In-the-wild detection of speech affecting diseases

Maria Joana Ribeiro Folgado Correia

CMU-LTI-21-009

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:

Bhiksha Raj (Co-chair), Carnegie Mellon University
Isabel Trancoso (Co-chair), Instituto Superior Técnico, University of Lisbon
Tanja Schultz, University of Bremen
António Teixeira, University of Aveiro
Rita Singh, Carnegie Mellon University
Alberto Abad, Instituto Superior Técnico, University of Lisbon

Submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
In Language and Information Technologies

In-the-wild detection of speech affecting diseases

Maria Joana Ribeiro Folgado Correia

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, USA

Instituto Superior Técnico
University of Lisbon
INESC-ID
Lisbon, Portugal

Professor Bhiksha Raj (Co-advisor), Carnegie Mellon University
Professor Isabel Trancoso (Co-advisor), Instituto Superior Técnico, University of Lisbon
Professor Tanja Schultz, University of Bremen
Professor António Teixeira, University of Aveiro
Professor Rita Singh, Carnegie Mellon University
Professor Alberto Abad, Instituto Superior Técnico, University of Lisbon

Submitted in partial fulfillment of the requirements for the degrees of
Doctor of Philosophy in Language and Information Technologies and
Doctor of Philosophy in Electrical and Computer Engineering

June 2021

to my grandparents
para os meus avós

Acknowledgements

I would like to begin by acknowledging my outstanding advisors, Professor Isabel Trancoso, and Professor Bhiksha Raj. I have had the immense privilege of working under them, as a doctoral student, for the past six years. These have been years of intense growth, professionally as a researcher, and also personally, so I owe Professor Isabel Trancoso and Professor Bhiksha Raj a great dept of gratitude for their teachings, for guiding me, believing in my work, and supporting me on a personal level as well, during this period of time. For the great amount of time and energy that they have spent discussing research ideas, and providing technical insight, I will be forever grateful to them.

I would also like to acknowledge Professor Tanja Shultz, and Professor Rita Singh for their many insightful discussions, particularly since the thesis proposal.

Furthermore, I would like to acknowledge all the Professors and colleagues that I had the pleasure to work alongside, and collaborate with, at the MLSP group, at CMU, and the HLT group, at IST. It has been an invaluable experience to share ideas with such a bright, and diverse group of researchers and friends. I am grateful and indebted to each of them for what they have taught me. A special thank you to my colleagues who have also been my co-authors. In particular, to Francisco Teixeira, and Catarina Botelho for their thoughtful review of this thesis, and to Professor Alberto Abad for his guidance and insight since my years as a master student.

It is also important to acknowledge the Portuguese Foundation for Science and Technology (FCT), for partially funding this PhD; and the CMU Portugal program, for creating such a unique program for young researchers. The experience of learning and working at two leading research institutes has been an invaluable one.

On a personal note, I would like to acknowledge my friends. An enormous thank you to Bill, Luís, Miguel, Joli, Shenglan, António, Eliezer, Telmo, Paul, Anuva, Carlos, Leonor, Maria Cremilde, and Figueiredo for their support through good and bad.

Finally, my family, who has loved and supported me unconditionally through my life, which has allowed me to stand here today. To them, I say: whatever accomplishments I achieve in life, they are yours. My deepest gratitude and love goes to my mother Maria João, brother João, grandparents Maria das Dores, Maria Filomena, João, and José, godmother Maria, mother-in-law Edna, grandmother-in-law Lourdes, brother-in-law Rui, stepfather-in-law Amadeu, cousins Raquel and Matthew, and to my life partner João.

Abstract

Speech is a complex bio-signal that is intrinsically related to human physiology and cognition. It has the potential to provide a rich biomarker for health, allowing a non-invasive route to early diagnosis and monitoring of a range of conditions that affect speech. The scientific community has shown consistent interest in automating the diagnosis and monitoring of speech affecting diseases, but advances in this area have been limited by the small size of the available speech medical corpora, as these can be prohibitively difficult and expensive to collect.

At the same time, the problem of diagnosing and monitoring speech affecting diseases specifically in in-the-wild contexts has been neglected, as the few existing speech medical corpora only contain recordings made in controlled conditions. These are typically conditions in which the channel is known, the background noise is minimized, or the content of the recordings is controlled by either speaking exercises or clinical interviews. They do not provide a good representation of real life scenarios.

In this thesis we address the problem of detecting SA in in-the-wild contexts by, on one hand proposing novel strategies to collect and annotate speech medical corpora of arbitrary size, for arbitrary speech affecting (SA) diseases, from pre-existing massive online multimedia repositories. On the other hand, by proposing novel strategies to detect speech affecting diseases in both controlled and in-the-wild conditions, thus expanding the scenarios in which the detection of such diseases is possible.

At the same time, we perform the first study of the limitations of both the existing speech medical corpora and current speech affecting disease detecting techniques when faced with in-the-wild data.

In the scope of this thesis we also collect and annotate the in-the-wild speech medical (WSM) corpus, a first of its kind, ever growing corpus of in-the-wild multimodal recordings, featuring examples of several speech affecting diseases, including depression and Parkinson's disease.

Resumo

A fala é um bio-sinal complexo, que está intrinsecamente ligado à fisiologia e cognição humana. Tem o potencial de ser um bio-marcador importante para determinar o estado de saúde, permitindo o diagnóstico precoce e a monitorização de um leque do doenças que afectam a fala. A comunidade científica tem consistentemente mostrado interesse na automatização do diagnóstico médico e monitorização de doenças que afectam a fala, mas o progresso desta área têm sido travado pelo tamanho limitado das *corpora* de oradores com doenças que afectam a fala, uma vez que o custo e a dificuldade de recolha e anotação das mesmas tende a ser proibitivo.

Em simultâneo, o problema de diagnosticar e monitorizar doenças que afectam a fala, particularmente em contextos *in-the-wild*, tem sido negligenciado, uma vez que as *corpora* existentes apenas contêm exemplos recolhidos em condições controladas. Estas condições são, tipicamente, aquelas em que o canal é conhecido, o ruído de fundo é minimizado, e também nas quais o conteúdo dos exemplos presentes no *corpus* é determinado por exercícios de fala, ou guiado por entrevistas clínicas. Estas condições não consistem numa representação fiel das condições encontradas na vida real.

Nesta tese procuramos endereçar o problema da detecção de doenças que afectam a fala, particularmente em contextos *in-the-wild*. Por um lado, propomo-nos a fazê-lo através do desenvolvimento de novas estratégias de recolha e anotação automática de corpus de oradores com doenças que afectam a fala, para doenças que afectam a fala arbitrárias, a partir de repositórios massivos e multimodais já existentes. Por outro lado, desenvolvendo novas estratégias de detecção automática de doenças que afectam a fala, tanto em condições controladas, como em condições *in-the-wild*, alargando assim o leque de cenários em que estas podem ser automaticamente detectadas.

Também realizamos aquele é, tanto quanto sabemos, o primeiro estudo que mede as limitações tanto das *corpora* actualmente existentes, como das técnicas actuais de detecção de doenças que afectam a fala, quando confrontadas com exemplos provenientes de condições *in-the-wild*.

No âmbito desta tese, também recolhemos e anotamos o *in-the-wild speech medical (WSM) corpus*. Um corpus com características únicas, em permanente crescimento, com exemplos de vídeos *in-the-wild*, e que contempla várias doenças que afectam a fala, incluindo depressão e doença de Parkinson.

Contents

List of Tables

List of Figures

Acronyms

AMT Amazon Mechanical Turk.

AVEC Audio/Visual Emotion Challenge and Workshop.

BERT Bidirectional Encoder Representations from Transformers.

BoAW Bag-of-AudioWords.

BoW Bag-of-Word.

CC Controled Conditions.

CNN Convolutional Neural Network.

ComParE Computational Paralinguistics challengE.

DAIC Distress Analysis Interview Corpus.

DAIC-F2F Distress Analysis Interview Corpus - Face to Face.

DAIC-WOZ Distress Analysis Interview Corpus - Wizard of Oz.

DDK diadochokinetic.

DL Deep Learning.

DNN Deep Neural Networks.

EER Equal Error Rate.

eGeMAPs extended Geneva minimalistic acoustic parameters.

EM Expectation Maximization.

F2F Face-to-face.

GeMAPs Geneva minimalistic acoustic parameters.

GMM Gaussian Mixture Model.

HC Healthy Controls.

JFA Joint Factor Analisys.

LDA Linear Discriminant Analysis.

LLD Low-Level Descriptor.

LLR Log-Likelihood Ratio.

LR Logistic Regression.

MAE Mean Absolute Error.

MFCC Mel-Frequency Cepstral Coefficients.

mi/MI-SVM Multiple Instance Support Vector Machine.

MIL Multiple-Instance Learning.

MLP Multi-Layer Perceptron.

NN Neural Network.

OSA Obstructive Sleep Apnea.

PC-GITA Parkinson's disease Corpus from the Applied Telecommunications Group.

PCC Pearson correlation coefficient.

PD Parkinson's Disease.

PHQ-9 Patient Health Questionnaire.

PLDA Probabilistic Linear Discriminant Analysis.

PTSD Post-Traumatic Stress Disorder.

RBF Radial Basis Function.

RMSE Root Mean Squared Error.

RNN Recurrent Neural Networks.

RNTN Recursive Neural Tensor Network.

SA Speech Affecting [Disease].

SBERT Sentence Bidirectional Encoder Representations from Transformers.

SVM Support Vector Machine.

td-idf term-frequency times inverse document-frequency.

TD-NN Time-Delay Neural Network.

UAR Unweighted Average Recall.

UBM Universal Background Model.

VAD Voice Activity Detection.

WOZ Wizard-of-Oz.

WSM in-the-Wild Speech Medical [Corpus].

Part I

Introduction

Chapter 1

Thesis Overview

1.1 Motivation

Speech is a complex bio-signal that is intrinsically related to human physiology and cognition. It has the potential to provide a rich bio-marker for health, allowing a non-invasive route to early diagnosis and monitoring of a range of conditions that affect speech, including several mood disorders (depression, bipolar disorder, anxiety, *etc.*), several degenerative diseases (including Parkinson's disease, Alzheimer's disease, Huntington's disease, amyotrophic lateral sclerosis, among others), sleep related conditions (such as sleep apnea), some forms of autism, and diseases of the respiratory system (such as the asthma, COVID-19, or influenza).

The scientific community has done extensive work, and has shown consistent interest in automating the diagnosis and monitoring of such diseases, to which we will refer to as speech affecting (SA) diseases, using a plethora of approaches, not necessarily based on speech analysis. ML-based tools for diagnosis range from those that analyse medical images, including magnetic resonance imaging [3] [4] [5], to electroencephalograms [6] [7], to videos, including eye-tracking [8] [9], and motion tracking [10], among others. All of them have their own advantages and disadvantages, namely in terms of four parameters: cost, invasiveness, accessibility, and performance.

The advantages of performing automatic diagnosis based specifically on speech, over other techniques, include the following: non-invasiveness, and easy availability, both of which because the only necessary material for the diagnosis is an external microphone.

However, cost can become an issue, in fact, it is one of the most important factors that is limiting the progress towards creating robust and accurate automatic speech based diagnosis

technologies. In this context, cost is related both to human and financial resources associated to collecting, and labeling speech medical data. The usual setup to collect any given speech medical dataset involves finding eligible and willing subjects, assigning healthcare specialists, and ensuring the technical, logistic and legal requirements for the data collection process. After that, it is necessary to have a team of specialists process the raw collected data and annotate it manually. As a consequence, the existing speech medical datasets are few in number, and small in size.

In turn, because of the limited size of existing speech medical datasets, any models or techniques developed using them are limited in complexity, which can be translated into limitations on the performance that can be achieved using this data.

Furthermore, currently existing speech medical data are collected in controlled conditions (CC), which corresponds to one or several of the following criteria: patients have a script or guidelines for what to say, as determined by specific speech exercises or via clinical interviews; the channel conditions are known; the noise conditions are controlled or minimized. These conditions may, at first glance, seem the most desirable conditions to collect the speech medical data in, given that the constraints under which the data were collected make the problem easier to solve: *e.g.* specific speaking exercises are designed to make isolated aspects of articulation, phonation or prosody, that are characteristic of a given SA disease stand out, when compared to spontaneous speech; or clinical interviews may guide the subject to an emotional state that is characteristic of a given SA disease (typically only applicable to mood disorders, and diseases from the autism spectrum), but which the subject would not spontaneously demonstrate.

In contrast, real-life applications for detecting SA diseases should operate in vastly different conditions, where the subjects are not constrained in terms of what they say or how they say it, and where, at the same time, there is no knowledge about the channel and background noise, *i.e.* in-the-wild conditions.

While detecting SA diseases in in-the-wild conditions is a more difficult problem than the equivalent task in CC, it is arguably a more relevant one. This is because the former has the potential to create a more realistic characterization of SA diseases, as well as be applicable in more scenarios, and be made available more broadly, beyond what is possible in CC. However, the problem of detecting SA diseases in in-the-wild conditions, is yet to be addressed.

We believe that by improving the detection of SA diseases, particularly in in-the-wild conditions, we are making a small contribution towards democratizing the access to healthcare worldwide.

1.2 Thesis Statement

With the previous motivation in mind, the goal of this thesis is to address the limitations in the state-of-the-art in the detection of SA diseases based on speech, both in terms of lack of data that faithfully represents real life scenarios, as well as techniques that automate the detection of such diseases in any conditions, both CC and in-the-wild.

In essence, the main goal of this thesis is to:

Push the state-of-the-art of automatic detection of SA diseases based on speech, by proposing a set of tools that would ultimately allow the detection of any SA disease, in real-life scenarios.

Along the way, we hope to answer the following research questions:

- How does the problem of detecting SA diseases in CC differ from the same problem in in-the wild conditions?
- Do speech medical data collected from existing in-the-wild sources provide a good representation of non-healthy speech?
- Can speech medical data collected from existing in-the-wild sources be an effective resource for training and evaluating SA detectors?

1.3 Contributions

To the best of our knowledge, this is the first work to address the in-the-wild detection of SA diseases, from the data collection and labeling stage, to the diagnosis. Over the course of this thesis, we will focus on two SA diseases as our working examples: depression, and Parkinson's disease (PD). Nevertheless, it is not the goal of this thesis to present solutions that are optimized for these two SA diseases. Rather, our aim was to develop solutions that remained "disease agnostic", *i.e.*, that do not leverage from domain specific knowledge about a target SA disease, and that, therefore, can be easily reused for any SA disease, not only depression and PD.

This thesis' main contributions are:

1. The in-the-wild speech medical (WSM) corpus, an ever growing, first of its kind corpus that features in-the-wild recordings of subjects affected by several SA diseases;
2. The development of novel frameworks that automate the collection and annotation process of these in-the-wild datasets of arbitrary size, that are, at the same time, easily

translatable to other tasks;

3. A study comparing and measuring the differences between the tasks of detecting SA diseases in CC and in-the-wild conditions;

4. The development of frameworks to tackle the in-the-wild detection of SA diseases, while remaining agnostic to the target SA disease.

Finally, the work presented on this thesis has resulted in the following peer reviewed publications:

- J. Correia, I. Trancoso, and B. Raj, "*Detecting psychological distress in adults through transcriptions of clinical interviews*," in IberSPEECH 2016, Lisbon, Portugal, November 2016

- J. Correia, I. Trancoso, and B. Raj, "*Adaptation of SVM for MIL for inferring the polarity of movies and movie reviews*," in 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, USA, December 2016

- J. Correia, I. Trancoso, B. Raj, and F. Teixeira, "*Mining multimodal repositories for speech affecting diseases*," in 19th Annual Conference of the International Speech Communication Association (INTERSPEECH), Hyderabad, India, September 2018

- J. Correia, I. Trancoso, and B. Raj, "*Querying depression vlogs*,", in IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, December 2018

- J. Correia, I. Trancoso, and B. Raj, "*End-to-end in-the-wild detection of speech affecting diseases*," in IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) 2019, Sentosa, Singapore, December 2019

- J. Correia, I. Trancoso, and B. Raj, "*Automatic in-the-wild dataset annotation with deep generalized multiple instance learning*," in 12^{th} International Conference on Language Resources and Evaluation (LREC), Nice, France, May 2020

- J. Correia, C. Botelho, F. Teixeira, I. Trancoso, and B. Raj, *The In-the-Wild Speech Medical Corpus*,", in 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada, June 2021

This work has also been included in several keynote papers.

1.4 Thesis Organization

This thesis is organized in four parts. Part I is dedicated to introduce, motivate and contextualize the problems addressed in this thesis. Specifically, within Part I, Chapter 2 provides an overview of the characteristics of depression and PD, and the physiological changes that are associated to these diseases, particularly regarding speech production, thus providing the necessary context to understand how it is possible to detect them through speech. This Chapter also provides a summary of the existing speech medical corpora for these two target SA diseases. Chapter 3 is dedicated to reviewing the state-of-the-art of the detection of depression and PD. However this review is limited to experiments performed in CC conditions, which are the only conditions that have yet been studied. Additionally, this Chapter also reviews some of the work previously developed in the scope of automating the annotation of corpora.

Part II lays out our proposed strategies to automate the process of collecting and annotating speech medical corpora for arbitrary SA diseases, and provides examples for depression and PD. Chapter 4 describes the WSM Corpus, a corpus of in-the-wild recordings of subjects affected by several SA diseases, collected in the scope of this thesis. This Chapter also provides some details regarding different versions of the corpus, from proof-of-concept to final, along with their collection methodology. Chapter 5 proposes several techniques to automate the annotation of corpora, with varying ratios of labeled to unlabeled data during training. In this Chapter we perform several experiments where we apply the proposed techniques to annotate the WSM Corpus.

Part III covers several techniques to detect SA diseases both in CC and in-the-wild conditions, both from a theoretical as well as experimental perspectives. Chapter 6 describes several proposed approaches to detect SA diseases, by adopting different strategies do formulate the problem, from knowledge based approaches, to approaches based on speaker modeling and end-to-end deep learning (DL). This Chapter also features the experimental verification of these techniques, using, among others, the WSM Corpus.

Finally, Part IV is dedicated to reflect on the work that has been accomplished over the course of this thesis, drawing some conclusions, which can be found in Chapter 7, as well as laying out suggestions for future work, in Chapter 8.

Additionally to the main body of this thesis, we include three Appendices. Two contain ideas, experiments, and results, that, while not directly in the scope of this thesis, provide additional insight to problems adjacent to the ones presented and addressed in this thesis. Appendix A explores a novel strategy to detect depression, anxiety and post-traumatic stress

disorder (PTSD) based on a quantity deemed the "connotation" of a word. Appendix B provides an experimental verification to the problem of dataset annotation, using one of the techniques proposed in Chapter 5, applied to the domain of written movie reviews. Finally, the third, Appendix C, provides the supporting documentation for the distribution of the WSM Corpus for academic, research, and other non-commercial purposes.

Chapter 2

Speech affecting diseases: depression and Parkinson's disease

Before addressing the main problem of this thesis, which is related to the detection of SA diseases, it is important to have a basic understanding of what the target SA diseases are, and what symptoms are typically associated with them, particularly regarding speech production. We will dedicate this Chapter to covering these aspects, specifically for the SA diseases which were chosen to illustrate our thesis statement: depression and PD.

In Section 2.1 we review the main acoustic changes that these impairments impose on otherwise healthy, or "normal" speech. Then, in Section 2.2, we describe the most commonly used speech medical datasets to automatically train models that detect or assess the severity of depression and PD, as a way to provide context to the resources typically available in this research area.

2.1 Speech affecting diseases

2.1.1 Depression

Depression, otherwise known as major depressive disorder, or clinical depression, is a common and serious mood disorder, characterized by persistent feelings of sadness and hopelessness, as well as loss of interest or pleasure in activities previously enjoyed. It has a lifetime prevalence of over 16% [11], and is considered the world's fourth most serious health threat [12]. At the same time, it is the leading cause of disability worldwide in terms of total years lost due to disability [13]. An estimated 350 million people of all ages are affected by this

disorder, worldwide [13], and it is expected to become more prevalent as the average age of the worldwide population increases [12].

It is estimated that the total cost of depression per year in the European Union is €92 billion, out of which €54 billion is the amount lost due to lost work productivity [14]. Similar estimates were made in the United States, where the cost of lost work productivity per year due to depression is estimated to be between $US 36 billion [15] and $US 53 billion [16]. In Australia the annual cost associated to absenteeism, presenteeism, turnover and treatment costs caused by depression is of $AUD 12.6 billion [17].

Aside from the above mentioned emotional symptoms, depression can also cause physical symptoms such as chronic pain or digestive issues.

