sqrt{n_{topics}}} \tag{3.7}$$

3.4 Summary

This chapter focused on the work developed, from pre-processing of the interviews to the development of a fusion model. In here, an overview of the pre-processing phase, an explanation of the features and their corresponding selection for each modality was presented. Then, the developed experiments were detailed, while mentioning the applied machine learning models.

CHAPTER 4

Results and Discussion

In this chapter, the achieved results from the experiments mentioned in Section 3.3 will be presented. Besides, a comparison between non-depressed and depressed participants will be shown. In the end, a comparison between the AVEC baseline and the best performing models will be made.

4.1 Analysis of features

The training subset is composed of 31 depressed and 76 non-depressed participants, while the test subset has 13 of 46 depressed participants. On average, depressed participants spoke less (148 sentences) than non-depressed participants (161 sentences) in the train subset, but the same does not happen on the test subset (201 vs 176). Also, depressed participants spoke slightly shorter sentences (8.1 on train and 8.7 on test to 9.1 non-depressed). When performing sentiment analysis with transformers, results showed that depressed participants on average spoke just slightly more negatively (negative:51.8%) when compared to non-depressed (negative:49.3%). On the test subset, the difference is even shorter (depressed 50.8% to non-depressed 50.3%). With VADER, the results showed that on the train subset, depressed participants had 38.2% positive and 14.6% negative sentences, while non-depressed participants had 40.7% positive and 14.1% negative sentences. On the test subset, depressed participants showed 37.7% positive and 13.2% negative and non-depressed 39% positive and 14.2% negative. The rest of the text features had similar results between depressed and non-depressed, which makes them irrelevant for quantifying or detecting depression. When analysing the audio features, the most discriminative features were the descriptors related to MFCCs, especially the skewness. On average, the biggest difference between depressed and non-depressed participants in the final feature matrix was the standard deviation of the first MFCC in both subsets.

4.2 Experimental Results

This section is dedicated to the results of the experiments. First, the results from the topic modelling experiment will be presented. Then, the results from the development subset will be presented. Next, the results from the test subset are presented. In the end, a comparison between both splits and AVEC baseline [50] is made.

4.2.1 Topic Modelling Results

As mentioned in Section 3.3.1, a LDA model was implemented. With the visualization tool pyLDAvis [1], it was possible to get a visual representation of the topic distribution. Figure 4.1 is a representation of the LDA model with best Coherence. When analysing Figure 4.1,