According to the American Diagnostic and Statistical Manual of Mental Disorders, 5^{th} Edition (DSM-5) [18] (published by the American Psychiatric Association), the most widely used resource in the diagnosis of mental disorders, the diagnostic for depression must verify that the individual being diagnosed experiences five or more of the following symptoms during the same two week period, and at least one of the symptoms should be either depressed mood, or loss of interest and pleasure:

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

The criteria-based diagnosis of depression can also be accomplished with other instruments, including several self administered questionnaires, such as the Patient Health Questionnaire

(PHQ-9) [19], which scores each of the nine DSM-5 criteria as "0" (not at all) to "3" (nearly every day), thus being a tool that allows for a measure of severity of depression as well.

From a perspective of automating the diagnosis, it is important to understand which symptoms of depression can be quantified and measured. Typically, psychomotor disturbances in depression are a good candidate, since they account for a significant portion of the physical symptoms that occur in depression. Additionally, there is growing evidence that psychomotor disturbances are the earliest and most consistent indicators of mood disorders [20]. Psychomotor disturbances can be broadly classified into four subgroups of symptoms and signs based on three available clinical rating scales designed to characterize them (CORE [21], motor agitation and retardation scale [22], Widlöcher scale [23]): retardation, agitation, non-interactiveness, and mental slowing.

In turn, all of the above mentioned subgroups of psychomotor disturbances have an impact on speech production abilities, the acoustical properties of the speech of the depressed individual. These differences in acoustic properties between the speech of healthy and depressed individuals have been widely studied over the last few decades. Often, depressed speech is characterized as dull, monotone, monoloud, lifeless and metallic. These perceptual qualities can be associated with measurable acoustic properties such as the fundamental frequency (F_0), amplitude modulation (AM), formant structure, power distribution, pause frequency, pause duration, and jitter.

Specifically, several works comparing healthy and depressed speech have shown that depressed speech, in comparison to healthy speech, has:

- Lower mean F_0, as a paralinguistic marker of a person's underlying mood [24]
- Smaller range of formant frequencies, as a consequence of psychomotor retardation that leads to a tightening of the vocal tract [25]
- Reduced variation in loudness due to lack of speaking effort [26]
- Higher jitter and shimmer caused by issues in the spontaneous control of the glottal production mechanism [27] [28]
- Higher harmonic-to-noise ratio, caused by changes in the patterns of the air flow during speech production [29]

Diagnostic devices based on speech acoustics, which measure and quantify the above mentioned differences between the speech of healthy and depressed individuals, give the medical community useful tools to aid in the diagnosis and monitoring processes of depression.

2.1.2 Parkinson's disease

PD is a progressive multi-system neurodegenerative disease of the central nervous system, with multiple subgroups (including but not limited to multiple-system atrophy, and progressive supranuclear palsy), that causes partial or full loss in motor reflexes, speech, behaviour, mental processing, and other vital functions [30].

Its cause remains unknown, however, there is some evidence that the disease arises from an interaction between genetic and environmental factors that leads to progressive degeneration of neurons in susceptible regions of the brain.

PD is the second most common neurological problem in the elderly, after Alzheimer's disease [31]. The prevalence of PD in industrialised countries is generally estimated at 0.3% of the entire population, about 1% in people over 60 years of age, and about 4% in people over 80 years of age [31][32]. Therefore, as the average life expectancy increases, so will the impact of PD in future years. At the same time, the estimated yearly economic burden of PD in the US alone, including direct, indirect, and non-medical costs, is $US 52 billion [33]. A number that is predicted to increase to $US 79 billion by 2037 [33].

James Parkinson's original description of "the shaking palsy" in 1817 focused on the motor features of the disorder: tremor, bradykinesia, rigidity, micrographia, and different speech impairments [34][35]. Over time, a more complete picture of the clinical phenotype of PD has emerged, revealing it to be a multi-system disorder with a wide variety of motor and non-motor symptoms, with the non-motor symptoms being categorized into disturbances in autonomic function, sleep disturbances, cognitive and psychiatric disturbances, and sensory symptoms.

The first step for a PD diagnosis is to detect slowness of initiation of voluntary movements with progressive reduction in speed and amplitude of repetitive actions (bradykinesia) and one of the following additional symptoms: muscular rigidity, resting tremor or postural instability. Then, the diagnosis also has to ascertain at least three supportive criteria for PD, such as unilateral onset of symptoms, persistent asymmetry of clinical symptoms, good response to levodopa treatment, and induction of dyskinesias by the dopaminergic treatment.

In the present days, the diagnosis of PD is based on the criteria defined on the UK PD Brain bank [36], and the level and characteristics of motor impairments are currently evaluated according to the Movement Disorder Society – Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [37]. However, this scale only contains one item that is related to speech impairments.

An alternative scale to assess only the speech deficits of PD patients is the Frenchay Dysarthria Assessment (FDA), introduced in [38] and revised in [39], which was designed to assess dysarthria, a symptom that is frequently found in PD patients. The FDA scale includes several items to evaluate dysarthria such as reflexes, respiration, lips movement, palate movement, laryngeal capacity, tongue posture/movement, intelligibility, and others. This tool covers a wide range of aspects. However, it requires the patient to be with the examiner, which is not possible in many cases due to their reduced mobility.

As is the case of depression, studying the changes in speech produced by healthy individuals and individuals with PD can present a supplementary route to perform not just early diagnosis of PD, but also as a tool to monitor the progression of the disease.

The study of speech disturbances known to occur in individuals with PD is especially important as it occurs in about 90% of the cases, and affects the three principal "dimensions" of speech: phonation, articulation, and prosody [40]. They include:

- Dysarthria (difficulty in articulation) [41]
- Hypophonia (reduced loudness) [42]
- Hurried speech [43]
- Dysphagia (difficulty in swallowing) [44]
- Sialorrhoea (excessive salivation) and subsequent dribbling of saliva [45]

Dysarthria specifically, is also a common symptom in other neurological disorders such as bulbar palsy, pseudobulbar palsy, amyotrophic lateral sclerosis, cerebellar lesions, dystonia, and choreoathetosis. However, [46] have studied the perceptual characteristics of dysarthric speech in patients with seven different types of neurological disorders, including PD, and were able to conclude that dysarthria is manifested differently in patients with different neurological disorders.

The perceptual characteristics of dystarthric speech specifically caused by PD typically include reduced loudness, monopitch, monoloudness, reduced stress, breathy, hoarse voice quality, and imprecise articulation [46].

Later works focused on studying the specific phonetic changes and misarticulations of PD patients with dysarthria [47]. They showed that the phoneme classes that were most affected were stop-plosives, affricates, and fricatives. The former two were typically misarticulated as fricatives, and the latter ones were perceived as fricatives with reduced "sharpness". Further analysis of the articulatory deficits revealed inadequate tongue elevation to achieve complete

closure on stop-plosives and affricates; and inadequate close constriction of the airway, which cause misarticulations in lingual fricatives. Both phenomena represented inadequate narrowing of the vocal tract at the point of articulation.

Some of the effects of PD on the vocal tract have been observed through video stroboscopy, namely laryngeal tremor, vocal fold bowing, and abnormal glottal opening and closing [48].

Besides dysarthria, speech affected by PD is also characterized by variable rate and frequent word finding difficulties, referred to as "tip-of-the-tongue" phenomenon [49].

2.2 Existing Speech Medical Corpora

2.2.1 Depression: Distress Analysis Interview Corpus

The Distress Analysis Interview Corpus (DAIC) [50] is a multi-modal collection of semi-structured clinical interviews. It was designed to simulate the standard protocols created for identifying people at risk for depression, anxiety and post-traumatic stress disorder. The interviews were collected as part of a larger effort to create a computer agent that interviewed people and identified verbal and non-verbal indicators of mental illness [51].

The DAIC contains four types of interviews:

- Face-to-face (F2F) interviews between a participant and a human interviewer
- Teleconference interviews, conducted by a human interviewer over a teleconferencing system
- Wizard-of-Oz (WOZ) interviews, conducted by an animated virtual interviewed named Ellie, controlled by a human interviewer out of the participants sight
- Automated interviews, where participants are interviewed by Ellie, operating as a fully automated agent

Out of the four types of interviews, only one was made publicly accessible, the DAIC-WOZ, through the Audio/Visual Emotion Challenge and Workshop (AVEC 2016), and in the subsequent editions of this challenge.

The DAIC-WOZ contains 189 interviews, ranging from about 5 to 20 minutes. The participants were recorded by a camera, high-quality close-talking microphone, and Kinect. As such, the corpus contains audio, video, and depth sensor recordings of all the interactions. The interviews were automatically transcribed, and subsequently reviewed for accuracy by a senior transcriber. Utterances were segmented at boundaries with at least 300 milliseconds

Table 2.1: Summary of the DAIC-WOZ and DAIC-F2F in terms of number of interviews and labels.

Corpus	Partition	# Interviews	# Depressed	# PTSD	# Anxiety
DAIC-WOZ	Train	107	30	N/A	N/A
	Devel.	35	12	N/A	N/A
	Test	45	N/A	N/A	N/A
DAIC-F2F	N/A	65	26	29	43

of silence. All of them were subsequently anonymized (references to the patients' names, address or other such personal information was redacted).

Furthermore, the DAIC-F2F, although not publicly available, was kindly made available to us. It contains 65 videos recorded using the same setup as the DAIC-WOZ. However, the interviews in the DAIC-F2F are on average longer than the DAIC-WOZ ones, lasting from 30 to 60 minutes.

In terms of labels, the DAIC codes the participants for depression, PTSD and anxiety, based on accepted standard psychiatric questionnaires. Respectively, they were based on the Patient Health Questionnaire, depression module (PHQ-8) [19] (which is the same questionnaire as the PHQ-9 without the suicidal ideation questions), the PTSD Checklist – Civilian Version [52], State-Trait Anxiety Inventory [53]. The results of all the questionnaires were highly correlated, reflecting the co-occurrences found in these clinical conditions.

In Table 2.1 we present a summary of the DAIC-WOZ (divided in train, development and test partitions, as distributed by the authors), the DAIC-F2F, and their respective label distributions for depression, and, when available, for anxiety and PTSD. We note that the test labels of the DAIC-WOZ are not publicly available, as they are part of the running AVEC Challenge. Furthermore, the labels for PTSD and anxiety of the DAIC-WOZ have also never been publicly distributed.

2.2.2 Parkinson's disease: New Spanish Parkinson Corpus

The Parkinson's disease Corpus from the Applied Telecommunications Group (GITA) at the Universidad de Antioquia, Colombia [54] (PC-GITA), sometimes also referred to as the New Spanish Speech Corpus by some works of the corpus' authors, is a collection of speech recordings from 50 PD patients and 50 control subjects, 25 men and 25 women on each group, where the subjects perform a number of speech exercises.

In terms of age of the subsets in the corpus, the men with PD have ages ranging from 33 to

77 years old, with a mean of $62.2 \pm .2$; and the age of the women with PD ranges from 44 to 75 years old, with a mean of 60.1 ± 7.8. Among the healthy controls (HC), the age of the men ranges from 31 to 86, with a mean of 61.2 ± 11.3; and the age of the women ranges from 43 to 76 years old, with a mean of 60.7 ± 7.7. As such, this dataset is balanced both in terms of gender as well as age.

The dataset is in Colombian Spanish, and the recordings were captured in noise controlled conditions, in a sound proof booth that was built at the Clinica Noel, in Medellin, Colombia. All of the patients were diagnosed by neurology experts, and were labeled according to standard clinical protocols: the unified Parkinson's disease rating scale (UPDRS) [55], and Hoehn and Yahr (H&Y) [56]. These scales also provide a measure of the severity of PD.

The recording protocol considered different tasks which were designed to analyze several aspects of the voice and speech of people with PD. Such tasks were grouped into three aspects: phonation, articulation and prosody.

The evaluation of phonation was performed through the task of performing three repetitions of the five Spanish vowels uttered in a sustained manner; and the task of uttering the five Spanish vowels changing the tone of each vowel from low to high.

The evaluation of articulation included the following tasks: Three repetitions of the five Spanish vowels uttered in a sustained manner (same as in the phonation evaluation); the rapid repetition of the words and phonemes */pa-ta-ka/*, */pa-ka-ta/*, */pe-ta-ka/*, */pa/*, */ta/*, */ka/* (dysdiadochokinesia analysis, or DDK); and the repetition of a given list of Spanish words.

Finally, prosody was evaluated with the tasks of: repeating sentences with different levels of syntactic complexity; reading a pre-written dialog between a doctor and a patient; reading sentences with additional emphasis in specific words; and spontaneous speech.

The complete evaluation protocol amounts to less than 10 minutes of speech per patient.

Table 2.2 summarizes the battery of tasks and the number of exercises per task. Each of the 50 PD patients and HCs completed the full battery of exercises, for a total of 4800 recordings.

2.2.3 Other related corpora

It is also of interest to acknowledge other popular corpora for SA diseases not contemplated in this theses that are frequently used by the research community. Although their detailed description falls out of the scope of this work, we still provide a brief, and non-exhaustive list of them. The Upper Respiratory Tract Infection Corpus (URTIC) [57] is a corpus of subjects

Table 2.2: Summary of the battery of tasks and the number of exercises per task for each participant in the New Spanish Parkinson Corpus.

Task	**# Exercises**	**Aprox. duration [sec]**
Monologue	1	60
Read text	1	20
Read sentences	6	5
Vowels	5	3
DDK analysis	6	5

affected by the cold/flu. The Dementia Bank [58] is a corpus dedicated to Alzheimer's disease, mild neurocognitive disorder, and primary progressive aphasia. The TORGO is database of dysarthric articulation, which consists of aligned acoustics and measured 3D articulatory features from speakers with either cerebral palsy or amyotrophic lateral sclerosis [59]. The Child Pathological Speech Database (CPSD) [60] covers diseases from the autistic spectrum, specifically in children. More recently, the Coswara, is a corpus of respiratory sounds, such as cough, breath, and voice, of COVID-19 patients [61].

We note that all the listed corpora contain CC recordings that impose the same limitations as the DAIC and PC-GITA regarding real-life applications to diagnose, assess the severity, and monitor SA diseases. To different extents, each of this corpora verifies at least on of the conditions to be considered collected in CC: the content is determined by speaking exercises or clinical interviews; the channel is known and/or controlled; the noise is controlled and/or minimized.

To the best of our knowledge, there exist no speech medical datasets that mimic real-life scenarios. In fact, datasets in such conditions, also referred to as in-the-wild, are rare. One of the few examples of a task somewhat related to detection of SA diseases is the Acted Facial Expressions In The Wild (AFEW) Corpus [62], which claims to mimic real-life scenarios through close to real world environments extracted from movies.

Chapter 3

Automatic, speech-based detection of Depression and Parkinson's disease

Chapter 2 described the physiological mechanisms of depression and PD, as well as how they can affect speech production from a perceptual point of view. Now we move on to look at SA diseases from an automation perspective. In this case it is necessary to translate the previously described perceptual characteristics of speech affected by depression or PD into objective measures that help differentiate it from healthy speech.

This Chapter, specifically Section 3.1, begins by reviewing the literature for the most relevant studies performed towards improving the automatic, speech-based detection and severity assessment of depression and PD (in 3.1.1, and 3.1.2, respectively). In both cases, we adopt a historical perspective and begin by summarizing the earliest relevant works in this topic, which tended to be heavily dependent on handcrafted features, and advance chronologically to more recent ones, which tend to be dominated by data driven approaches.

Finally, Section 3.2 reviews the state-of-the-art for the automatic annotation of corpora, not necessarily in a speech or health care context. This review will provide context for the work developed in this thesis regarding the automatic annotation of the WSM Corpus, which is described in more detail in Chapter 5.

3.1 Automatic detection of SA diseases: Depression and Parkinson's disease

3.1.1 Depression

The early days:

Using speech to detect signs of depression in individuals dates back to as early as the 1930s, when the earliest paralinguistic investigations into depressed speech were carried out. The first efforts to systematically use recordings as the patients read passages and answered psychiatrist's questions, to allow the review and repetition of their speech samples were performed by [63]. They did not have the technology to apply acoustic methods and found that even a skilled speech pathologist required multiple repetitions of a speech sample to develop reliable impressions. The measures used in this work included ratings of tempo, and pause frequencies, and rather than providing summary data, they described a few prototypic cases. These authors argued that the monotony of depressed voice, was a result of the reduced prosodic variability. The authors also identified rate differences between read and free speech, and rate differences between different topics of conversation. They called for the use of objective and reliable measures of voice, and in many ways, were ahead of their time.

Later on, works in the 1960s and 1970s showed that depressed speech was negatively correlated to rate of productivity and filled pauses, and positively correlated to silent pauses [64]. Depressed speech was typically characterized by decreased loudness and pitch variability [65][66].

In [67], the authors used measures of amplitude and frequency variability to reflect the monotone quality and the previously mentioned "flatness" of the speech of both depressed and schizophrenic patients. They seem to have been the first ones to have suggested the possibility of developing voice profiles to assist diagnosis in psychiatry.

The Signal Processing and Machine Learning days:

Nowadays, there is a plethora of approaches using diverse machine learning based strategies do detect signs of depression from speech. However, the results presented in different works are generally difficult to compare, given the lack of standard datasets of speech for depression detection, other than the DAIC, previously described in detail in Section 2.1. As such, this document will focus on works that report their results on DAIC, and a few other notable works that use other corpora.

Furthermore, this document will focus on reviewing the approaches that use speech alone to

determine the presence or severity of depression, and will disregard multimodal approaches that include visual, natural language, or other types of cues. While there has been significant success in using other modalities to detect depression, these exceed the scope of this work.

The baseline provided in the AVEC 2016 [68], for determining the severity of depression in the DAIC-WOZ, based solely on speech, was obtained via Support Vector Machines (SVMs). The proposed baseline used prosodic, spectral and voice quality features, as well as the four first formants at every 10ms. The acoustic features were extracted with the COVAREP toolkit [69], and the resulting 79-dimensional feature set was used to fit a linear SVM trained with stochastic gradient descent (SGD). The model was validated on the development set, and the hyperparameters were optimized via grid search. Temporal fusion, to obtain a final interview level prediction, was achieved through simple majority voting of the predictions for all the frames within an entire screening interview. This baseline yielded an F1 score (defined as the harmonic mean of precision and recall) performance on the test set of 0.410 for the depressed class and 0.582 for the not depressed class.

Later on, this baseline was beat in the 2016, 2017, and 2019 editions of the AVEC by several research teams. Notably, there were several approaches that successfully predicted depression or accessed its severity.

In [70], the authors proposed a gender specific decision tree, constructed according to the distribution of the multimodal prediction of PHQ-8 scores (same as PHQ-9 without the item for suicidal ideation) and participants' characteristics (PTSD/Depression Diagnostic, sleep-status, feeling and personality) obtained via the analysis of the transcript files of the participants. At each node of the tree, there is a separate Support Vector Regression (SVR) model with Radial basis function (RBF) kernel to predict the PHQ-8 score. The single stream decision tree for the speech modality proposed by the authors obtained a root mean squared error (RMSE) in the development set of 6.224 and 6.910 for females and males, respectively, and a mean absolute error (MAE) of 4.842 and 5.750, for females and males, respectively.

Another work proposed a Gaussian staircase modeling approach, which generalizes the use of Gaussian distributions for binary classification into the domain of multivariate regression [71]. This is accomplished by partitioning the outcome variable into multiple nested ranges with binary class labels for "lower" and "higher" being associated with complementary ranges at each nested partition. A multivariate normal distribution is used to model the class-conditioned features in each partition, and the class-conditioned likelihoods are computed by summing the likelihoods across all the partitions. The authors used correlation structure (CR) formant features, CR δ Mel-frequency cepstral coefficients (MFCC) features, spectral

energy, and peak-to-rms. The performance they obtained on the development set of the DAIC-WOZ was a RMSE of 6.38 and a MAE of 5.32. Furthermore, the authors report very interesting findings after performing a thorough analysis of the DAIC-WOZ. They report finding several limitations of the dataset: significant audio-transcript misalignments for some speakers; change in the protocol of the virtual interviewer's behavior after one third of the interviews (in terms of turn duration, and questions asked to the patients); inconsistent signal to noise ratio (SNR) between interviews of different subjects.

In [72] the authors beat the challenge baseline using two approaches: training a linear SVM model with SGD where the input features were Teager energy cepstral coefficients (TECC); and performing a Gaussian Probabilistic Linear Discriminant Analysis (G-PLDA) with i-vector modelling based on MFCC features. The last approach was the most successful of the two. The authors reported an F1 score on the development set of the DAIC-WOZ of 0.43 and 0.86 for the depressed and non-depressed classes, respectively, using the TECC features with the linear SVM model; and an F1 score of 0.57 and 0.89 for the depressed and non-depressed classes, respectively, using the i-vector features with the G-PLDA.

The authors of [73] proposed a deep learning based, DepAudioNet, to encode the depression related characteristics in the vocal channel, combining Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) Recurrent Neural Networks (RNN). They introduced a random sampling strategy in the model training phase to balance the positive and negative samples, which helped alleviates the bias caused by uneven sample distribution. The input to the proposed network was raw spectrograms and Mel-scale filter bank features. The authors reported an F1 score on the development set of the DAIC-WOZ of 0.52 and 0.70 for the depressed and non-depressed classes, respectively.

Another solution is based on a gender dependent model to address the challenge [74]. They took advantage of the transcription and timestamps that were made available by the challenge organizers to re-compute a number of low level descriptors according to a new protocol that excluded frames with laughter, sighs, unvoiced segments ($VUV = 0$), and voiced segments lasting less than 5 ms. All of the frames that verified any of the above mentioned conditions were considered to be non-informative for the task of detecting depression. The remaining frames were used to compute statistical descriptors from the low-level descriptors provided by the challenge organizers, 10 Discrete Cosine Transform (DCT) coefficients, and 8 high level features computed at interview level (pause ratio, voiced segment ratio, speaking ratio, mean laughter duration, mean delay in response, mean duration of pauses, maximum duration of pauses, and fraction of pauses in overall time). The authors do not provide many details in terms of the modeling strategies they chose, but reported an F1 score on the development set

of the DAIC-WOZ of 0.59 and 0.87 for the depressed and non-depressed classes, respectively.

Two works adopted a deep CNN (DCNN) based strategy [75] [76], using the Geneva minimalistic acoustic parameters (GeMAPS) [77]. From those, the first and second order derivatives were also computed, when relevant, which was used as the network input. The authors chose to train two separate models, one for each class, depressed and non-depressed. They report the performance the development set of the DAIC-WOZ, and claim an RMSE of 4.516, 2.767, 1.467, and 2.694 for depressed females, non-depressed females, depressed males and non-depressed males, respectively; and a MAE of 3.633, 2.350,1.226 and 2.092, for the same gender and model combination.

Another work proposed a solution based on multi-scale temporal dilated CNNs (MS-TDCNN) [78]. These used a special case of dilated convolution, also called convolution with holes, where the multi-scale filters skip the input values with a certain step along the temporal dimension. Given the set of features of MFCCs, first and second derivatives, and extended Geneva minimalistic acoustic parameters (eGeMAPS) [77], for an interview, these were divided into N spans of t frames. For each span, the authors computed maximum, minimum, average and standard deviation of each feature, and they appended these to the existing set of features to the span. The concatenation of the original set of features and the span-wise statistical audio features were the input to the network. They reported an RMSE of 6.20 and an MAE of 4.88 on the 2019 extended DAIC (e-DAIC), which is an extended version of the original DAIC corpus. The baseline reported for this dataset obtained an RMSE of 6.43 for the speech modality only.

The authors of [79] proposed a multi-level attention network, which the authors claimed reinforced overall learning by selecting the most influential features within each modality. For the speech modality the authors trained four models, one for each type of features: MFCCs, eGeMAPS, Bag-of-AudioWords (BoAW) [80], and a high dimensional deep representation of the audio sample, extracted by passing the audio through a Deep Spectrum and a visual geometry group (VGG) network. The authors only reported per network performances, and did not report the results for the fusion of the four networks for the audio modality. The best performing model was the one based on the MFCCs, which yielded a RMSE of 5.11, on the development set of e-DAIC.

In [81] the authors propose the use of deep convolutional generative adversarial networks (DCGANs) to overcome the limited amount of annotated on the DAIC, by developing a model that generates new examples of feature vectors, thus augmenting the available training data. The quality of the generated examples is measured in terms of characterizing the spatial,

frequency and representation learning of the augmented features. They were able to achieve a RMSE of 5.520 and MAE of 4.634.

Finally, in [82], the authors propose a Hierarchical Attention Transfer Network, a novel cross task approach which transfers attention mechanisms from speech recognition to aid depression severity measurement. The transfer is applied in a two-level hierarchical network which mirrors the natural hierarchical structure of speech. Their experiments based on the DAIC, demonstrated the effectiveness of their model. On the development set, the approach achieved a RMSE of 3.85, and a MAE of 2.99.

3.1.2 Parkinson's disease

As in the case of depression, PD can be diagnosed through multiple bio-signals, as well as through visual cues. The goal of this Section is to review the works that are dedicated to detection of PD from speech alone.

The earliest efforts to automate the detection of PD using speech were motivated by previous perceptual studies where differences in phonation, articulation and prosody between healthy individuals and individuals with PD were clearly detected by trained medical specialists.

In contrast with depression, however, there is no standard dataset that is consistently used to diagnose, or access the severity of PD. The most commonly adopted strategy is to have the authors collect their own, typically very small dataset, with a handful of PD patients and HCs, performing one or several short speech tasks, such as sustaining vowels, repeating predetermined sequences of syllables, reading sentences, or doing short monologues. Usually the data collected for each patient range from a few seconds up to a few minutes of speech. Therefore, the results of different works are not directly comparable, and it is not trivial to determine which are the most promising strategies to diagnose or access the severity of PD. Furthermore, most works report results on datasets that are not balanced in terms of age, or gender, and do not address the biases that such imbalances may contribute to their findings. Nevertheless, the remainder of this Section will be dedicated to summarize some of the most relevant works and trends in this field.

As mentioned in Chapter 2.1, PD can cause speech impairments in patients in terms of three principal dimensions: phonation, articulation, and prosody. Some works focus exclusively in capturing the differences in one of these dimensions, while others consider all three at the same time.

The symptoms related to *phonation* impairments are related to the stability and periodicity

of the vocal fold vibration. They have been analyzed in terms of perturbation measures, and the most relevant features used in the literature are derived from jitter (absolute and average absolute difference between cycles), the amplitude perturbation quotient, shimmer (calculated as the average absolute difference between the amplitudes of consecutive periods), pitch perturbation quotient, harmonics to noise ratio (HNR), noise to harmonics ratio (NHR), MFCCs, and non-linear dynamics measures. Typically these features are computed over recordings of sustained vowels. That was the case of [83], which performed an analysis of some of the above mentioned features using recurrence period density entropy, detrended fluctuation analysis, correlation dimension, and the pitch period entropy. The authors reported an accuracy of 91% in a test set of 23 PD patients and 8 HCs. In [84], the authors also used a similar set of features computed over sustained vowels, but on a different speech dataset. The authors adopted random forests and support vector machines as their classification strategies, and reported, in the best case, a performance of 94.4% accuracy in detecting phonations by PD patients in a test set of 263 speech samples from 43 subjects (33 with PD and 10 HCs). Although the results seemed promising, the authors did not guarantee speaker independence between the train and test sets, and as such biased and optimistic conclusions may be drawn.

Articulation symptoms in patients with PD are related to the modification of position, stress, and shape of several limbs and muscles to produce speech. These symptoms have been modeled mostly by spectral features, including vowel space area, vowel articulation index, formant centralization ratio, diadochokinetic analysis (DDK), onset energy, and MFCCs and its derivatives. These features tend to be computed over a monologue, reading text, sentence repetition, or some other form of running speech [85][86]. In [85], the authors were able to find a strong correlation between features derived from different formant quotients and the presence of dysarthric speech in patients with PD. In [86], the authors studied a group of 35 Czech native speakers (20 early PD patients and 15 HCs), and were able to show that even at early stages of PD, it is possible to use features that capture the characteristics of articulation, to discriminate between PD patients and HCs with an accuracy of about 80%.

Prosodic differences between the speech of PD patients and HCs are manifested as monotonocity, monoloudness, and changes in speech rate and pauses. These changes need to be captured over time, using features derived from pitch energy contours, and duration. Some of these features include the F_0, its mean and standard deviation, intensity, and its standard deviation, all of these computed over a recording of running speech. Additionally, other useful features to measure changes in prosody are the speech rate, which can be measured using the length of each syllable and each pause, the net speech rate (NSR), which is measured in syllables per second related to the net speech time in milliseconds, percent pause time, articulation

rate, number of pauses, *etc.*. In [87], the authors studied prosodic differences, and observed that there is a correlation between several PD symptoms and prosodic variables, such as the number of pauses in speech. They also showed that the variation of F_0 is lower in PD patients than in HCs.

More recent works tend to consider the differences over the three dimensions of speech, between PD patients and HCs. These works combine the features and approaches used by the above mentioned works. In [88], the authors considered a total of 46 participants (23 with PD and 23 HCs), and performed an analysis of their phonation, articulation, and prosody. The authors concluded that 78% of the patients evidenced speech problems: prosody was the most affected dimension of speech, even in the initial stage of the disease, and articulation was the second most affected dimension. They also found that the variation of the fundamental frequency measured on the monologues and emotional sentences contained very useful information for separating HCs from PD speakers. In [89] and [90], the authors used the publicly available OpenSMILE [91] toolkit to extract a set of 1582 acoustic features per utterance that describe the phonation, articulatory and prosodic characteristics of a speech signal. The former trained separate models for phonation, articulation, and prosody using appropriate subsets of the OpenSMILE feature set, and after combining the three tasks were able to obtain an unweighted average recall (UAR) of 81.9% on a dataset of 176 German native speakers (88 with PD and 88 HCs). The latter attempted to model all of the features in the OpenSMILE feature set as a single task, using several regression techniques, including ridge, lasso and support vector regression, to access the neurological state and the severity of patients with PD. According to their reports, features extracted from the reading texts are the most effective and robust to quantify the severity of the disease.

Besides the classic features extraction methods that use with hand crafted features, there has been some interest in exploring deep learning approaches to detect and monitor PD using speech. An example of such efforts was the "2015 Computational Paralinguistics challengE (ComParE)" [92], which had a sub-challenge dedicated to the automatic estimation of the neurological state of PD patients according to the MDS-UPDRS-III score. The corpus used in this challenge was the subset of the PC-GITA, previously described in Section 2.2.2, corresponding to the speakers affected by PD.

The winners of this challenge [93] adopted Gaussian processes and deep neural networks (DNN) to predict the clinical scores and reported a correlation of 0.65. In [94], the authors proposed a deep learning model to assess dysarthric speech. The model aimed to predict the severity of dysarthria and proposed introducing an intermediate interpretable hidden layer in a DNN that contained four perceptual dimensions: nasality, vocal quality, articulatory

precision, and prosody. The authors presented an interpretable output that was highly correlated (Spearman's correlation of up to 0.82) with subjective evaluations performed by speech and language pathologists. In [95], the authors modeled the composition of non-modal phonations in PD. The authors computed phonological posteriors using DNNs to predict the dysarthria levels. In [96], the authors modeled articulation impairments of PD patients with time-frequency representations (TFR) and CNNs. The authors classified PD and HC speakers considering speech recordings in three languages: Spanish, German, and Czech. They reported accuracies from 70% to 89%, depending on the language. In [97] the authors used articulation features extracted from continuous speech signals to create i-vectors, used train a model to predicted the dysarthria level according to the FDA score. In [98] the authors take advantage of X-vectors [99], to generate DNN based speaker embeddings that are then used to train a probabilistic linear discriminant analysis (PLDA) model that distinguishes healthy from non-healthy speech. X-vectors are known to be able to capture meta-information besides the speaker identity, such as gender or speech rate [100], and now also, some articulatory, prosodic or phonatory information characteristics that characterize PD affected speech.

In [101] the authors propose not only detecting PD, and distinguishing it from healthy speech, but also distinguishing PD from Amyotrophic Lateral Sclerosis (ALS), which is another prevalent neuro-degenerative movement disorder. Speech related complications that ALS patients typically experience include dysphagia, dyspnoea, orthopnea and dysarthria, which have some overlap with the symptoms experienced by PD patients. This makes the task of distinguishing these two diseases more difficult than either of them from healthy speech. The authors opted to use a modeling strategy based on CNN-LSTMs and transfer learning, where they leveraged from the information from the ALS corpus to detect PD, and *vice-versa*.

3.2 Related work: automatic corpora labeling

The task of automatically annotating corpora is typically associated with scenarios where manual annotation is not a feasible approach. This tends to be the case in one of the following situations: when the number of examples that need to be annotated is larger than what can be processed by the available human annotators; when the examples need to be annotated faster that human annotators can achieve; when the cost of manually annotating the examples is prohibitively high; or simply, when there are no human annotators available for a given annotation task.

In any of these cases, there is a compromise between the performance that would be achieved by the human annotators, and overcoming whatever limitation exists in the annotation

process.

Most automatic corpora annotation strategies are based on using a small "gold dataset" of manually annotated examples as a starting point for the automatic annotation process, and/or exploit the specific characteristics of the data that are being annotated, by imposing restrictions based on their knowledge of the data on the automatic annotation model, thus making the overall task less ambiguous.

In this Section we will summarize some of the most popular works on automatic corpora annotation, particularly those that propose solutions that do not heavily rely on prior domain knowledge. Since automatic corpora annotation is a transversal task, popular in several different domains where annotated data are scarce, we will provide a review of the techniques used to solve this problem, regardless of the applications (this means that we cover works outside the speech or medical domains).

In [102], the authors introduce a method for the automatic annotation of images with keywords from a generic vocabulary of concepts or objects for the purpose of content-based image retrieval. Each image is represented as sequence of feature-vectors characterizing low-level visual features such as color, texture or oriented-edges, and is modeled as having been stochastically generated by a hidden Markov model (HMM), whose states represent concepts. The parameters of the model are estimated from a set of manually annotated (training) images. Each image in a large test collection is then automatically annotated with the a posteriori probability of concepts present in it.

In [103] the authors revisit a well-known active learning algorithm: uncertainty sampling. They propose an adaptation of this technique with lower computational complexity: approximate uncertainty sampling, which is applied in the context of finding spam e-mails in large e-mail corpora. The onus of active learning strategies is optimizing the strategies of iteratively finding the few examples in the dataset that should be sent for human annotation. In this case, at each iteration, approximate uncertainty sampling selects only a subset of examples to re-evaluate and then chooses the best m examples amongst this limited subset. The key to the effectiveness of this technique is that, at each iteration, the model, rather than reevaluating the uncertainty of each message, performs the sampling using the uncertainties calculated in previous iterations.

In [104], the authors propose a graph-based semi-supervised learning approach that incorporates embedding techniques. In their work, they propose that the embedding of an instance is jointly trained to predict the class label of the instance and the context in the graph. Then they concatenate the embeddings and the hidden layers of the original classifier and feed them

to a softmax layer when making the prediction. They formulate a transductive solution where the embeddings are learned based on the graph structure, as well as an inductive solution where they define the embeddings as a parameterized function of input feature vectors, *i.e.*, the embeddings can be viewed as hidden layers of a neural network. This work was applied in the context of annotating new examples for text classification, entity extraction and entity classification.

In [105], the authors present an approach capable of training deep neural networks on large-scale weakly-supervised web images, which are crawled from the Internet, using text queries, without any human annotation. Their learning strategy leverages from curriculum learning, with the goal of handling a massive amount of noisy labels and data imbalance effectively. They design the learning curriculum by measuring the complexity of data using its distribution density in a feature space, and rank the complexity in an unsupervised manner. This allows for an efficient implementation of curriculum learning on large-scale web images, resulting in a high-performance CNN model, where the negative impact of noisy labels is reduced substantially.

Overall we have reviewed some works that use generative models, active learning strategies, weakly supervised learning strategies, and semi-supervised learning strategies for automatic corpora annotation in very different domains. There are a number of variants for each of these approaches. The choice for each approach and variant should be based on the constraints of the specific problem being solved, and there is not a single approach that is necessarily the best for every case.

Part II

Towards automating the collection and annotation of speech medical corpora

Chapter 4

The In-the-wild Speech Medical Corpus

In Chapter 2 we have identified that there is a lack of resources, specifically speech corpora, that faithfully represent SA diseases in in-the-wild contexts. The existing ones are small in size, and collected in CC. In an attempt to overcome this, we have collected the WSM Corpus. This is an audiovisual corpus of videos collected from the online multimedia repository YouTube, mostly featuring recordings in the vlog format, of subjects potentially affected by SA diseases.

Vlogs, short for both video blog and video log, are a popular video format where subjects record themselves talking about one or several topics of their choosing, ranging from products or media reviews, to video diary entries, among many others. These videos are typically recorded in very informal settings, such as at home, in a car, *etc.*. Furthermore, the recordings are usually made with a smartphone, laptop computer, or other non professional camera and microphone equipment.

This category of videos, vlogs, and more generally, informal videos, are a valuable resource that portrait human behaviour and speech in real life conditions, which provides a window to study SA diseases specifically in an in-the-wild context. Figure 4.1 shows some examples of screenshots of vlogs included in the WSM Corpus, to provide a sense of the nature of this corpus.

It is important to clarify the meaning of our prior claim that the WSM is a corpus of people potentially affected by SA diseases. In the context of this corpus, the videos are categorized into two classes: videos containing a subject making an explicit claim that they are currently affected by the target SA disease, to which we refer to as *self-reported health status*, and videos that do not contain such claims. We emphasise that self-reported health status does

Figure 4.1: Frames from six videos of the WSM Corpus, showing what the setting of the typical video is. Usually in a vlog, or other informal video, the subject is addressing the camera, and records him/herself in a familiar environment, such as the house, car, or a nearby park.

not depend on the presence of any symptom related to the target SA disease, nor it is the same thing as the *true health status* of the subject. The latter would have to be determined by an appropriate healthcare specialist. However, for both practical and ethical reasons, that determination is not made for the videos of the WSM Corpus, therefore its videos are only classified in terms of the self-reported health status.

We acknowledge that the self-reported health status is not the same as the true health status, but, at the same time, it is our claim that the subject's self-reported health status in vlogs tends to be accurate and truthful, as there is no incentive for the subjects to be deceitful about their health status. We will show later, in Chapter 6, several experimental results that support this claim, that in the context of the WSM Corpus, self-reported health status is, in fact, a good proxy for the true health status, and therefore, they can be used interchangeably.

In Section 4.1 we begin by describing the collection methodology of the videos pre-selected to be part of the WSM Corpus, from our multimedia platform of choice, YouTube. We also summarize what metadata we extracted, along with the videos.

Then, in Sections 4.2, 4.3, and 4.4 we describe each of the three versions of the WSM corpus, where each one was created with a specific purpose in mind. Specifically, in Section 4.2 we will describe the earliest version of the WSM Corpus, which was our first effort of collecting in-the-wild speech medical data. The goal of this version of the corpus was to create a small

collection of videos to use as proof-of-concept in showing the usefulness of in-the-wild speech medical data to detect SA diseases and to mimic real life conditions. Our target for this version was to collect and annotate approximately 60 videos per SA disease. The WSM Corpus v.1 contains subsets for depression, PD, and the common cold/flu, all manually annotated by one non-expert for self-reported health status.

The second version of the WSM Corpus, described in Section 4.3 was collected with the intent of adding more HC examples to the corpus, therefore the corpus contains not only vlogs related to depression and PD, but also additional vlogs of topics unrelated to SA diseases. This version of the corpus was also manually annotated for the self-reported health status of the speakers by one human non-expert.

Finally, Section 4.4 describes the third and final version of the WSM Corpus, which is our attempt at creating the first large scale corpus of in-the-wild SA diseases, particularly for depression and PD, with several hundreds of videos per SA disease. Additionally, this version of the corpus also contains age and gender information about the subjects in the videos, which was not present in prior versions of the corpus. Furthermore, the corpus is balanced in terms of both self-reported health status and demographic characteristics (age and gender). This version of the corpus was annotated through the crowdsourcing platform Amazon Mechanical Turk (AMT), where each video was reviewed by at least five non-expert annotators. The final labels for each video correspond to the aggregation of the accepted answers of the annotators.

We emphasise that the thorough description of each version of the WSM Corpus is necessary as the experiments presented in Chapters 5 and 6 use distinct versions of the corpus (the most current one at the time of the work).

4.1 Collection Methodology

The videos in all of the versions of the WSM Corpus were pre-selected by using a combination of the official YouTube API and scrapping tools to retrieve lists of search results for sets of desired queries, and time windows for the date of publishing.

For each pair of query and time window, a list of up to 50 YouTube videos were retrieved. An example of a possible query and time window could be: [“depression vlog”, “01/01/2020-12/31/2020”]. Each video in the search result was accompanied by the following information:

- A unique identifier
- A title, as assigned by the uploader

- A description, as assigned by the uploader (optional, can be left blank)
- The transcription (when available. This transcription is automatically generated for videos in English, unless provided by a user)
- The uploader channel's unique identifier
- The playlist's unique identifier, if the video is part of a playlist
- The timestamp of the time of publishing
- The video category, as assigned by the uploader (one out of a closed set of 14 categories, including "News", "Music", "Entertainment", *etc.*)
- The total number of of times the video was viewed
- The total number of "thumbs up" given to the video by the viewers
- The total number of "thumbs down" given to the video by the viewers
- The comments to the video, including the unique identifier of the user that made the comment, and the timestamp of when it was posted

We note that the video's transcription was automatically generated by YouTube, using a large scale, semi-supervised deep neural network for acoustic modeling [106], unless a human transcription was provided by a user.