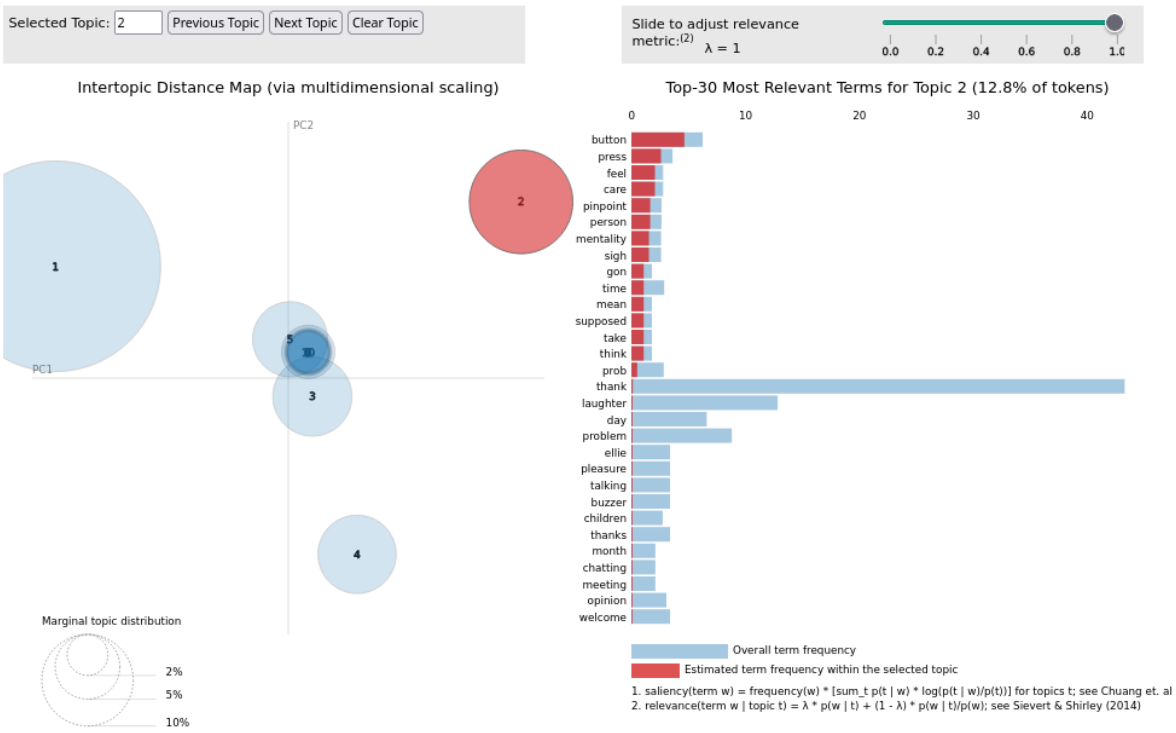

Figure 4.1: LDA representation with 10 topics.

it is possible to see that Topic 1 is the most extensive topic, by covering 53% of the tokens. Besides, Topic 2 covers 12.8%, Topic 3 to 5 cover approximately 6.7% to 7.5% each, leaving the remaining tokens to the remaining five topics. This LDA model was evaluated and presented 0.782 Coherence and -5.08 Perplexity. Perplexity is a measure of predictive likelihood, used for language model evaluation. Optimization for this measure would be inappropriate, because it would produce human non-interpretable topics, unlike coherence, which measures the score of a topic by evaluating the degree of semantic similarity between high scoring terms in the topic. Table 4.1 presents the most frequent words per topic, with the terms being ordered by weight in descending order.

[1] https://pyldavis.readthedocs.io/en/latest/readme.html

Topics	Relevant Words
Topic 1	*thank, laughter, day, problem, people, welcome, thanks, buzzer, talking, pleasure*
Topic 2	*button, press, feel, care, pinpoint, person, mentality, sigh, time, gon*
Topic 3	*laugh, laughter, things, time, improve, themself, help, thank, problem, day*
Topic 4	*fun, others, opinion, store, kind, gadgets, play, kid, giving, candy*
Topic 5	*problem, see, prob, thank, laughter, day, pleasure, button, ellie, children*
Topic 6	*thank, laughter, day, problem, button, ellie, pleasure, talking, buzzer, thanks*
Topic 7	*thank, laughter, day, problem, button, ellie, pleasure, talking, buzzer, children*
Topic 8	*thank, laughter, day, problem, button, pleasure, talking, buzzer, ellie, children*
Topic 9	*thank, laughter, problem, day, button, pleasure, talking, buzzer, ellie, hon*
Topic 10	*thank, laughter, day, problem, month, see, pleasure, children, button, buzzer*

Table 4.1: Most relevant words per topic.

As it is possible to see, there is a huge overlap of the most relevant terms from Topic 6 to Topic 10, which do not provide enough information. The first three topics and the 5^{th} topic are related to a clinical interview, as expected. Topic 1 is very general and neutral, by having positive (e.g. *pleasure*) and negative (e.g. *problem*) terms as the most relevant. Topic 2 seems to be feeling-oriented and more negative, with words like *sigh*, *feel*, *mentality*, etc.. Topic 3 seems to be slightly positive, with *improve* as a relevant term. Topic 4 is clearly more positive, since it is related to relaxing activities (e.g. *fun*, *play*, *kind*). Topic 5 is also neutral, just like Topic 1, with both positive (e.g. *laughter*, *pleasure*) and negative (e.g. *problem*) terms.

4.2.2 Results with Development subset

The results from the Development subset are presented from Table 4.2 to Table 4.4. Table 4.2 has the results from the textual model, Table 4.3 the results from the audio model and Table 4.4 the results from the fusion model.

Results Text	RMSE	MAE	F1
LR-SGD	**6.18**	5.25	0.78
SVR	6.60	5.22	0.79
RF	6.54	5.42	0.77
AdaBoost	6.38	**4.97**	0.79
GBDT	6.47	5.17	0.79
MLP	7.34	6.22	0.70

Table 4.2: Results from text model on dev subset.

Results Audio	RMSE	MAE	F1
LR-SGD	6.36	**4.82**	0.79
SVR	6.34	5.17	0.79
RF	6.34	5.17	0.78
AdaBoost	**6.04**	5.34	0.75
GBDT	6.15	4.91	0.78
MLP	6.54	5.14	0.78

Table 4.3: Results from audio model on dev subset.