The set of pre-selected videos was further narrowed down to be part of each version of the WSM Corpus according to different criteria, which is described in the corresponding Sections.

The choice to collect a dataset specifically of vlogs originated from empirical experiences when manually searching for videos of people affected by SA diseases. We were able to verify that searching simple for a "[*target disease*]" would yield vastly different results compared to "[*target disease*] vlog", in terms of finding content that included people currently affected by the "*target disease*". As an example, Figure 4.2 shows the differences in the search results for the queries "depression", on the left side, and "depression vlog", on the right side. Each search result in this Figure marked in red corresponds to a video that is somehow related to depression, but does not necessarily contain a subject currently affected by it. This includes lectures about depression, motivational videos, videos of therapists, medical doctors or other healthcare specialists sharing knowledge about depression, videos of caretakers or partners of people suffering from depression, and even videos of people who are describing their past experiences with depression, who are currently not affected by it. We observed the same phenomena for PD, cold/flu, Alzheimer's disease, bipolar disorder, and obstructive sleep

apnea (OSA). Marked in green are the videos of the people who verbally confirm to be currently affected by depression. As it can be seen, the target videos are found much more commonly when the search query includes the term "vlog".

4.2 WSM Corpus, v.1

The first version of the WSM Corpus contains a total of 182 videos, published between January 2017 and January 2018. The videos in the WSM Corpus v.1 were pre-selected using the queries "Depression vlog", "Parkinson's disease vlog", and "Flu vlog".

The videos retrieved with the query "depression vlog" tend to display subjects that are currently suffering from depression, or have suffered in the past (and are recalling their past experiences), or videos from therapists and other healthcare specialists, among others. In the case of PD, the videos retrieved with the query "Parkinson's disease vlog" tend to show either PD patients relaying their current experience with the disease (*e.g.* discussing symptoms, treatments, lifestyle changes, the progression of the disease, *etc.*), or videos of informal caregivers, healthcare professionals, and others. Finally, the videos retrieved with the query "Flu vlog" tend to include videos where the speaker is affected with a cold or a flu and is describing how they are feeling, or experiences from parents and other guardians describing the experiences of their children going through a cold or flu.

A total of 177, 164, and 174 videos were pre-selected for depression, PD, and flu, respectively, and approximately 60 randomly chosen videos for each SA disease were manually annotated. All the annotated videos of the corpus are in English, but not restricted to native speakers. These videos have an average duration of approximately 10 minutes, totaling approximately 30 hours of raw recordings.

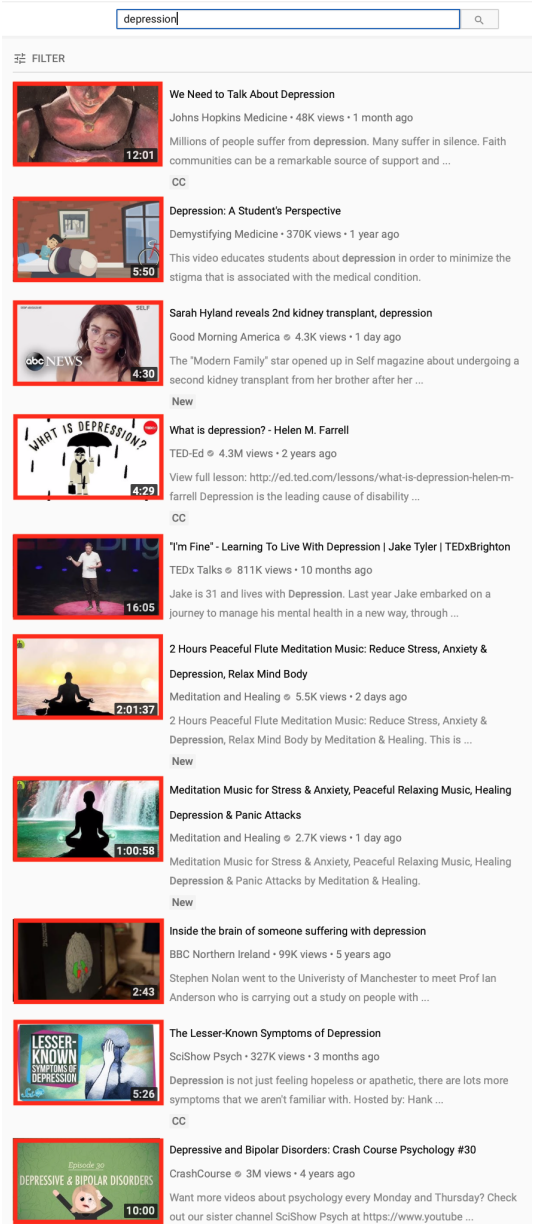
Each annotated video in this version of the corpus was manually annotated by one non-expert annotator, with five binary labels, corresponding to a yes or no answer to each question of the following questionnaire:

Q1: *Is the video in a vlog format?*

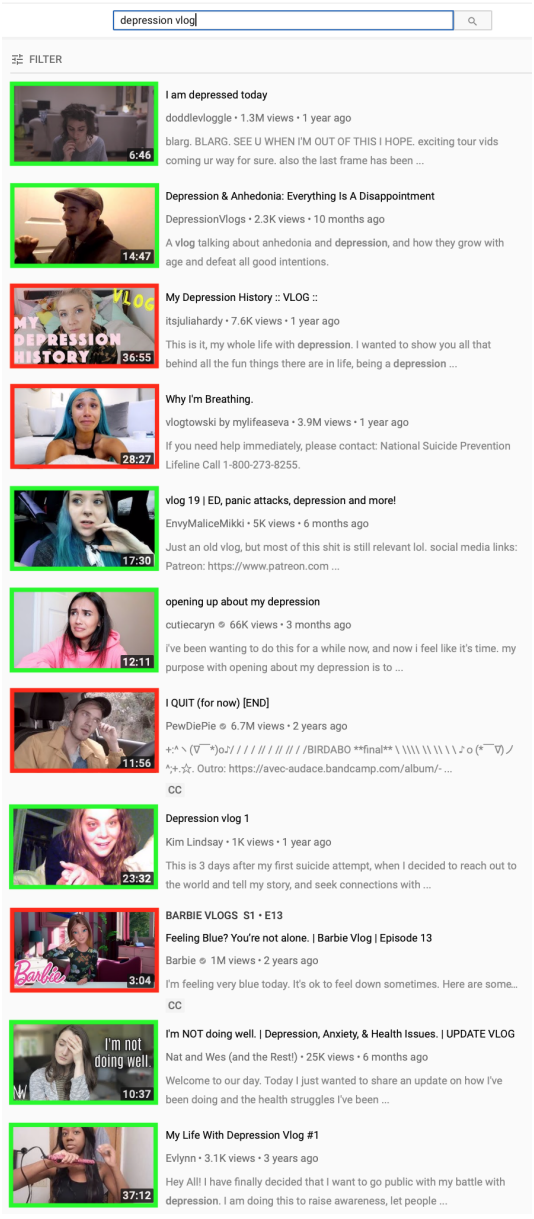
Q2: *Regardless of the topic, is the main speaker of the video talking mostly about themself?*

Q3: *Regardless of the topic, is the main subject of the video talking mostly about present events/opinions/situations/etc. or from a recent past (over the last few hours/days)?*

Q4: *Is the main topic of the video related to [target disease]?*



(a) "Depression"



(b) "Depression vlog"

Figure 4.2: Example of the search results for the query "Depression" (left), and "Depression vlog" (right) on the multimedia repository YouTube. Next to each video thumbnail is the title in bold, and below is the channel's unique identifier, the number of video views, and how long ago the video was posted. The last lines show a preview of the video's description, as written by the uploader. Video thumbnails outlined by a red box correspond to videos of people who do not claim to be currently affected by depression, and video thumbnails outlined in green correspond to target videos featuring subjects who claim to be currently affected by depression.

Table 4.1: Positive class incidence per label, per disease for the WSM v.1.

Target disease	Query	# Annotated examples	# Is vlog	# 1st person	# Present	# Topic is [target disease]	# Positive [target disease] self-reported diagnosis	# Non annotated examples
Depression	*"depression vlog"*	58	53	45	32	35	18	119
Parkinson's disease	*"Parkinson's disease vlog"*	61	35	34	35	43	18	103
Flu/Cold	*"flu vlog"*	63	62	51	58	40	30	111

Q5: *Does the subject claim to be currently affected by the [target disease]?*

Where [*target disease*] can be depression, PD, or flu/common cold.

We note that the last question in the questionnaire in practice corresponds to an intersection of the second, third, and fourth questions. This is because a positive self-reported diagnosis must contain a subject referring to themself, talking about their present situation, and also about the target disease. An example of this would be the sentence *"I'm currently suffering from depression."*.

Table 4.1 summarizes the class distribution for each of the five binary labels for each SA disease. The most important information to note is that, out of the annotated videos, averaged over the three queries and respective SA diseases, 36.3% contain a subject claiming to be currently affected by the respective [*target disease*].

4.3 WSM Corpus, v.2

The second version of the WSM Corpus contains 550 videos, published between January and July of 2016, of English speakers (native and non-native). The dataset includes two subsets for depression, and PD. This version of the corpus was collected with the intent of containing more examples of control data, mostly of vlogs unrelated to depression or PD, that were included as control data for both datasets. The videos were annotated by one non-expert annotator.

The depression subset contains 100 videos collected with the query "depression vlog". These videos were annotated only for the self-reported health status for depression encoded as a binary label. Out of the 100, 49 contained a subject that self-reported to be currently affected by depression.

The PD subset contains 150 videos collected with the query "Parkinson's disease vlog". They were annotated for the self-reported health assessment of PD, which was only verified for 26

Table 4.2: Positive class incidence, per disease and query for the second version of the WSM Corpus.

Target disease	**Query**	**# Annotated examples**	**# Positive [target disease] self-reported diagnosis**	**# Non annotated examples**
Depression	*"depression vlog"*	100	49	0
Parkinson's disease	*"Parkinson's disease vlog"*	150	26	0
None	*"daily vlog"*	0	0	100
	"vlog"	0	0	100
	"Parkinson's disease lecture"	0	0	100

videos.

The remainder 300 videos of the dataset were retrieved with the queries "Parkinson's disease lecture", "vlog", and "daily vlog", 100 videos for each query, and were not manually annotated. They were considered HCs by default, given the low probability of containing a positive self-reported health status for any of the target SA diseases. These 300 videos were included as HCs for both the depression and PD datasets.

Table 4.2 summarizes the videos on this version of the WSM Corpus, by SA disease and by query used to retrieve it, along with their respective class distribution.

4.4 WSM Corpus, v.3

The third version of the WSM Corpus is our attempt at creating what is, to the best of our knowledge, the largest in-the-wild multimodal corpus of SA diseases. Currently the WSM Corpus is focused on depression and PD, but with more contributions over time, we expect that new datasets will be added, dedicated to more SA diseases such as OSA, cold, *etc.* WSM v.3 currently contains 956 videos collected from YouTube, published between January 2016 and January 2019. The language of the videos was restricted to English, however it was not restricted to native speakers.

The videos in the corpus correspond to a subset of a larger pre-selected set of over 1800 videos, collected by using a combination of the official YouTube API and scrapping tools to retrieve the results for relevant queries relative to the target SA disease (such as "depression vlog" and "Parkinson's disease vlog", for depression and PD, respectively) and irrelevant queries (such as "diary vlog", "vlog knitting", "book review", *etc.*). As such, the set of pre-selected videos corresponds to a series of vlogs and other informal videos, some of which are related to

the target SA disease and others are not. From the irrelevant queries we obtain generic vlogs with a single speaker, not related to depression or PD, discussing a broad category of topics, from video diary entries, to daily routines, errand running, media reviews, short tutorials, among many others.

We will proceed to summarize the protocol that was used to annotate the pre-selected videos retrieved with both relevant and irrelevant queries, and the criteria that was used to select which ones to include in the final version of the corpus.

4.4.1 Annotation protocol via crowdsourcing

The protocol for the annotation follows below. The videos were annotated in batches, via the crowdsourcing platform AMT by at least five distinct human non-expert annotators. The annotators were given a short questionnaire where they were asked to watch the video and estimate the age and gender of the speaker in the video. The age was estimated via the following question:

Q: *What is the apparent age group of the subject in the video?*

A: *[0-18, 18-30, 30-40, 40-50, 50-60, 60-70, 70-80, 80+, Ambiguous/difficult/not possible to answer]*

The gender was estimated with the following question:

Q: *What is the apparent gender of the subject in the video?*

A: *[Male, Female, Ambiguous/difficult/not possible to answer]*

For both questions the annotator was able to select a single answer. This allowed us to obtain a simple demographic characterization of the subjects in the pre-selected videos. We note that the option "Ambiguous/difficult/not possible to answer" was made available to the annotators in case there were no speakers or more than one speaker in the video, or in case of a single speaker, if it was difficult to determine their age or gender, *e.g.* in videos where the speaker is talking from behind the camera, or the sound is muffled.

Additionally, for the videos pre-selected via relevant queries only, *i.e.* "depression vlog" and "Parkinsons's disease vlog", for depression and PD, respectively, the questionnaire included an additional question where they were asked to retrieve the self-reported health status of the speaker.

We chose to get this annotation only for the videos pre-selected via relevant queries because we assume that, in videos of topics unrelated to the target SA diseases there will not exist

speakers that claim to be discussing, and more importantly, claim to be currently affected by the target SA disease. This design decision is similar to the ones taken in the scope of the WSM Corpus v.2 describes in Section 4.3.

The self-reported health status was obtained by having the annotators chose one of the answers to the following question:

Q: *Does the subject in the video claim to be currently suffering from [target disease]?*

A: *[Yes, currently; No, but claimed they suffered in the past and got cured; No, but claims to suffer from another disease; No, the speaker makes no claims about their health status]*

The options provided to the annotators were more specific than a simple yes or no answer regarding the subject's self-reported health status. This was a design choice to incentivize them to pay closer attention to the content of the video, rather than simply providing a simple binary answer. Naturally, these labels can be merged into broader categories (a binary yes or no answer) to reflect if the subject is currently suffering from the target SA disease or not.

To illustrate why more granular answers are useful in the context of retrieving self-reported health status, we can use PD as an example. We recall that PD is an incurable, degenerative disease, and that its progress cannot be reversed. As such, in the best case scenario, its symptoms can be managed through appropriate medication, and, in some cases, treatments that include deep brain stimulation. Therefore, if a subject claims to have had PD in the past but to have subsequently gotten cured, we know that this statement does not accurately reflect their experience with PD. However, to non-expert annotators, this scenario could lead to missannotations if presented simply with yes or no annotation options. Although the videos that contain such scenarios are rare, it is important to find them and avoid mistakenly flagging them as target videos. Having more granular options than a simple yes or no answer is helpful for annotators to correctly describe what is happening in the video, as well as to maintain their engagement.

In summary, the questionnaires associated to videos pre-selected with irrelevant queries contained two questions, to estimate age and gender of the subject in the video. The questionnaire associated to videos pre-selected with relevant queries contained three questions, to estimate age, gender, and assess the subject's self-reported health status.

The annotations obtained via AMT questionnaires were subject to approval. We note that the approval or rejection of the annotations is provided at questionnaire level. We established

the approval criteria as follows:

For each video, at least three out of the five annotators had to agreed on gender for the questionnaires to be accepted. Otherwise, the questionnaires were rejected starting from the worker with the lowest acceptance rate from previous batches of videos. Ties were broken by one in-house experienced annotator. The rejected questionnaires were redistributed until a total of five questionnaires per video are accepted. We did not impose agreement restrictions for age.

Additionally, for the questionnaires associated to videos pre-selected with relevant queries we imposed additional constraints for their acceptance. Firstly, we imposed a minimum watch time of 30 seconds of the video, for the annotations to be considered for acceptance. This is because the questionnaire for the videos pre-selected with relevant queries included the more difficult question of determining the self-reported health status of the subject, which requires more work time than estimating age and gender (these two can be usually accomplished in a few seconds). The second restriction that was imposed was that at least three out of the five annotators had to agree on the annotation for the self-reported diagnosis, similarly to the acceptance criteria for the gender estimation question.

At the end of the annotation process, all of the pre-selected videos contained five accepted annotations from different workers for apparent gender, and additionally, the videos retrieved with relevant queries contained five accepted annotations for the self-reported self assessment.

After this, we computed the final labels for each video. The annotations for self-reported health status were simplified to a binary yes or no answer by aggregating the three possible negative answers into a single negative class. The final label for apparent gender and self-assessed health status were obtained via majority voting. Given that the annotation protocol requires that at least three out of the five annotators agree on one answer, that guarantees a lower bound for the inter-annotator agreement for gender and diagnosis. In the case of age, its estimation was obtained by averaging the accepted annotations of the speaker.

Table 4.3 summarizes the total number of videos that were annotated, and the number of questionnaires that were obtained, and accepted, per query.

4.4.2 Video selection

After the annotation process of all the pre-selected videos was completed, we moved on to perform the final selection of videos that would constitute each dataset of this version of the corpus. For this, we began by excluding all the ones without a speaker, or with more that

Table 4.3: Summary of the number of videos and questionnaires given in the scope of the annotation of the WSM Corpus v.3, per query.

Query	**# Videos**	**# Questionnaires**	**# Accepted questionnaires**
Depression vlog	546	3758	2730
Parkinson's disease vlog	716	4819	3580
HCs (vlog, daily vlog, ...)	548	2781	2740

one speaker were removed, leaving only the videos with exactly one speaker. Then, for each SA disease, we divided the videos into two classes: the ones with a positive final annotation for self-reported health diagnosis, and those with a negative or non-existent self-reported health status. Videos with a negative self-reported health status were obtained from the annotations of videos pre-selected with relevant queries for the target SA disease, and videos with non-existent self-reported health status from videos pre-selected with irrelevant queries (this meant that, for example, a video of a subject that had reported to having had depression in the past, but not currently (*i.e.* with a negative self-reported health status for depression), was not considered as a candidate to be an HC for the PD dataset).

From this point on, we considered both negative or non-existent self-reported diagnosis as belonging to the same class: HCs.

Then, for each SA disease, we selected the largest possible subset of videos from the two classes such that the amount of data (in number of videos, and number of hours) was similar, as well as the demographics were similar, *i.e.*, similar apparent gender distribution, and estimated age.

The resulting subset of selected videos contained to 928 videos, between the two diseases, 543 for depression, and 413 for PD. A total of 28 videos of the HC class are shared between the two datasets.

To summarize, at this stage, each video of the WSM Corpus v.3 contained three labels: one determining the subject's apparent gender; one determining the subject's apparent age group; and one determining the subject's self-assessed health status for the target SA disease.

Table 4.4 summarizes the size of the final dataset. In this table it can also be noted that the final set of videos was partitioned into train, development and test partitions, following roughly a $8:1:1$ ratio. Again, these partitions have similar label distribution across age,

Table 4.4: Summary of the WSM Corpus for depression and PD datasets, per partition and group.

WSM Corpus dataset	Partition	Group	# Videos	# Hours	Age	Gender (m:f)
Depression	train	*D*	191	27.6	30 ±5	86 : 105
		HC	199	29.5	30 ±5	93 : 106
	devel	*D*	39	6.0	30 ±5	19 : 20
		HC	40	5.4	30 ±6	19 : 21
	test	*D*	37	6.7	29 ±5	18 : 19
		HC	37	7.8	29 ±5	18 : 19
Parkinson's disease	train	*PD*	157	18.5	45 ±10	79 : 78
		HC	155	20.7	43 ±13	76 : 79
	devel	*PD*	24	1.8	45 ±10	12 : 12
		HC	23	2.6	42 ±10	11 : 12
	test	*PD*	28	4.1	45 ±9	14 : 14
		HC	26	5.8	43 ±12	11 : 15

gender and self-reported health status.

In order to minimize the probability of having the same speaker across different partitions, the data were split without channel overlap, *i.e.* videos from the same author or YouTube channel could only be present in one partition of the dataset.

From table 4.4, we can see that the depression and PD datasets of the WSM Corpus consist of a total of 543 and 413 videos, and 83 and 53.5 hours, for depression and PD respectively. The average of the estimated age for the speakers is 30 ± 5 years and 44 ± 11 years, for depression and PD, respectively. In the depression dataset, a total of 267 videos, corresponding to 40.3 hours, belong to speakers that have self-diagnosed depression, and the remaining videos belong to healthy speakers (this category includes both speakers that have suffered from depression in the past, but are not currently affected by it, or speakers that never suffered from depression). In the PD dataset, 209 videos, corresponding to 24.4 hours, belong to speakers that have self-reported as currently suffering from PD.

We emphasise that the WSM Corpus, particularly the PD dataset, is radically different from others that are publicly available, including the PC-GITA, described in Section 2.2. This is due to the conditions in which the data were collected, and also due to the age of the speakers in the corpus. While the average age of the speakers in the PC-GITA is 61 years old, the average age of the speakers in the PD dataset of the WSM Corpus is 17 years less. We argue that the lower average age may be related to digital literacy as well as milder cases of PD, where the subjects still retain most of their autonomy. This makes the PD dataset of

the WSM Corpus an invaluable resource to possibly study phenomena such as early stages of PD, and subgroups of PD such as early onset PD.

4.4.3 Insights on the WSM Corpus and its annotations

After having annotated all the examples and selected the subset that would integrate the final corpus, we move on to perform an analysis of the annotations collected in the previous stage.

In this Section we will, on one hand, provide some metrics of quality of the annotations in this corpus, based on measures of inter-annotator agreement, and on the other hand, explore the consequences of our annotation acceptance criteria.

We begin by providing a summary of the inter-annotator agreement for each label on the corpus (*i.e.* estimated age, estimated gender, and, when applicable, self-reported health status). Traditionally, the two options to measure the agreement between annotators for non-ordered categorical items are the Cohen kappa [107], and the Fleiss kappa [108]. The former is used when the number of annotators is exactly two, and the latter when there is an arbitrary but fixed number of annotators. Neither of these scenarios correspond to the annotation framework that was used on the WSM Corpus v.3. In fact, for each example we collected a minimum of five annotations. Of all the annotations for one example, five are accepted, and the remaining are rejected. As an alternative to the Cohen and Fleiss kappa, we use agreement ratio as the inter-annotator agreement metric, computed as the quotient between the frequency of the most common answer and the total number of answers. This measure, while less robust than Cohen and Fleiss kappa, which takes into account the possibility of the agreement occurring by chance, is the only one out of the three that is applicable to our scenario. Under the metric of the inter-annotator agreement ratio, higher values represent a better inter-annotator agreement, and the maximum value this metric can take is one, which represents a perfect inter-annotator agreement (where all annotators select the same answer).

We chose to report the average and the median inter-annotator agreement ratio for three sets of annotations per SA disease: the set of all annotations, the set of accepted annotations, and the set of selected annotations (which is the subset of accepted annotations associated to videos that were selected for the final dataset). This analysis will concern the two non-ordered categorical items to be annotated - gender and diagnosis - since no agreement restrictions were imposed for age. The summary of the measures of the inter-annotator agreement ratio can be found on Table 4.5 and 4.6, for depression and PD, respectively.

Table 4.5: Mean and median inter-annotator agreement ratio for the labels of gender and self-reported diagnosis for several subsets of data collected in the scope of the depression dataset of the WSM Corpus v.3.

Annotations subset	**Gender**		**Self-reported health status**	
	Average agreement	**Median agreement**	**Average agreement**	**Median agreement**
All	0.93	1.00	0.64	0.60
Accepted	0.94	1.00	0.69	0.60
Selected	0.97	1.00	0.72	0.80

Table 4.6: Mean and median inter-annotator agreement ratio for the labels of gender and self-reported diagnosis for several subsets of data collected in the scope of the PD dataset of the WSM Corpus v.3.

Annotations subset	**Gender**		**Self-reported health status**	
	Average agreement	**Median agreement**	**Average agreement**	**Median agreement**
All	0.92	1.00	0.77	0.78
Accepted	0.93	1.00	0.89	1.00
Selected	0.98	1.00	0.93	1.00

From Tables 4.5 and 4.6, we can immediately see two phenomena that, intuitively, we would have expected. On one hand, we can see that, for both depression and PD, the inter-annotator agreement ratio (either average or median) is higher for the gender annotations than for the diagnosis annotations. This phenomena goes in line with our expectation that estimating gender is an easier task than determining the self-reported diagnosis. On the other hand, we can observe that, for both depression and PD, as well as for gender and diagnosis, the inter-annotator agreement improves after removing the rejected answers (comparing the inter-annotator agreement for the subset with all the annotations to the one with only the accepted ones). This is also expected, given the annotation approval criteria described before, requiring that at least three of the five annotators select the same answer for a given question. This forces the minimum annotator agreement for the accepted and selected subsets to be 0.6.

Overall, the labels for gender and diagnosis of the WSM Corpus (computed based on the subset of selected annotations) were obtained from annotations with an average inter-annotator agreement of: 0.72 and 0.93 for the self-reported diagnosis label for depression and PD, respectively; and 0.97 and 0.98 for the gender labels for depression and PD, respectively.

It is also interesting to gauge the performance of each annotator and demonstrate the correlation between their performance and average work time (the amount of time spent watching one video and providing their respective annotations), particularly for the self-diagnosis label, which as we have observed, was the most difficult question to answer. In practice, during the annotation of one video, the annotators are not obligated to watch the full video from beginning to end to provide the annotations, and they can watch as much of the video as they find necessary to complete the questionnaire accurately. In order to demonstrate the relation between work time and performance, we will use specifically the annotations related to the self-diagnosis label, as, out of the three, these are the most difficult ones to annotate, and therefore, should be the ones that demonstrate this relationship most clearly.

In Figures 4.3 and 4.4 we plot the annotators average work time against their performance measured in F1 score (computed between the annotator's answers and the aggregated answers obtained via majority voting), for the self-diagnosis label, for the depression and PD datasets, respectively. Each dot represents one worker. We note that the figures only show annotators with more than 7 annotations, to reduce variance. From these figures, we can observe an interesting phenomena that for increasing minimum work times, the performance of the worst annotators improves. This phenomenon is particularly clear in the case of depression. A way to visualize this phenomena is to draw a diagonal line in the plot, and observe that

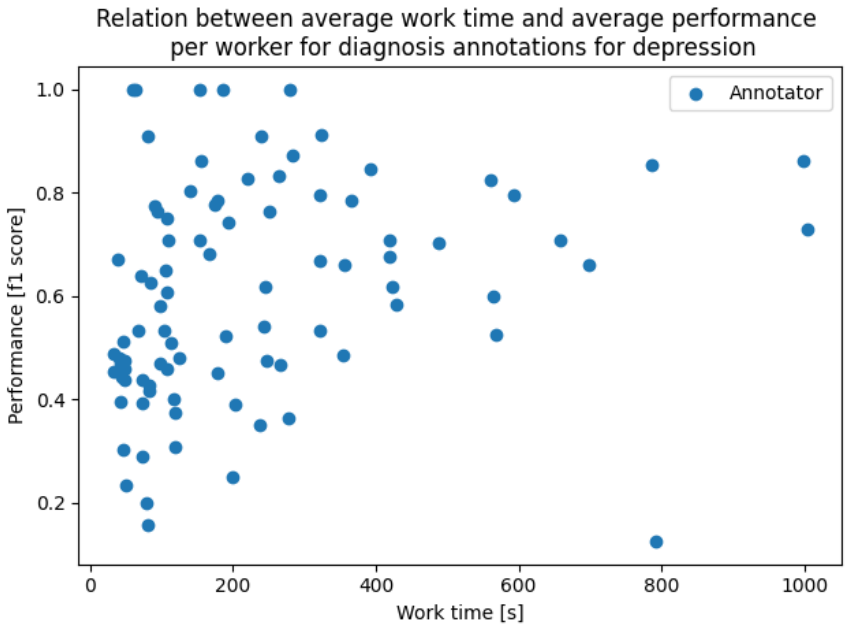


Figure 4.3: Average work time measured in seconds versus average performance measured in f1-score of annotators (each dot represents one annotator), for annotations related to self-reported diagnosis of the depression data of the WSM Corpus v.3.

most workers fall above the diagonal line. However, the reverse relationship (that better performance is related to higher work time) is not valid.

Finally, it is also interesting to compare the distributions of accepted and rejected annotations for the subset of selected videos, per final label, for age, gender and self-reported diagnosis, for depression and PD. With this comparison we hope to further illustrate the difference in difficulty of each of obtaining the annotations for each of the three labels.

Figures 4.5 and 4.6 show the distributions of accepted and rejected annotations for gender, for depression and PD, respectively. Then, Figures 4.7 and 4.8, for the annotations related to the self-reported diagnosis, for depression and PD, respectively. Finally, Figures 4.9 and 4.10 show the distributions of accepted and rejected annotations for each age group, for depression and PD, respectively.

We can observe, in all the above-mentioned figures that the majority of the annotations coincide with the final label, even in the case of the annotations for age (which were not subject to any annotation acceptance criteria). More specifically, from Figures 4.5 and 4.6, the figures relative to the annotations obtained for the gender, we can observe that there is a high level of agreement in both the accepted and rejected answers. This phenomena shows us, once again, that obtaining annotations for gender is an easy task. Furthermore, since both

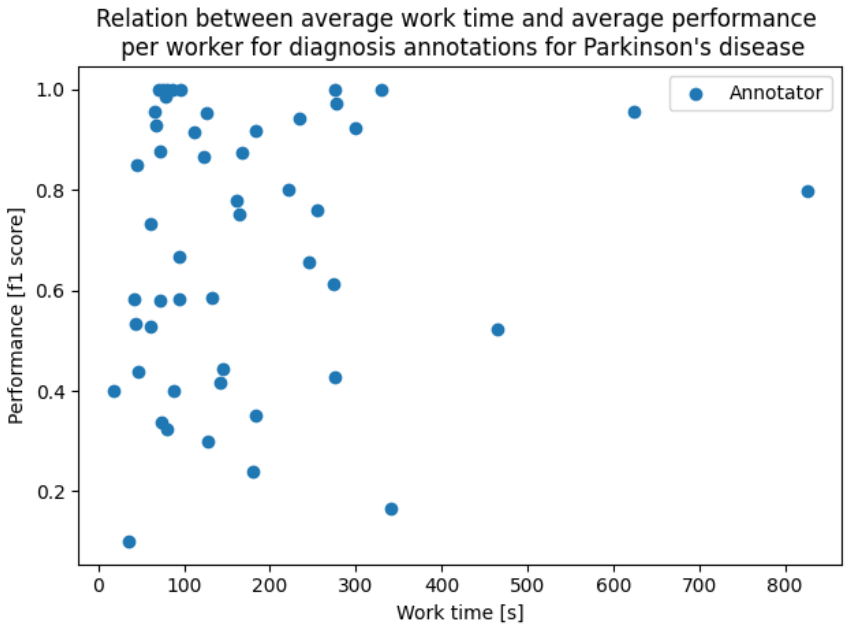


Figure 4.4: Average work time measured in seconds versus average performance measured in f1-score of annotators (each dot represents one annotator), for annotations related to self-reported diagnosis of the PD data of the WSM Corpus v.3.

accepted and rejected annotations present a relatively similar distributions, we can conclude that the rejection of the questionnaires from which these annotations were obtained, did not occur based on the lack of annotator agreement for the estimated gender.

In the case of annotations for self-reported diagnosis, shown in Figures 4.7 and 4.8, we observe a slightly different phenomena, where the distribution of the accepted and rejected annotations have a lower agreement that the ones for gender. This shows that determining the self-reported diagnosis was more difficult to do accurately than determining the gender of the speaker. At the same time, the distributions of accepted and rejected labels differ more than in the case of the gender, which shows that the lower inter-annotator agreement for self-reported diagnosis tended to be a more common criteria to reject questionnaires.

Finally, regarding the annotations for age, shown in Figures 4.9 and 4.10, we can observe that, for the case of depression the distributions of accepted and rejected annotations are similar, for all present age groups. In the case of PD, the accepted annotations tend to slightly more dominated by the most common answer than the rejected ones. This shows that the rejection criteria, based on the remaining questions, indirectly increased the inter-annotator agreement for the question related to age.

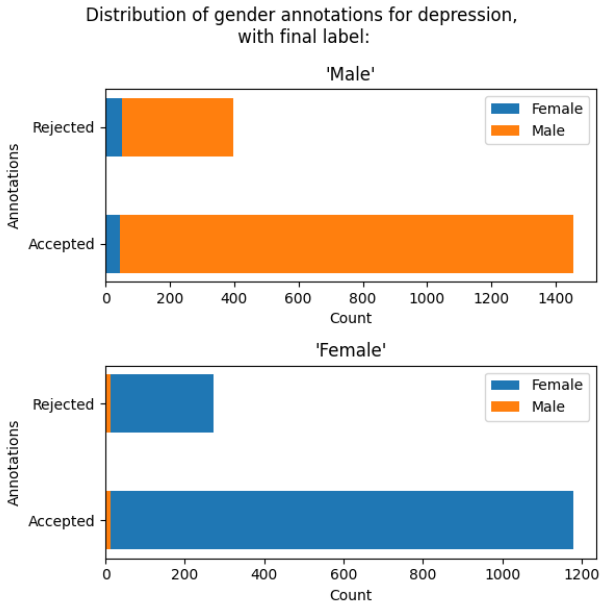


Figure 4.5: Distribution of the accepted and rejected gender annotations for the depression dataset of the WSM Corpus v.3.

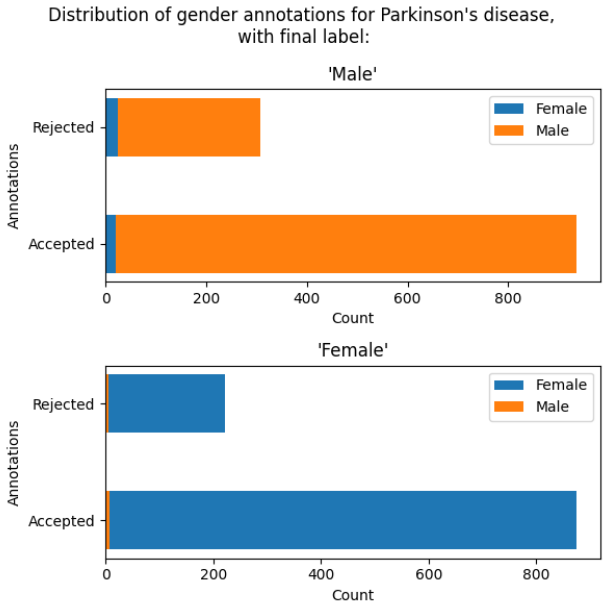


Figure 4.6: Distribution of the accepted and rejected gender annotations for the PD dataset of the WSM Corpus v.3.

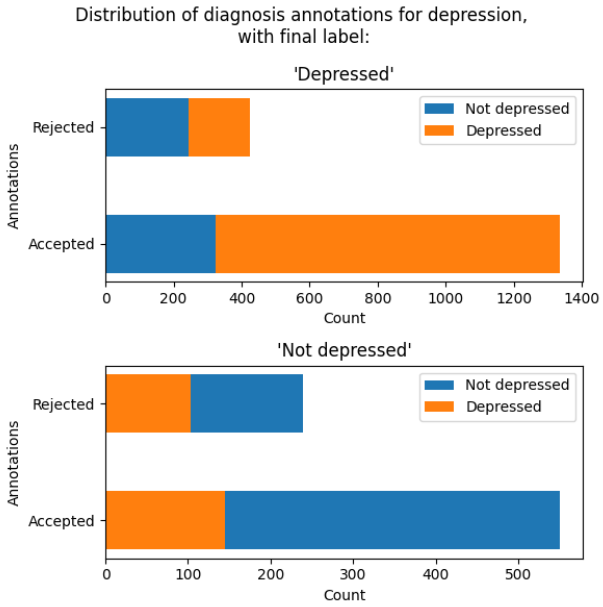


Figure 4.7: Distribution of the accepted and rejected self-reported diagnosis annotations for the depression dataset of the WSM Corpus v.3.

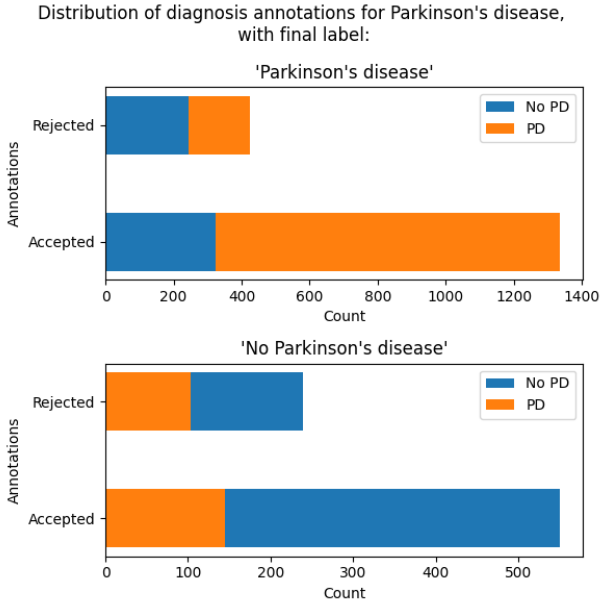
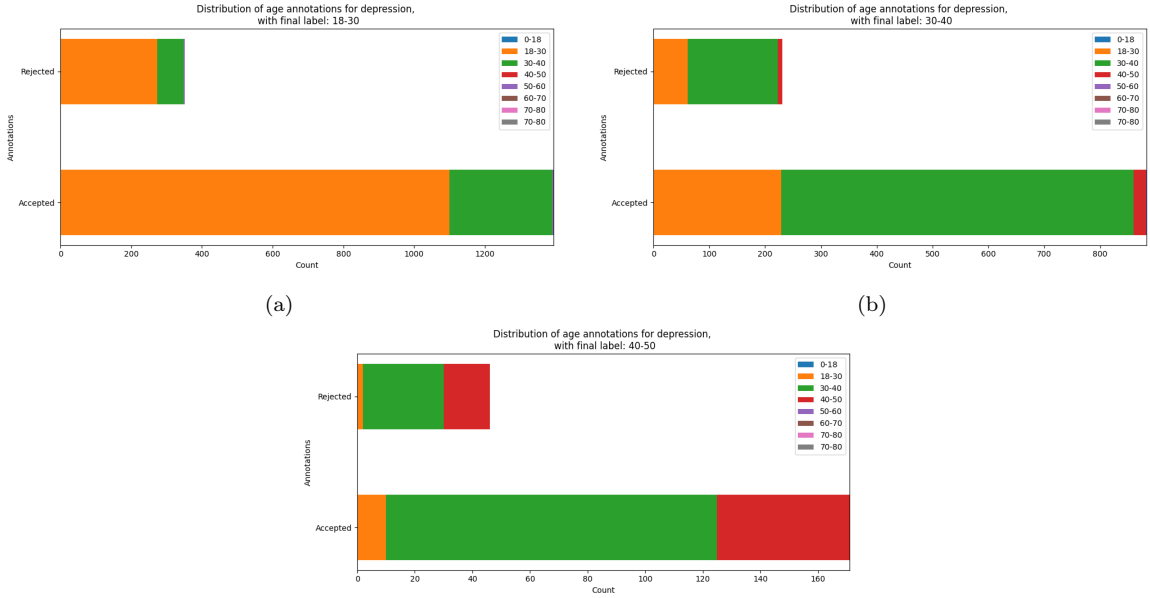


Figure 4.8: Distribution of the accepted and rejected self-reported diagnosis annotations for the PD dataset of the WSM Corpus v.3.



(a)

(b)

(c)

Figure 4.9: Distribution of accepted and rejected annotations per age group for the depression dataset of the WSM Corpus v.3.