Results Fusion	RMSE	MAE	F1
LR-SGD	6.26	5.02	0.79
SVR	6.54	5.31	0.79
RF	6.34	5.17	0.78
AdaBoost	6.39	5.14	0.79
GBDT	**6.15**	**4.91**	0.78
MLP	6.61	5.22	0.77

Table 4.4: Results from fusion model on dev subset.

After analysing Table 4.2, the algorithm with the best result in terms of RMSE was LR-SGD, while AdaBoost had the best result in MAE. Regarding F1, the results are in general similar between algorithms. As for the Table 4.3, Adaboost performed better in terms of RMSE while LR-SGD had the best result on MAE. In Table 4.4, GBDT outperformed the other algorithms in both regression metrics. It is noteworthy that RF and GBDT had the same results for the audio model and for the fusion model because feature selection discarded the textual features. This happened because as mentioned in 4.1, most of the extracted textual features are not discriminant enough to estimate depression.

4.2.3 Results with Test subset

The results from the test subset are presented from Table 4.5 to Table 4.7, similarly as in 4.2.2.

Results Text	RMSE	MAE	F1
LR-SGD	6.95	5.67	0.80
SVR	**6.48**	**5.34**	0.83
RF	7.44	6.00	0.82
AdaBoost	6.75	5.71	0.83
GBDT	6.71	5.76	0.83
MLP	6.95	5.91	0.79

Table 4.5: Results from text model on test subset.

After analysing the results from the test subset, it is possible to say that the results are worse than expected, when comparing with the ones obtained in the development subset. SVR outperformed the remaining algorithms in every modality and in every metric. One possible explanation is that the features extracted from the test subset are too different when compared with the ones from the train subset, implying that there's a lack of discriminative

Results Audio	RMSE	MAE	F1
LR-SGD	6.98	5.82	0.80
SVR	**6.56**	**5.36**	**0.83**
RF	6.80	5.76	0.81
AdaBoost	6.80	6.06	0.78
GBDT	6.73	5.47	0.82
MLP	6.93	5.63	0.81

Table 4.6: Results from audio model on test subset.

Results Fusion	RMSE	MAE	F1
LR-SGD	6.79	5.43	0.82
SVR	**6.64**	**5.34**	**0.83**
RF	6.80	5.76	0.81
AdaBoost	6.99	5.80	0.81
GBDT	6.73	5.47	0.82
MLP	7.00	5.50	0.81

Table 4.7: Results from fusion model on test subset.

features. One example of this was presented in 4.1, when analysing the textual features. Also, the fusion model was affected by this issue, which resulted on worse results than the separated models. Other possibility is a slight overfit of the models.

4.2.4 Comparison with AVEC baseline and articles

In Table 4.8 and 4.9, the baseline from AVEC'17 and the best results from some of the papers presented in Section 2.1.2 are presented with the best performing model for the development subset and the best performing model for the test subset.

Dev	RMSE	MAE
AVEC-Audio	6.74	5.36
AVEC-Video	7.13	5.88
AVEC-Audio+Video	6.62	5.52
Fusion-GBDT	6.15	4.91

Table 4.8: Baseline AVEC for dev subset and best performing model.

After analysing the AVEC baseline results, it is clear that the developed models still outperformed them, even though they had worse results on the test subset. When analysing the results from some of the articles presented in Section 2.1.2 for the test subset, the results from the best performing model (Text-SVR) are in some way comparable with them. It is worth mentioning that Zheng et al. did produce the best results with their model for the DAIC dataset.

Test	RMSE	MAE
AVEC-Audio	7.78	5.72
AVEC-Video	6.97	6.12
AVEC-Audio+Video	7.05	5.66
Yang et al.[12]	9.10	6.70
Yang et al.[14]	5.97	5.16
Zheng et al.[18]	**3.28**	**2.62**
Alhanai et al.[24]	6.27	4.97
Text-SVR	6.48	5.34

Table 4.9: Baseline AVEC, State-of-the-Art models results and best performing model(Text-SVR) for test subset.

4.3 Summary

In this chapter, an analysis of the textual and audio features was presented. It was demonstrated that the textual features were not discriminative enough, with some of them providing small variance between depressed and non-depressed participants. Audio features like the MFCC have proven that are reliable features to identify depression. The topic modelling with LDA revealed interesting results, by finding various topics. It was possible to distinguish negative topics from positive topics. The results from the models were presented for development and test subsets, with worse results on the latter. Nonetheless, the results were still better than the baseline and are not too far off of some State-of-the-Art models, as presented in Table 4.9.