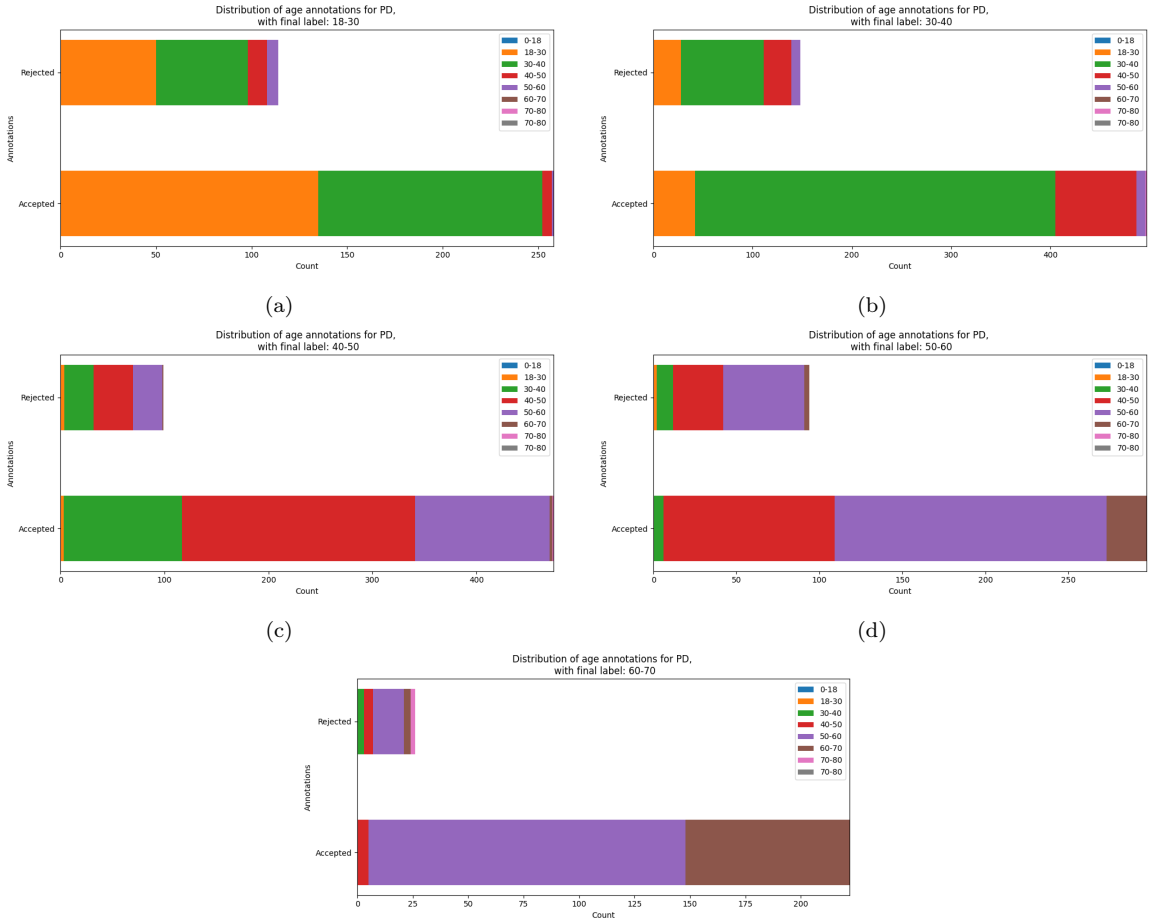


Figure 4.10: Distribution of accepted and rejected annotations per age group for the PD dataset of the WSM Corpus v.3.

Chapter 5

Automatic annotation of speech medical datasets

After having described in detail the collection and manual annotation process of the WSM Corpus in Chapter 4, in its three versions, we move on to explore strategies to automate the annotation process. By developing a successful automatic annotation strategy for these corpora, we are able to significantly reduce the costs typically associated with annotation tasks. In fact, with the constant addition of new videos to the existing online multimedia repositories, having access to an automatic annotation strategy, enables the possibility to continue growing the WSM Corpus indefinitely at no extra cost.

In this Chapter, we will explore several strategies to perform automatic annotation of corpora, particularly the WSM Corpus. It is important to be aware that this task is more specific than separating videos featuring people who self-report to be currently suffering by either depression or PD from videos featuring people who do not. It is doing so in a context where the negative examples are somehow also related to the target disease as well as the positive ones. In most cases, the negative examples belong to one of the following categories: featuring subjects that suffered from the target disease in the past (only in the case of depression), featuring caretakers of subjects with the target disease; featuring healthcare specialists sharing knowledge about the target disease. As such, the strategies presented throughout this Chapter also take the intricacies of the task into consideration.

One of the design choices we made in the work presented in this Chapter, was to choose natural language as the main modality used to automatically annotate the WSM Corpus. The choice of this modality over speech was based on the specific characteristics of the data that were pre-selected for annotation via scrapping tools and implemented retrieval

algorithms: the retrieved examples are mostly of subjects who are aware of their health status, and who are either discussing it or discussing a related topic (as it is also the case of many of the interviews on the DAIC-WOZ). Therefore, we will take advantage of this, and focus on looking for *explicit* cues that the person is affected by depression or PD. As we will show along this Chapter, using the natural language modality allows us to achieve a good annotation quality. Furthermore, there are also practical considerations to take into account: text based information such as the video's transcription, comments, or the metadata of the video, are typically lighter (can be downloaded faster, stored in less space, and be processed faster); versus the audio of the videos, which is heavier (requires downloading the full video and splitting it into video without audio and audio, it takes more storage, and it takes longer to process).

In this Chapter we will present several different solutions to automate corpora annotation, where each strategy has different requirements for the availability of labeled data during training. We begin by exploring the most straightforward case, the fully supervised one, where the training data are labeled. Then we move on to explore a semi-supervised strategy, and finally we adopt a weakly supervised solution for this problem. The later does not require any labeled data during training, and only takes advantage of the existing underlying structure of the data.

In a way, the different strategies represent some of the possible operating points on the scale of the cost of manually annotating none to all of the corpus. Depending on the resources available, one can chose the most appropriate approach. At the same time, all the techniques presented in this Chapter are presented in such a way that they may be easily translatable to corpora annotation problems from other domains.

Specifically, in Section 5.1 we adopt a naive strategy based on fully supervised learning strategies and off-the-shelf tools, in which we establish a baseline for annotating new examples on the WSM Corpus when all the previously existing ones have already been (manually) annotated. We study the influence of a number of features derived from the transcription and the metadata of the video.

Then, in Section 5.2 we explore a greedy semi-supervised strategy to automatically annotate the WSM Corpus. This strategy attempts to partition the unlabeled subset of the data into positive and negative examples, such that the performance of a model trained on these data and labels, evaluated on the labeled subset of the corpus, is optimized. This strategy depends on the existence of a small "gold" dataset containing manual annotations.

Finally, in Section 5.3 and 5.4, we propose more sophisticated strategies, more specifically,

we propose two weakly supervised learning strategies based on original generalizations of the multiple-instance learning (MIL) framework, that take advantage of the data structure implicit in the videos collected from YouTube, and do not depend on any manual labeling: θ-MIL and deep-θ-MIL. Specifically, the fact that each video is associated with the search term that was used to retrieve it, and the hypothesis that the set of videos retrieved with the same search term is made up of videos that have some sort of commonalities among themselves. θ-MIL is based on the application of SVMs for MIL, and deep-θ-MIL is an implementation of a similar solution with neural networks. We test the proposed deep-θ-MIL solution to annotate the WSM Corpus v.2 in Section 5.5.

5.1 Leveraging from transcriptions and metadata in a fully supervised context

In this Section we will show some experiments, first reported in [109], where we attempted to annotate the WSM Corpus (v.1) by finding videos deemed "ideal candidates". This means that they verify the following characteristics: be in the vlog format, featuring a single subject; who is typically talking about themself; and typically referring to present and not past events, situations, or emotions; and which has to verify all the prior labels at the same time, which equated to explicitly self-report their current health status of depressed or suffering from PD.

This work was done as a proof-of-concept to show a possible avenue to annotate new examples, and automatically grow the WSM Corpus with a fully supervised learning setting as the starting point.

We took advantage of the binary labels for which the corpus is annotated (the video is in a vlog format; regardless of the topic, the main speaker of the video talks mostly about themself; regardless of the topic, the main subject is speaking mostly about present events/opinions/situations/*etc.* or from a recent past (over the last few hours/days); the main topic of the video is related to the [*target disease*]; and finally, the subject claims to be currently affected by the [*target disease*]) to find these "ideal candidates".

As mentioned, we worked on a fully supervised learning scenario, where the modeling strategies we adopted were Logistic regression (LR), and SVMs, and where the the input to these models was a set of features and descriptors extracted from the transcription, title, description, top comments, and metadata of the videos.

We trained a separate model for each binary label and for each type of feature in order to study their different contributions to find the target videos.

Again, we emphasise that we opted for simple, straightforward techniques, both for the feature extraction stage as well as for the modeling stage, using mostly off-the-shelf tools in order to present a modular baseline in which the features or the modeling techniques can be easily changed. We deferred replicating state of the art techniques used to solve related problems, including multimodal emotion recognition [110][111][112], and techniques that perform the synchronization of the features across different modalities [113][114][115] to future work.

5.1.1 Feature extraction

From the information extracted for each video, we computed the following features derived from the video's transcription, title and description as provided by the uploader:

- **Bag-of-Word (BoW) features:** We extract these from the video's transcription. The BoW model was used to convert a transcription in to a frequency vector of tokens in the transcriptions. In this scheme, we obtained one feature vector per transcription, in which each feature was the normalized frequency of an individual token. The length of the vector was the total size of the vocabulary of the corpus of transcriptions. This model ignores the ordering of the tokens in the transcription. In order to reduce the weight of very common words, (e.g. "the", "a", "is" in English), which carry very little meaningful information about the actual content of the document, we used the term-frequency times inverse document-frequency (tf-idf) transform.

- **Sentiment features:** We derived these from the title, description, transcription and top n comments of the video using the Stanford Core NLP tool [116]. This tool is based on a Recursive Neural Tensor Network (RNTN). RNTNs take as input phrases of any length, and represent them through word vectors and a parse tree. They then compute vectors for higher nodes in the tree using the same tensor-based composition function. This RNTN was trained on a corpus of movie reviews [117], and parsed with the Stanford parser [118].

- **Metadata:** We extracted features that were derived from the collected metadata including: a one hot vector representing the video category out of fourteen possible categories; the video duration; the number of views; the number of comments; the number of thumbs up; and the number of thumbs down at the time of collection.

At this early stage, and given the limited size of the WSM Corpus v.1, we did not include topic modeling, nor semantic word embedding models.

5.1.2 Classifiers

We used two alternatives to predict each of the five binary labels of the videos in the WSM Corpus: LR, and SVMs. For the case of the SVM we trained 3 distinct models with linear, polynomial of degree 3, and RBF kernels.

5.1.3 Datasets

For the experiments reported in the following Section, we used the labeled examples of the depression and PD subsets of the WSM Corpus v.1, as described in Section 4.2. These subsets contained 58 and 61 examples for depression and PD, respectively. Out of which 18 and 18 were positive for the self-reported health status of depression and PD, respectively. The distribution of the remaining labels was reported in Table 4.1.

5.1.4 Experiments and Results

The training and evaluation of the proposed pipeline for each SA disease was accomplished via leave-one-out cross validation, *i.e.*, the model was trained n times, each time with $n-1$ examples as the training data, and the remaining example is left to evaluate the model. This process is repeated, always leaving a different example for evaluation, such that, by the n^{th} time, each example is used once during the evaluation. The final performance is the average of all the performances.

In order to understand the contribution of each type of feature to detect each of the five binary labels, we trained a distinct classifier for each type of feature, and another one with all the features.

The BoW features contributed with feature vectors of dimension 5096 and 5849 for depression and PD, respectively; the sentiment features extracted from the title, description, transcription and comments of the video contributed with a 28 dimensions feature vector (7 dimensions each); the metadata contributed with 19 dimensional feature vector. By concatenating features extracted from all modalities, the final feature vectors had 5914 and 6667 dimensions for depression and PD, respectively.

For practical reasons, given the limited amount of examples in our datasets, and the comparatively large number of features, the feature vectors were reduced in dimensionality by eliminating the features that had a Pearson correlation coefficient (PCC) with the corresponding label below 0.2, thus only the features that carried some linear correlation to the label were preserved.

In total, 160 models were trained: LR, linear SVM, polynomial SVM, and SVM with RBF kernel, for each one of the three types of features, and an additional one for the concatenation of all the features, for each of the five labels, for the two SA diseases (depression and PD). The models were trained in a leave-one-out cross validation fashion.

The results are reported in precision and recall. We considered that a good model would have at least a high precision measure, since we consider that it is more important to maximize the rate of true positives. At the same time, false negatives are also a concern but to a lesser extent, in this scenario: we assume that the repository being mined has a much larger number of target videos than the size of the desired dataset.

Tables 5.1 and 5.2 summarize the performance of the best overall model (SVM-RBF), for depression, and PD, respectively. The results of the remaining models are omitted, for the sake of brevity. The cells highlighted in gray mark models which performed equal or worse than simply choosing the majority class. The best performing models for each dataset achieve a 86%, and 100% precision, and 72%, 89% recall, for the depression, and PD subsets of the WSM Corpus v.1, respectively.

These Tables also show the contribution of each type of feature to the overall performance, as well as the performance of the model in identifying each label correctly. The type of features that had the most impact were the BoW features, for every SA disease, and for every label. They conveyed, in fact, sufficient information to achieve the best performance, without any other type of feature. The sentiment features were also capable, albeit to a lesser extent, to correctly detect all the labels for PD, but not for depression. Finally, the metadata features were only able to contribute to the detection of one labels for both depression and PD. The performance of each feature corresponded to our intuition that the features based on the transcriptions would be the most relevant to solve this task.

The hardest label to correctly predict was consistently the "present" one, which refers to videos describing mostly present and not past events, situations, or emotions.

It was an interesting and counter-intuitive result to observe that the sentiment features were only useful in the tasks related to PD, but not for depression.

Overall, with this set of experiments we were able to demonstrate that it is possible to annotate the WSM Corpus, using only simple off-the-shelf tools, and simple ML strategies, when in a fully supervised scenario.

Table 5.1: Performance of the SVM-RBF reported in precision and recall rate in detecting target content in the depression dataset of the WSM Corpus.

Modality	Features	Label				
		Vlog	**1st Person**	**Present**	**Target topic**	**All**
Text	BoW	0.98 / 1.00	0.98 / 1.00	0.73 / 0.94	0.89 / 0.89	0.86 / 0.67
	Sentiment	0.91 / 1.00	0.77 / 0.96	0.52 / 0.66	0.52 / 0.71	0.33 / 0.17
Metadata	Metadata	0.91 / 1.00	0.77 / 0.98	0.62 / 1.00	0.60 / 0.97	0.00 / 0.00
All	All	0.981 / 1.00	0.93 / 0.96	0.83 / 0.91	0.89 / 0.91	0.86 / 0.67

Table 5.2: Performance of the SVM-RBF reported in precision and recall rate in detecting target content in the PD dataset of the WSM Corpus.

Modality	Features	Label				
		Vlog	**1st Person**	**Present**	**Target topic**	**All**
Text	BoW	1.00 / 0.86	0.74 / 0.82	0.81 / 1.00	0.91 / 1.00	1.00 / 0.89
	Sentiment	0.71 / 0.71	0.69 / 0.71	0.77 / 0.49	0.73 / 0.95	0.88 / 0.39
Metadata	Metadata	0.73 / 0.77	0.49 / 0.76	0.56 / 0.77	0.70 / 0.98	0.00 / 0.00
All	All	0.97 / 0.91	0.87 / 0.82	0.80 / 0.91	0.90 / 1.00	1.0 / 0.89

5.2 Greedy set partitioning for corpora annotation

The second proposed strategy to automatically annotate the WSM Corpus was based on a greedy semi-supervised approach that depends on the existence of a small "gold" dataset with a ground truth. The work presented in this Section is based on a prior work [119]. In this Section, we assume that we are working with a partially annotated dataset, instead of a fully annotated dataset, as was the case in Section 5.1, and that the objective is to label the unlabeled portion of the dataset. We wanted to do this, such that the resulting partitions of the unlabeled data into positive and negative subsets, optimizes the performance of a binary classifier trained on it, and evaluated on the labeled portion of the dataset.

The proposed framework is based on the following steps, which are explained in more detail below. First, we used a small labeled portion of the corpus to train a binary classifier that detects the target videos; we then used it to predict the labels of the remaining unlabeled portion of the corpus. On a second stage, we used the unlabeled portion of the corpus and predicted labels estimated in the previous step to train a noisy model. Finally, the noisy model's performance was evaluated against the labeled portion of the corpus, attempting to reconstruct the original labels. While we could not evaluate the exact performance of the original model in predicting the labels of the unlabeled portion of the corpus, we could use the reconstruction rate of the noisy model as an estimate.

We tested this framework on the WSM Corpus v.1, including the unlabeled examples, and on the DAIC-WOZ, for which we could actually evaluate the performance of both the base and noisy models, since there is a ground truth available for the later.

We addressed the classification problem using multiple SVMs trained on BoW features computed from the transcriptions of the videos. The features and classifiers chosen were disease agnostic, so as to make this approach trivial to apply to corpora of other SA diseases.

5.2.1 Proposed framework

The problem we are faced with is a usual variant of semi-supervised learning. We are given a dataset $\mathbf{X} = \{X_L, X_U\}$, composed of two subsets: X_L, of labeled examples with the corresponding Y_L labels, and X_U, for which we have no labels. We must combine the two datasets effectively to train a classifier.

Conventionally, in semi-supervised learning the *labelled* data is used as the training data to learn parameters, and the *unlabelled* data as augmentation data to improve the learned parameters. In our solution, we cast the problem differently: We start by using the labelled data to train a base model that yields noisy labels for the unlabelled data. Then, we train a new model from scratch using the unlabelled data, *i.e.*, we treat our unlabelled data as our primary training data that parameters will be learned from. The labelled data are, finally, treated as validation data to evaluate the new model learned.

In this framework, in order to train a model, we require label estimates of the training data, X_U, $\hat{Y}_U$. We treat the identification of labels as a set partitioning problem. For a binary classification problem, our goal is to find the partition P that separates the data in the set of positive and negative examples: $P_{X_U} = X_{U+}, X_{U-}$, such that, for any classifier F_{noisy}, trained on X_U and $\hat{Y}_U$, the labels predicted for X_L, $\hat{Y}_L = F_{noisy}(X_L)$, are as similar as possible to Y_L. This idea can be formalized as follows:

$$P^* = argmin_P(L(Y_L, \hat{Y}_L)), \tag{5.1}$$

where $\hat{Y}_L$ is obtained from a generic model F_{noisy}, and L is a generic loss function.

There are many approaches that can be used to solve the partitioning problem, and even brute force is an option. In this work, we propose a greedy solution, based on SVMs with RBF kernel, that estimates the partition P for X_U, and consequently the labels $\hat{Y}_U$, by training a base model F_{base} with X_L and Y_L, which we then use to predict the labels for X_U by computing: $\hat{Y}_U = F_{base}(X_U)$. Using the unlabeled subset of the corpus, X_U, and the

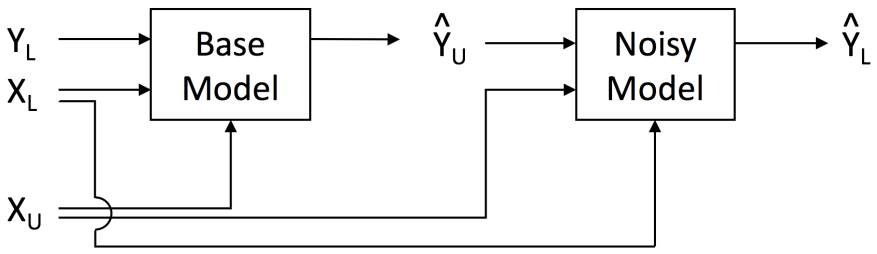


Figure 5.1: Proposed framework, using base and noisy model, to reconstruct labels of the labeled subset of a corpus and estimate labels for the unlabeled subset of a corpus.

respective noisy predictions computed before, $\hat{Y_U}$, we train a noisy model F_{noisy}. This model, F_{noisy}, is used to estimate the quality of the previously estimated partition P and labels $\hat{Y_U}$, by computing the loss $L(Y_L, \hat{Y_L})$, following Eq. 5.1.

The proposed two-step solution based on SVMs is summarized in Figure 5.1.

5.2.2 Feature extraction

We computed BoW features from the transcriptions of the videos, whether they were manually or automatically obtained. The BoW model, as explained before in Section 5.1, converts the transcription into a frequency vector of tokens in the transcriptions. Using this scheme, we obtained one feature vector per transcription, in which each feature was the normalized frequency of an individual token. The length of the vector was the total size of the vocabulary of the corpus of transcriptions. This model ignored the ordering of the tokens in the transcription. In order to reduce the weight of very common words, we used tf-idf transform.

5.2.3 Datasets

In the experiments described in the following Section we used the DAIC-WOZ (train and development), described in detail in Section 2.2, and the depression subset of WSM Corpus v.1.

The DAIC-WOZ was partitioned into two subsets with equal size of 71 interviews (not the partition originally suggested by the authors), where one was used to train the base model and the other to train the noisy model. Since all the examples of this dataset are labeled for depression, it can be used to evaluate the performance of the noisy models.

As for the depression subset of the WSM v.1, it contained 58 annotated videos for the self-reported health status of currently affected by depression, out of which 18 were positive.

The remaining 119 examples were unlabeled. The labeled examples were used to train the base model, and the unlabeled ones to train the noisy one.

We note that, at the time of this work, we did not have the HCs for the New Spanish Parkinson Corpus yet, as it was originally made available in the context of a PD severity assessment task, where all the subjects were known to suffer from PD. Therefore, we could not evaluate this approach on PD, as we have done with the experiments reported in other Sections.

5.2.4 Experimental results for the base model

To train the base models, we used a subset of 58 and 71 labeled examples from the WSM Corpus and DAIC-WOZ, respectively.

We also note that, since the subset of the DAIC-WOZ that was used to train the base model does not correspond to the train partition originally designed by the authors of the dataset, the results reported in this Section are not directly comparable to other works. In this work, we opted to shuffle the original train and development partitions of the corpus, and use one half of the total number of samples as the labeled subset and the other half as the unlabeled subset, so that the noisy model, trained on the second half subset, would have enough train examples.

The chosen parameters for the models were set as follows:

The SVMs using BoW features were trained using linear, polynomial of degree three, and RBF kernels, however we only report the best result, for brevity. The parameter C that penalizes error term on the cost function was set to 10.

Furthermore, we opted to reduce the dimensionality of the BoW feature vectors, given the limited amount of training data. The dimensionality reduction was achieved by computing the Pearson correlation coefficient of the features of the train examples to the respective labels and only keeping those with a coefficient over 0.3.

We report the performance of the models against the train data, and, for the DAIC-WOZ which has the ground truth labels available for the whole corpus, we report the performance against the remainder 71 unseen examples of the dataset, as can be seen in Table 5.3.

The results are reported in terms of UAR. This metric is defined as the unweighted average of the class-specific recalls, and is computed as follows:

Table 5.3: Performance, in UAR, of the base models trained on the labeled subsets of the WSM corpus and the DAIC, using BoW.

Modality	Features	Model	WSM		DAIC-WOZ	
			Train	Test	Train	Test
Text	BoW	SVM	1.000	N/A	0.972	0.570

$$UAR = \sum_{c=0}^{C} \frac{1}{C} \frac{TP_c}{TP_c + FN_c}, \tag{5.2}$$

where C is the number of classes, TP_c is number of true positives for class c and FC_c is the number of false negatives for class c. A UAR of $\frac{1}{C}$ is achieved by voting according to the prior probabilities of the classes. A higher UAR corresponds to a better performance, and a perfect performance corresponds to a UAR of 1. UAR is especially useful to report results for classification tasks with unbalanced data, rather than weighted average recall.

As can be seen from Table 5.3, models achieve a high UAR on the training data, an indication that the models were able to obtain some sort of information from the training data. When testing the models trained on DAIC-WOZ, we observed that the SVM with the BoW achieved a UAR of 0.570. We reiterate that since the unlabeled subset of the WSM Corpus has no ground truth, it was impossible to evaluate the base models trained on the WSM Corpus on test data.

5.2.5 Experiment results for the noisy model

The base models were used to predict labels for the respective 119 and 71 unlabeled examples of the corpora of the WSM Corpus and DAIC-WOZ corpora.

Using the noisy predictions and the unlabeled subsets of the corpora we computed noisy models for each corpus, analogous to the ones described before. The model parameters and architectures were the same as before.

The performance of the models was evaluated against the noisy predictions of the training data and, more importantly, against the labeled subsets of the WSM Corpus and DAIC-WOZ, which were unseen data for the noisy models. Their performance is summarized in Table 5.4, and is reported in UAR.

Table 5.4 reports the reconstruction rate, i.e. the capability that the noisy models have to correctly estimate the original labels of the labeled subsets of the corpora, in the second and

Table 5.4: Performance, in UAR, of the noisy models trained on the unlabeled subsets of the WSM corpus and the DAIC and respective noisy predictions estimated by the respective base models, using BoW.

Modality	Features	Model	WSM		DAIC-WOZ	
			Train	Dev	Train	Dev
Text	BoW	SVM	0.981	0.969	1.000	0.917

fourth columns, for the WSM Corpus and the DAIC-WOZ, respectively. We observe mostly good reconstruction rates. Furthermore, by comparing the fourth column of Tables 5.3 and Table 5.4, we can observe a possible correlation between the performances of the noisy models on unseen data and the base models for the DAIC-WOZ. From the comparison of these two columns we can hypothesize a similar relation between the performances of the noisy and base model trained with DAIC-WOZ, thus assuming a reasonable performance of the base models in estimating the labels of the dataset, even with so few examples to learn from.

This semi-supervised greedy approach could be further improved by using the labels estimated by the base model as the initialization for an iterative algorithm that would further optimize the partition of the unlabeled subset into positive and negative samples.

5.3 Generalizing the Multiple Instance Learning framework in a semi supervised context

In this Section we will explore a weakly supervised learning strategy to automatically annotate the WSM Corpus without the requirement of having any annotated data during training. Instead, we will take advantage of the existing underlying structure of the data to impose some constraints on the problem.

We will begin by describing the existing structure of the WSM Corpus, to motivate our solutions. After that, we move on to introduce the MIL frameworks, as well as its generalized version, θ-MIL, proposed in previous works [120]. Additionally, we proposed two solutions for the θ-MIL problem, the first based on SVMs, and the second on NNs. The solution based on SVMs is detailed in this Section, and the one based on NNs, in Section 5.4. Finally, we use the NN-based solution for θ-MIL to annotate the WSM Corpus v.2, and report the obtained results in Section 5.5.

The SVM based solution for the θ-MIL problem was tested in the context of inferring the polarity of written movie reviews, where the bags were considered to be the sets of reviews

for a given movie, and the bag labels were the binarized score of the movie. The task was to identify which reviews were positive or negative without access to any information, other than the movie scores (the bag labels) and the reviews for each movie (the instances, and their bag assignment). However, since this task is not directly relevant to the scope of this thesis, but the technical details are, we opted to describe them in the body of this document, and relay the experimental results to the Appendix B, where we experimentally confirm the usefulness of the proposed technique.

5.3.1 Underlying structure of the WSM Corpus

The videos of the WSM Corpus, as explained in Chapter 4, were pre-selected with either *relevant* or *irrelevant* queries to the target SA disease. For example, in the case of depression, the candidate videos were selected by querying the YouTube API for the relevant query "depression vlog", along with a time window. This process was repeated for several time windows. As a result, we obtained, for each combination of query and time window, a set of up to 50 videos, that are the top ranking results, within the given time windows, according to YouTube's recommendation algorithm. Intuitively, we can expect that the retrieved videos are either related to the concepts of "depression", "vlog", or both. However, while all the videos may relate to these concepts, not all of them verify the conditions to be a target video in the WSM Corpus. We recall that in order for a video to be included as a target video of the depression dataset of the WSM Corpus, it must contain a subject self-reporting as currently suffering from depression. From prior experiments, and also according to the small annotation task reported in 4.2, we know that about one third of the results from the query "depression vlog" consist of target videos for depression. The remaining videos contain subjects that are, for example, caregivers, partners, or other family members of people suffering from depression; people who have suffered from depression in the past, but not currently; medical specialists discussing some aspect of depression; or other videos.

Conversely, when querying the YouTube API for terms unrelated to the target SA disease, such as "vlog" of "book review", and a time window, it is reasonable to expect that either very few, or none, of the retrieved videos contain a positive self-reported health status for that target disease.

We illustrate, in Figure 5.2, how the results of the sets of results are organized based on queries.

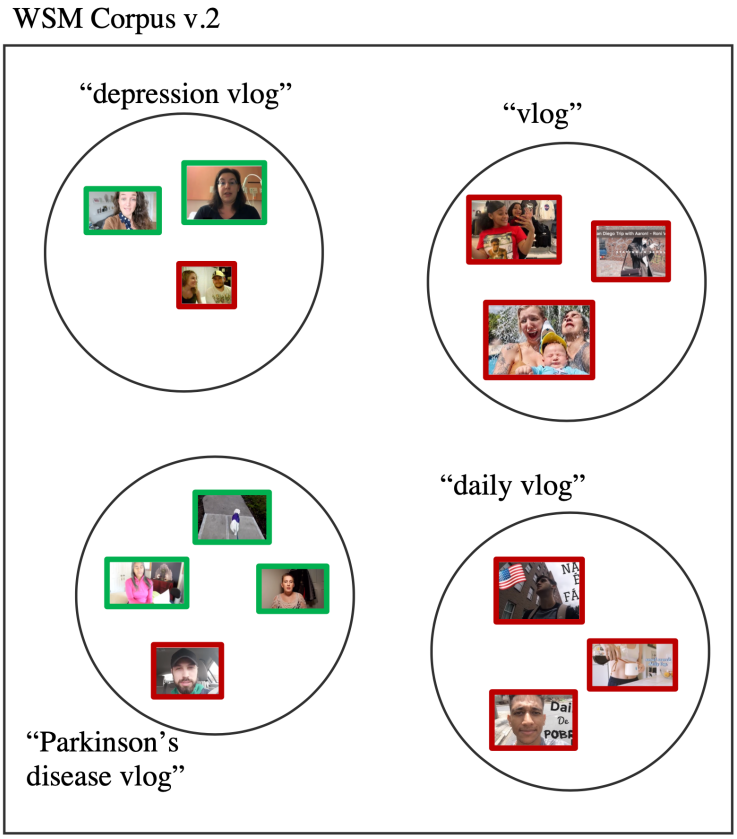


Figure 5.2: Example of the natural bag organization of videos retrieved with a given query. Circles represent a set of results for the query above the respective circle. Videos outlined in green contain a positive self-reported health status for the target SA disease, and red ones do not.

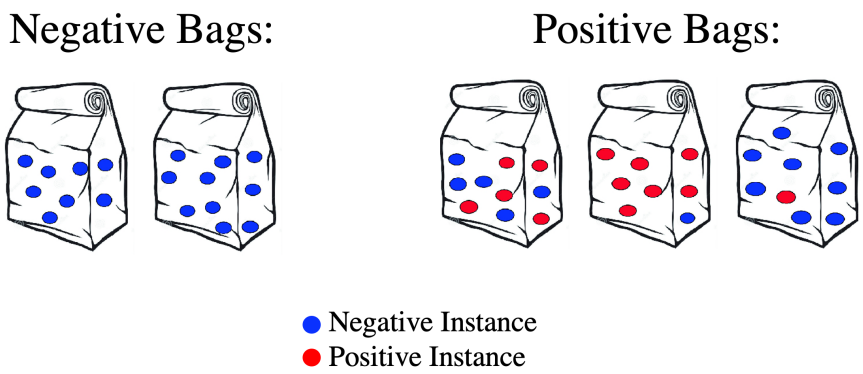


Figure 5.3: Illustration of the label assumptions under the MIL framework. Adapted from [1].

5.3.2 Multiple Instance Learning

MIL [121] is a generalization of fully supervised learning problems, where the training examples, called *instances*, are grouped into sets, called *bags*, and, at train time, only the labels of the bags are known, and the instance labels remain hidden. The main assumption of the MIL framework is that bags with a negative label only contain negative instances, while bags with a positive label contain at least one positive instance. This is illustrated in Figure 5.3.

Another way to explain the MIL framework is with a recurring analogy, that compares a bag with a set of keys, where the keys correspond to the instances: Given a set of keys and the information that we can use that set of keys to open a given lock, we will know immediately that at least one of the keys in the set will open the lock, but not necessarily how many, or which one(s). If we are told the opposite, we will know immediately that none of the keys open the lock.

In the context of the WSM Corpus, if we want to formulate it using MIL, we can think of the sets of results for a given query as the bags, and the videos as the instances. Given a target SA disease, we can expect the sets of results obtained with the relevant queries to contain some positive instances, *i.e.* target videos, although which and how many is unknown without manual annotation. Conversely, for sets of results obtained with queries unrelated to the target SA disease, we can expect that very few or none of the videos will be target videos.

Other than in the context of the WSM Corpus, MIL has been used in several other tasks such as medical imaging segmentation [122][123][124], where an image is typically described by a single label, but the region of interest is not given. In this case, the bag is the image and each segment of the image is an instance. Other examples include drug activity prediction [121], image annotation and retrieval [125], text categorization [126], and object detection

[127], among others.

There are multiple approaches to address the MIL problem. Arguably, one of the early, most popular ones is [2]'s solution, based on SVMs, that describes two algorithms that formulate the MIL problem as a maximum-margin problem. Then, this problem can be solved via mixed either of two integer quadratic programs: mi-SVM and MI-SVM. Both will be explained in furhter detail in Section 5.3.4.

Other, more recent works have proposed solutions for the MIL problem via deep neural networks, as is the case of the pioneering work of [128], and our own previous work [129]. Others, such as [124], proposed a formulation for the MIL problem as learning the Bernoulli distribution of the bag labels, which is parametrized by a neural network with an attention mechanism.

5.3.3 Intuition for generalizing the Multiple Instance Learning framework

The MIL framework has its limitations. Arguably the most significant one is its sensitivity to positive instances, in the sense that even one positive instance in a bag is enough to flip the bag's label. This can be an issue in certain domains where the labels can be noisy. Even in the context of the WSM Corpus, there is no guarantee that negative bags will not contain negative instances, *i.e.*, that queries unrelated to the target SA disease will not return videos with positive self-reported health status ever. The solution to this limitation can be found by generalizing the MIL framework, such that the generalization allows the user to specify how many instances, or which fraction of the instances in the bag, needs to be positive for the bag to be assigned a positive label. In this case, a bag would be positive if more than a fraction of the instances in the bag were positive, and negative otherwise, where this fraction could be determined either by the user or the problem's intrinsic constraints. This is illustrated in Figure 5.4.

5.3.4 MIL formulated as a maximum margin problem

The main concepts of MIL have been explained from a high level perspective in the previous Section. We will now formalize it, and compare it to fully supervised learning. In the later scenario, particularly in binary classification, it is necessary to have pairs of instances and labels, $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \rightarrow \mathcal{Y}$, generated independently from an unknown distribution, where $\mathcal{Y} = \{-1, 1\}$, as an example. In the MIL scenario, this problem is generalized by the ambiguity in the labeling of the instances. Instances, $\mathbf{x}_1, ..., \mathbf{x}_n$, are grouped into *bags*, $\mathbf{B_1}, ..., \mathbf{B_m}$, with

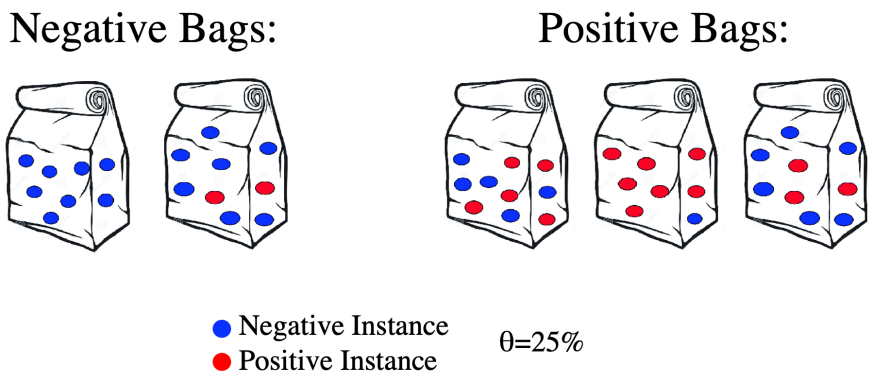


Figure 5.4: Illustration of the label assumptions under the generalized MIL framework, with the threshold of positive examples before the bag becomes positive, θ, set to 25%.

$\mathbf{B}_I = \{\mathbf{x}_i : i \in I\}$, for non-overlapping $I \subseteq \{1, ..., n\}$. Each bag, $\mathbf{B}_I$, is associated to a label, $\mathbf{Y}_I$. If $\mathbf{Y}_I = 1$, then there is at least one bag instance, $\mathbf{x}_i \in \mathbf{B}_I$ with $y_i = 1$, or if $\mathbf{Y}_I = -1$, then all $\mathbf{x}_i \in \mathbf{B}_I$ have $y_i = -1$.

One of the possible solutions for MIL is to formulate it as a maximum margin problem, which can then be solved by extensions of SVMs [2]. In [2], the authors propose two such approaches: the first treats the instance labels as unobserved integer variables, subject to the constraints defined by the positive bag labels; the second generalizes the notion of a margin from instances to bags and aims to maximize the bag margin.

In more detail, the first approach, *mi*-SVM, can have its mixed integer formulation of MIL as a generalized soft-margin SVM, and its primal form can be written as the following optimization problem:

$$\begin{aligned} \min_{\{y_i\}} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum_i \xi_i \\ \text{s.t.} \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \\ & y_i \in \{-1, 1\} \\ & \sum_{i \in I} \frac{1 + y_i}{2} \geq 1, \forall I \text{ s.t. } Y_I = 1 \\ & y_i = -1 \forall I \text{ s.t. } Y_I = -1 \end{aligned} \tag{5.3}$$

where the optimization variables $\mathbf{w}$, b, ξ_i, and y_i, are, respectively, the weight vector, a scalar, a scalar, and the predicted instance label for example i. Y_I is the bag label for bag I, and C

is a hyperparameter.

The second approach, *MI*-SVM, is formulated as a quadratic mixed integer problem, as follows:

$$\begin{aligned}
&\min_{s}\min_{\mathbf{w},b,\xi}\frac{1}{2}||\mathbf{w}||^2 + C\sum_I \xi_I \\
&\text{s.t.} \\
&\forall I : Y_I = -1 \wedge -\langle \mathbf{w}, \mathbf{x}_i \rangle - b \geq 1 - \xi_I \\
&\text{or } Y_I = 1 \wedge \langle \mathbf{w}, \mathbf{x}_{\mathbf{s}(I)} \rangle + b \geq 1 - \xi_I \\
&\xi_I \geq 0
\end{aligned} \tag{5.4}$$

where the optimization variables $\mathbf{w}$, b, ξ, and s, are, respectively, the weight vector, a scalar, a scalar, and the instance selector. Y_I is the bag label for bag I. Note that $s(I)$ acts as a selector among the instances of a bag. It will be active for one instance in each positive bag. C is a hyperparameter.

In this case, the positive bag margin is defined by the margin of the "most positive" instance.

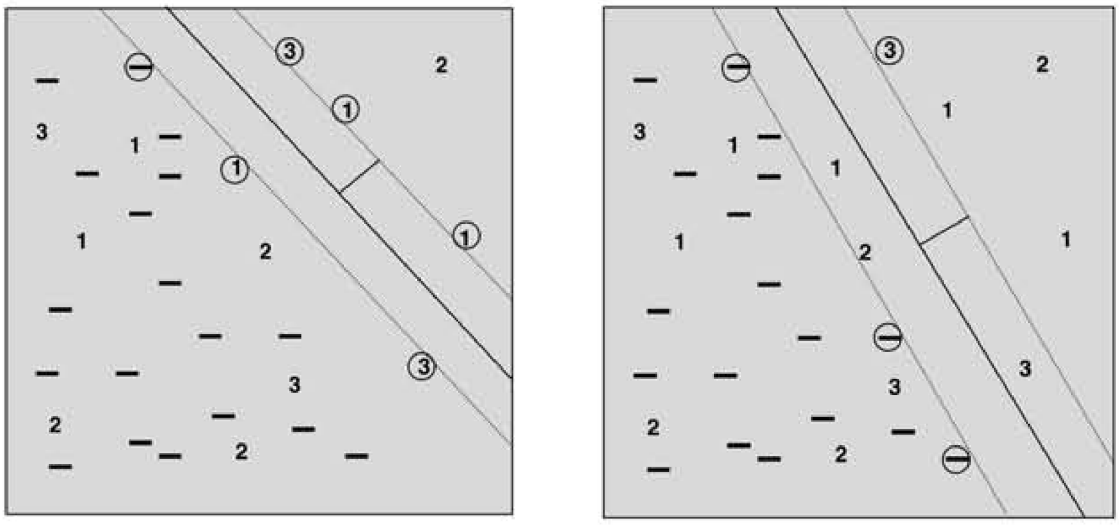
The difference between the two approaches is essentially that in MI-SVM, the negative instances of the positive bags are ignored, and at the same time, only one instance per positive bag contributes to the optimization problem. On the other hand, in mi-SVM, negative instances in positive bags, as well as one or more positive instances from a positive bag can be support vectors. This is illustrated in an example in Figure 5.5.

5.3.5 θ-MIL

In order to generalize the MIL framework, we can view it as a particular case of a more generic problem. An alternative for the assumptions regarding bag and instance organization would be to associate to each bag, $\mathbf{B}_I$, a label $\mathbf{Y}_I$, given by:

$$\mathbf{Y}_I = sign(\frac{\sum_i (y_i : i \in I + 1)}{2|I|} + \theta), \tag{5.5}$$

where $\theta \in [0, 1]$ is the minimum fraction of instances in the bag that has to be positive. That is, in this generalized version of MIL, θ-MIL, the label of a bag is positive if more than a fraction θ of the instances in that bag are positive, otherwise the bag is negative. This has been illustrated before in Figure 5.4. To reduce θ-MIL to the traditional MIL framework, one has simply to make $\theta = 0$.



(a) mi-SVM solution (b) MI-SVM solution

Figure 5.5: mi-SVM (left), and MI-SVM (right) solutions to an arbitrary MIL problem, where negative instances are denoted by "-" and positive instances by a number encoding their bag membership. Adapted from [2].