CHAPTER 5

Conclusions

In this work, a general overview of what depression is and some statistics of this disorder were presented. After a study on the state-of-the-art and on some specific audio and video concepts, two experiments were developed. The first one was the isolated text and audio models and the second one was a fusion model with text and audio features. For each model, multiple types of algorithms were experimented, ranging from linear regression to artificial neural networks. Also, a topic modelling analysis with LDA was conducted to evaluate if it could distinguish positive and negative topics. After an analysis on the extracted features, it was noted that the textual features did not meet with the expectations. Since the textual features were not discriminative enough, the fusion model did not present improvements when comparing with the isolated models. Nevertheless, the achieved results from the experiments outperformed the baseline defined by AVEC and are comparable with other results obtained in state-of-the-art models. The topic modelling experiment revealed positive results, by being able to identify negative and positive topics. As future work, an interesting approach would be to explore Deep Learning models, such as LSTM, in order to evaluate if it could provide better results than the ones presented in Chapter 4. Other approach would be to explore and extract more discriminative features, especially text-related. One example of this could be integrating the performed topic modelling experience into it. Another example could rely on analysing the context of the occurrences of non-verbal behaviour (e.g. participant sniffles when asked about family). Including video features in the fusion model would also be relevant to compare with state-of-the-art models.

References

[1] Stefan G. Hofmann, Mark A. Reinecke, David J. A. Dozois and Peter J. Bieling, *Cognitive-behavioral Therapy with Adults; A Guide to Empirically-informed Assessment and Intervention.* Cambridge University Press, 2010, ch. 1.

[2] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 4th. APA, 2000.

[3] GBD 2017 Disease and Injury Incidence and Prevalence Collaborators, «Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017.», *Global Health Metrics, Volume 392, Issue 10159*, 2018.

[4] Kessler RC, Chiu WT, Demler O, Walters EE., *Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication.* 2005.

[5] (2021). «7.2% of people in the EU suffer from chronic depression», Eurostat, [Online]. Available: `https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20210910-1`.

[6] De Vault et al., «SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support», *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France*, 2014.

[7] Alina Trifan, Pedro Salgado and José Luís Oliveira, «BioInfo@UAVR at eRisk 2020: on the use of psycholinguistics features and machine learning for the classification and quantification of mental diseases», *CLEF eRisk 2020*, 2020.

[8] Gratch et al., «The Distress Analysis Interview Corpus of human and computer interviews», *Proceedings of Language Resources and Evaluation Conference (LREC)*, 2014.

[9] M. Lhommet and S. C. Marsella, «Gesture with meaning», *Intelligent Virtual Agents*, 2013.

[10] S. Marsella et al., «Virtual character performance from speech», *Proceedings of the Symposium on Computer Animation, Anaheim*, 2013.

[11] Genevieve Lam, Huang Dongyan and Weisi Lin, «Context-Aware Deep Learning for Multi-Modal Depression Detection», *International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2019*, 2019.

[12] Le Yang et al., «Decision Tree based Depression Classification from Audio Video and Language Information», *International Workshop on Audio/Visual Emotion Challenge (AVEC'16)*, 2016.

[13] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, «The PHQ-8 as a measure of current depression in the general population», *Journal of Affective Disorders*, vol. 114, no. 1, pp. 163–173, 2009, ISSN: 0165-0327. DOI: `https://doi.org/10.1016/j.jad.2008.06.026`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0165032708002826`.

[14] Le Yang et al., «Multimodal Measurement of Depression Using Deep Learning Models», *International Workshop on Audio/Visual Emotion Challenge (AVEC'17)*, 2017.

[15] Q. Le and T. Mikolov., «Distributed representations of sentences and documents», *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1188–1196, 2014.

[16] (). «openSMILE», [Online]. Available: `https://audeering.github.io/opensmile/index.html`.

[17] Anastasia Pampouchidou et al., «Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text», *International Workshop on Audio/Visual Emotion Challenge (AVEC'16)*, 2016.

[18] Wenbo Zheng et al., «Graph Attention Model Embedded with Multi-Modal Knowledge for Depression Detection», *IEEE International Conference on Multimedia and Expo (ICME)*, 2020.

[19] Edward Choi et al., «Gram: Graph-based attention model for healthcare representation learning», *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 787–795, 2017.

[20] L. Yang, «Multi-Modal Depression Detection and Estimation», *8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2019.

[21] Anupama Ray et al., «Multi-level Attention Network using Text, Audio and Video for Depression Prediction», *International Workshop on Audio/Visual Emotion Challenge (AVEC'19)*, 2019.

[22] Daniel Cer et al., «Universal Sentence Encoder», *arXiv:1803.11175*, 2018.

[23] Florian Eyben et al., «The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing», *IEEE Transactions on Affective Computing*, 2015.

[24] Tuka Alhanai, Mohammad Ghassemi and James Glass, «Detecting Depression with Audio/Text Sequence Modeling of Interviews», *Interspeech 18*, 2018.

[25] Cenk Demiroglu et al., «Depression-level assessment from multi-lingual conversational speech data using acoustic and text features», *EURASIP Journal on Audio, Speech, and Music Processing*, 2020.

[26] M. Valstar et al., «Avec 2014: 3D dimensional affect and depression recognition challenge», *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (AVEC'14)*, pp. 3–10, 2014.

[27] David E. Losada, Fabio Crestani, Javier Parapar, «Overview of eRisk at CLEF 2020: Early Risk Prediction on the Internet (Extended Overview)», *eRisk CLEF 20*, 2020.

[28] Aaron T. Beck, R. A. Steer, G. K. Brown, *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation, 1996.

[29] Ana-Sabina Uban and Paolo Rosso, «Deep learning architectures and strategies for early detection of self-harm and depression level prediction», *eRisk CLEF 20*, 2020.

[30] S. M. Mohammad and P. D. Turney, «NRC Emotion Lexicon», NRC, Tech. Rep., Dec. 2013.

[31] Rodrigo Martínez-Castaño et al., «Early Risk Detection of Self-Harm and Depression Severity using BERT-based Transformers», *eRisk CLEF 20*, 2020.

[32] Diego Maupomé et al., «Early Mental Health Risk Assessment through Writing Styles, Topics and Neural Models», *eRisk CLEF 20*, 2020.

[33] Amina Madani et al., «Deep learning models to measure the Severity of the Signs of Depression using Reddit Posts», *eRisk CLEF 20*, 2020.

[34] Jana M. Havigerová et al., «Text-based Detection of the Risk of Depression», *Frontiers in Psychology*, 2019.

[35] Anas Belouali et al., «Acoustic and language analysis of speech for suicide ideation among US veterans», 2020. DOI: 10.1101/2020.07.08.20147504.

[36] G. Degottex et al., «COVAREP - A collaborative voice analysis repository for speech technologies», in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[37] P. F. Felzenszwalb et al., «Object Detection with Discriminative Trained Part Based Models», *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2010.

[38] T. Baltrušaitis, P. Robinson, and L.-P. Morency, «OpenFace: an open source facial behavior analysis toolkit», in *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[39] T. Baltrušaitis, L.-P. Morency and P. Robinson, «Constrained local neural fields for robust facial landmark detection in the wild», 2013.

[40] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, «Array programming with NumPy», *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. DOI: 10.1038/s41586-020-2649-2. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2.

[41] M. Honnibal, I. Montani, S. V. Landeghem, and A. Boyd, «spaCy: Industrial-strength Natural Language Processing in Python», 2020. DOI: 10.5281/zenodo.1212303.

[42] Bird, Steven, Edward Loper and Ewan Klein, *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[43] *librosa/librosa: 0.8.0*, version 0.8.0, Jul. 2020. DOI: 10.5281/zenodo.3955228. [Online]. Available: https://doi.org/10.5281/zenodo.3955228.

[44] C.J. Hutto and E. E. Gilbert, «VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text», *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, 2014.

[45] R. Řehůřek and P. Sojka, «Software Framework for Topic Modelling with Large Corpora», English, in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta: ELRA, May 2010, pp. 45–50.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, «Scikit-learn: Machine Learning in Python», *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[47] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, «Transformers: State-of-the-Art Natural Language Processing», in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[48] (). «LIWC2015», Pennebaker Conglomerates, Inc., [Online]. Available: https://liwc.wpengine.com/.

[49] H. Drucker, «Improving Regressors using Boosting Techniques», *Proceedings of the 14th International Conference on Machine Learning*, 1997.

[50] Shuller et al., «AVEC 2017- Real-life Depression, and Affect Recognition Workshop and Challenge», *International Workshop on Audio/Visual Emotion Challenge (AVEC'17)*, 2017.