The solution for θ-MIL can also be formulated as a maximum margin problem, as was [2]. In fact, both mi-SVM and MI-SVM can be adapted to a more general formulation to verify the conditions of θ-MIL as follows.

Given a set of bags $\mathbf{B}_I$, their labels $\mathbf{Y}_I$, and the instances of each bag, $\{\mathbf{x}_i : i \in I\}$, the optimal class separating hyperplane with parameters $\mathbf{w}$ and b, and instance labels $\{\mathbf{y}_i : i \in I\}$, can be found by minimizing the same objective as the Optimization Problem 5.3, subject to two new constraints: $\sum_{i \in I} \frac{1+y_i}{2|I|} \geq \theta, \forall I s.t. Y_I = 1$, and $\sum_{i \in I} \frac{1+y_i}{2|I|} < \theta, \forall I s.t. Y_I = -1$. More formally, the adaptation of the mi-SVM, to which we will refer to as θ-mi-SVM, can be written as:

$$
\begin{aligned}
\min_{\{y_i\}} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum_i \xi_i \\
\text{s.t.} \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \\
& \xi_i \geq 0 \\
& y_i \in \{-1, 1\} \\
& \sum_{i \in I} \frac{1+y_i}{2|I|} \geq \theta, \forall I \text{ s.t. } Y_I = 1 \\
& \sum_{i \in I} \frac{1+y_i}{2|I|} < \theta, \forall I \text{ s.t. } Y_I = -1
\end{aligned}
\tag{5.6}
$$

We note that the problem remains mixed integer, like the Optimization Problem 5.3. The first to third constraints remain the same. With this new formulation, we will have at least the fraction θ of the instances of each bag labeled positive in the positive halfspace, and at most the fraction θ of the instances of a negative bag in the negative halfspace. At the same time, the margin is maximized with respect to the complete dataset, according to the instance labels that were assigned.

The resulting mixed integer Optimization Problem 5.6 cannot be solved in closed form. So we employ the following heuristic:

Algorithm 1 θ-mi-SVM optimization heuristics

Input: $\mathbf{x}_i$, B_I, $\mathbf{y}_{B_I}$
Initialize $y_i = Y_I$ for $i \in I$
while *labels change from previous iteration* **do**
  Compute SVM solution $\mathbf{w}, b$ for the train instances and labels
  Compute outputs $f_i = \langle \mathbf{w}, \mathbf{x}_i \rangle + b$ for all $\mathbf{x}_i$ in all bags
  Update $y_i = sgn(f_i)$ for ever $i \in I$
  for *every positive bag* **do**
    if $\frac{\sum_{i \in I} 1+y_i}{2|I|} < \theta$ **then**
      compute $\mathbf{i}^* = \arg\max_{i,\theta|I|} f_i$
      set $y_{\mathbf{i}^*} = 1$
    end
  end
  for *every negative bag* **do**
    if $\frac{\sum_{i \in I} 1+y_i}{2|I|} \geq \theta$ **then**
      compute $\mathbf{i}^* = \arg\min_{i,\theta|I|} f_i$ set $y_{\mathbf{i}^*} = -1$
    end
  end
end
Output $\mathbf{w}, b$

In the above algorithm, we use the notation $\arg\max_{i,K} f_i$ to represent the set of indexes of the K highest valued f_i.

The heuristic to solve the Optimization Problem 5.6 involves alternating between two steps. In the first, given the instance labels, we solve the SVM and find the optimal separating hyperplane. In the second, for a given hyperplane, we update the instance labels in order to respect the constraints that at least or at most a fraction θ of the instances of positive or negative bags, respectively, will have the same label as their respective bag. Naturally, if a given bag already has the required fraction of instances with the same label as the bag, there is no need to update the instance labels for that bag, hence why there is no "else" clause in

Algorithm 1.

Secondly, we propose an adaptation of the MI-SVM, to which we will refer to as θ-MI-SVM to solve the θ-MIL problem using a maximum margin formulation again. In this case, the goal is to extend the notion of a margin from instance level to the bag level. As such, we define the functional margin of a bag with respect to only the instances with the same predicted label as the bag. So for the positive margin, the optimization problem uses the "most positive" instances and for the negative margin it uses the "most negative" instances, such that each positive and negative bag have at least or at most a fraction θ, of the instances being selected as key witnesses, respectively. The new optimization problem can be written as follows:

$$
\begin{aligned}
&\min_{s} \min_{\mathbf{w},b,\xi} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{I} \xi_I \\
&\text{s.t.} \\
&\forall I : Y_I = 1 \wedge (\langle \mathbf{w}, \mathbf{x}_{\mathbf{s}(I)} \rangle + b) \geq 1 - \xi_i \wedge |s(I)| \geq \theta |I| \\
&\text{or } Y_I = -1 \wedge (-\langle \mathbf{w}, \mathbf{x}_{\mathbf{s}(I)} \rangle - b) \geq 1 - \xi_i \wedge |s(I)| \leq \theta |I| \\
&\xi_i \geq 0
\end{aligned}
\tag{5.7}
$$

where the optimization variables w, b, ξ, and s are the weights vector, a scalar, a scalar, and the instance selector respectively. Note that $s(I)$ selects a fraction of the instances of the bag, an not just one, unlike in the Optimization Problem 5.4. Furthermore, since the margins are defined by the instances which have a label that matches their respective bag, this approach ignores the remaining instances within each bag. They are not contemplated in the optimization problem, contrary to the case of θ-mi-SVM.

There can be many initializations of the labels, however [2] recommends initializing the instance labels with the corresponding bag label.

This new problem, as in the MI-SVM formulation, is a mixed integer problem, without an easy solution. Therefore, we use the heuristic shown in Algorithm 2 to solve it iteratively.

Algorithm 2 θ-MI-SVM optimization heuristics

Input: x_i, B_I, y_{B_I}

Initialize $\mathrm{x}_I = \sum_{i \in I} \frac{\mathbf{x}_i}{|I|}$ for every bag

Initialize selector variables $s(I)$, where $\sum_{i \in I} s(i) \geq \frac{|I|}{2}$

while *s(I) changes from previous iteration* **do**

  Compute QP solution $\mathbf{w}, b$ for the train instances and labels

  Compute outputs $f_i = \langle \mathbf{w}, \mathbf{x}_i \rangle + b$ for all $\mathbf{x}_i$ in all bags

  for *every positive Bag* **do**

    set $\mathbf{x} = \mathbf{x}_{s(I)}$, where $s(I) = \arg\max_{i,K} f_i$

  end

  for *every negative Bag* **do**

    set $\mathbf{x} = \mathbf{x}_{s(I)}$, where $s(I) = \arg\min_{i,K} f_i$

  end

end

Output: $\mathbf{w}, b$

Similarly to the heuristic of Algorithm 1 for θ-mi-SVM, Algorithm 2 also alternates between two steps: the first step is, for the given selected instances of every bag, compute the quadratic problem solution and find the optimal separating hyperplane; the second step is, given a separating hyperplane, update the selected instances according to the problem constraints. Once the selected variables do not change from the previous iteration, the algorithm stops. We note that, unlike in the MI-SVM algorithm, in this case the instance selector, s, will select one or more instances of a bag, such that at least, or at most a fraction θ of the instances in the bag are positive or negative, for a positive or negative bag, respectively.

The initialization of the instances can be the bags centroids, as suggested in [2].

As mentioned before, the proposed solutions θ-mi-SVM and θ-MI-SVM have been tested in the context of detecting the polarity of movie reviews, a problem that we were able to formulate using the θ-MIL framework. As these experiments fall out of scope of this thesis, they are not presented in the body of this thesis. However, they can be found in Appendix B.

5.4 Deep Generalized Multiple Instance Learning

In Section 5.3 we have introduced the MIL framework, as well as its generalization θ-MIL. We have explored how it relates to the problem of automating the annotation of the WSM Corpus, without requiring any manual annotation of the training data. In this Chapter we will describe in detail the NN-based formulation of the solution for MIL and θ-MIL.

The motivation to use NN to solve the θ-MIL problem, is a consequence of the success that deep learning approaches have had over the course of the last decade in most problems solved with machine learning. It is only natural that we adopt them over the traditional SVM based approaches to solve the θ-MIL problem. They have the additional advantage of allowing flexible training strategies, since they can be trained end-to-end, via backpropagation, so long as all the transformations in the network can be computed with equations that are differentiable. In the θ-MIL framework, however, there is usually one step that does not respect this restriction: the pooling stage that converts instance labels into bag labels. In this Section, we will propose a novel strategy to approximate the instance label pooling step with a differential transformation, thus formulating a new solution for the MIL framework that is end-to-end differential. At the same time, our proposed formulation is also capable of solving the MIL problem that was introduced before, simply by setting the parameter θ to zero.

5.4.1 Proposed differentiable approximation

Let us assume for the sake of simplicity that instances are represented by features, that are obtained via arbitrary transformations, parametrized by neural networks, such that $h = f_\psi(x)$, where h is the hidden representation of the instance x. h is obtained after performing the transformation $f_\psi(.)$, where ψ are the transformation parameters.

Given a bag of hidden representations of instances, and its respective label $B = \{\{h_1, h_2, ..., h_k\}, Y\}$, the proposed approach is to define two more fully differentiable transformations that convert the hidden representation of the instances h_i, to instance labels y_i: $y = g_{1\phi}(h_i)$, where $y \in -1, 1$, and ϕ are the transformation parameters; and another transformation that pools the instance labels $\{y_1, y_2, ..., y_k\}$ into a bag label Y, following the constrains of the generalized MIL framework, as stated in Eq. 5.5: $Y = g_{2\rho}(y_1, y_2, ..., y_k)$, where ρ are the transformation parameters.

The first of the two transformations, $g_1(.)$, is trivial and any multi-layer perceptron (MLP) can be trained to parameterize it, with the constraint that the output layer contains only one unit and a sigmoid activation function.

For the second transformation, we propose a pooling scheme that guarantees the constraints of not only the MIL framework, but also the θ-MIL framework, as follows:

$$Y = \frac{e^{\sum_{i=0}^{k} \frac{(y_i+1)}{2} - k\theta} + 1}{2(e^{\sum_{i=0}^{k} \frac{(y_i+1)}{2} - k\theta} + 1)}, \tag{5.8}$$

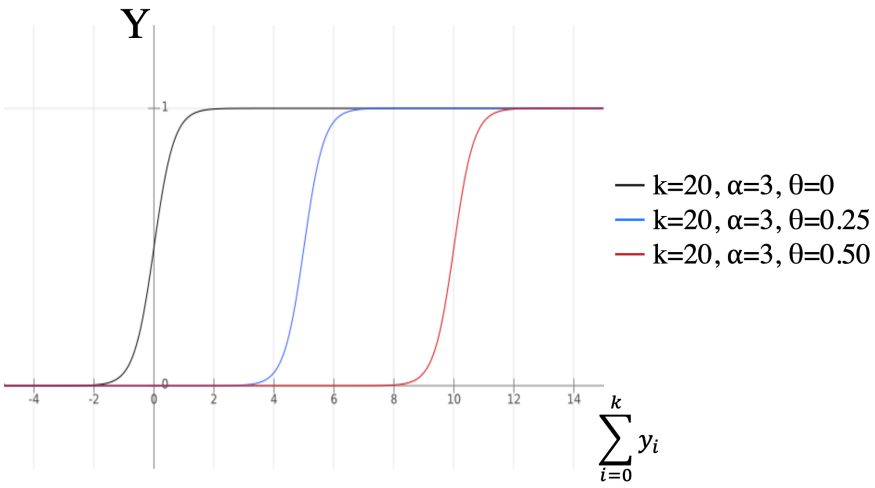


Figure 5.6: Illustration of the different smooth differentiable maximum approximation, with different sets of hyperparameters.

where Y is the bag label, y_i are each of the k instances in the bag. This transformation is essentially the sigmoid function for the sum of the labels of the bag instance, with a horizontal skew of θ. The parameter $\theta \in [0, 1]$ determines the fraction of the instances in the bag that have to be positive to assign the bag a positive label. When $\theta = \frac{1}{k}$, the horizontal skew disappears, and the sigmoid of the sum of the instance labels can be interpreted as a differentiable approximation of the maximum operator. For other feasible values of θ, the horizontal skew on the sigmoid of the sum of the instance labels, will ensure that the function output, *i.e.*, the bag label, is zero when the sum of the instance labels is smaller than $k\theta$, i.e., the number of positive instances in the bag is smaller than the minimum number of instances necessary to assign a positive label to the bag. Figure 5.6 gives a few examples of the differentiable approximation of the maximum operator with different sets of parameters.

In order to train a neural network with parameters ρ that performs the transformation $g_{2\rho}(.)$, the loss to be minimized during training must reflect the θ-MIL constraints. A possible loss based on the binary cross-entropy of the bag labels (as computed in Eq.5.8) is:

$$\begin{cases} L = -\frac{(t+1)}{2}\log(\frac{(s+1)}{2}) - (1 - \frac{(t+1)}{2})\log(1 - \frac{(s+1)}{2}), \\ s = \frac{e^{\sum_{i=0}^{k} p_{yi} - k\theta}}{e^{\sum_{j=0}^{k} p_{yi} - k\theta} + 1} \end{cases} \tag{5.9}$$

where t is the ground truth bag label, and s is the predicted score for a bag label. s is computed by applying Eq. 5.8 to the individual scores of the predicted instance labels p_{yj}.

5.5 Application of Deep θ-MIL for the automatic annotation of the WSM Corpus

In this Section, we test the proposed deep θ-MIL solution introduced in Section 5.4 and published in [129] to automatically label the videos of the WSM Corpus v.2, which has been described in detail in Section 4.3, featuring subjects possibly affected by depression and PD, and using only semantic and metadata information.

As a reminder, we take advantage of the existing structure on the WSM Corpus, where the videos are associated to the search term that was used to retrieve them, as well as a time window for the upload date. Therefore, in this context, a bag is the set of search results obtained for a given search term and time window, and the instances are the retrieved videos. The bag labels are assigned according to whether the search term used to retrieve the set of videos is relevant or irrelevant for the target SA disease.

For example, in the context of the WSM Corpus, if the target disease is depression, and the search term is "depression vlog", then we can assume that among the retrieved search results there will be some that will be target videos, although it is impossible to know which ones without further analysis. This means that the set of these results form a positive bag.

Conversely, if the search term is not related to depression, such as just "vlog", it is generally safe to assume that among the search results there will be no target ones. Thus, this set of results would form a negative bag.

A more subtle case occurs when using search terms like "Parkinson lecture". The results are still related to PD, but from both common sense and empirical evidence, we know that the retrieved videos are very unlikely to contain target videos, *i.e.*, videos of subjects that claim to be currently affected by PD, and are much more likely to contain healthy subjects, such as medical doctors, therapists, researchers or journalists, describing some aspect of PD, mostly from a third party perspective.

We argue that generating such negative bags creates a more interesting and nuanced problem than simply generating negative bags of videos with content that is less related. Models that are successfully trained with such data should be more useful in real life situations to detect subjects affected by a target disease than if the control examples were completely unrelated. From a class separation perspective, our argument is that these negative examples are much closer to the true decision boundary between the two classes than unrelated ones, which allows for a better estimation when training a model to learn it.

It is trivial to formulate this problem as either MIL and θ-MIL, by simply allowing for some positive examples to be in negative bags, which is a reasonable assumption in the context of this problem. In fact, there is no guarantee that irrelevant queries, *i.e.* negative bags, will not produce any positive instance. For generalization purposes, we will describe the following experimental setup and results assuming a θ-MIL formulation, which we can reduce to MIL by setting the threshold θ to zero.

Before presenting the experiments and results of the previously proposed deep θ-MIL solution, we establish an upper bound for the performance that can be achieved on the task of automatically labeling the WSM corpus using the chosen features, by training a fully supervised model on the dataset with the manually obtained labels. After this, we compare the performance of the fully supervised model to the performance obtained by a similar model but in a MIL scenario, where the instance labels are unavailable during training. We also study the contribution of different types features to the performance of the model, extracted from the transcription, title, description and comments from the video, which will be described in detail below.

Finally we study the influence of the bag size in the performance of the models, and as a sanity check, show that when bags are smaller, the MIL problem is easier to solve, and that in the extreme case where the bag size is one, the problem is reduced to a fully supervised learning problem, since all the instance labels correspond to their respective bag label.

5.5.1 Dataset

This work used the WSM Corpus v.2 as described in Section 4.3. Each set of search results was retrieved by a pair of search term and time window, corresponds to one bag, and its size is 50 videos. In the following experiments the data are organized as follows: both the depression and the PD model use all the 550 videos available in the dataset, organized into bags of size 50. In the case of depression, all bags were assigned a negative label except the two that were retrieved with the search term "depression vlog", which were assigned a positive label. For PD all the bags except the three retrieved with the search term "Parkinson's disease vlog" are assigned an negative label, and *vice-versa*.

5.5.2 Feature extraction

In regards to the information derived from the WSM Corpus v.2, we used exclusively information derived from the text-based components of the video. In particular, we used the transcription of the video, the title, the description, and the top 5 comments.

Relatively to the feature extraction process used in this work, we adopted a process derived from Bidirectional Encoder Representations from Transformers (BERT) [130], the Sentence-BERT (SBERT) [131], which we used to encode natural language cues.

For context, BERT was considered the state-of-the-art in encoding language representations at the time of this work, and is based on bi-directional transformers. It is designed to generate representations from unlabeled text by jointly conditioning on both its left and right context. However, it is not optimized for long sentences or even full text documents. SBERT, is a modification of the BERT network, using siamese and triplet networks, in order to derive meaningful sentence embedding of fixed sized, for arbitrarily sized sentences, converting them into feature vectors of 768 dimensions. With SBERT, the similarity between sentences can be computed by any similarity measure such as cosine similarity.

As such we adopt a pre-trained version of SBERT that was first trained on Natural Language Inference (NLI) data, then fine-tuned on AllNLI, and on the semantic text similarity (STS) benchmark training set, obtaining an STS score of 85.29, out of 100 possible points, as reported by the authors.

We use this pre-trained model to embed the three documents associated with a search result: 1) the transcription of the video; 2) the title and description of the video provided by the user; and 3) the top 5 comments on the video, sorted by popularity. Thus each search result is characterized by three 768-dimensional vectors. We repeat this for the complete dataset of 550 search results, which amounts to a total of 1650 embeddings.

Some of the videos did not have any comments. In such cases, a random 768-dimensional vector was generated, which in practice represents a random sentence/document.

5.5.3 Fully supervised upper bound

As mentioned before, the WSM Corpus contains manual annotations. While it is not realistic to assume that these will be available in a real-life application, they are useful in this context to perform a number of tasks, such as for evaluation purposes, for semi supervised learning tasks, *etc.*. In this case, we use the manual annotations of the WSM corpus to perform the fully supervised learning task of predicting the presence of subjects who claim to be currently affected by two diseases, depression and PD, from SBERT embeddings of the transcription, title and description, and top 5 comments of YouTube videos. This experiment is useful to determine what is the upper bound of the performance that could be obtained with a model trained on the WSM corpus, if all the instance labels were available. In other words, by comparing the following results with the performance of a similar model in a (θ)-MIL

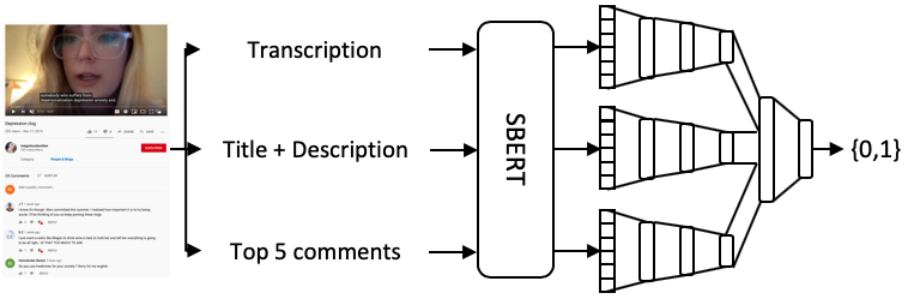


Figure 5.7: Architecture of the fully supervised model that estimates the upper bound of the performance that can be obtained in labeling the WSM Corpus, given the feature choice and model architecture.

learning scenario, it is possible to quantify the loss of knowledge when the instance labels are not available and only the bag structure and bag labels of the dataset are.

The architecture of the fully supervised model is shown in Figure 5.7. It comprises three streams of MLPs, one for each type of embedding of the three documents available, with 768 dimensions each. Each stream contains two fully connected layers, with 256 and 64 units, respectively, and both with ReLU activations, and a dropout rate of 0.2. The streams are fused by concatenating the three hidden representations. The network has two additional fully connected layers, the first of them with 64 units and a ReLU activation and a dropout rate of 0.5, and finally, a 1-unit output layer with sigmoid activation.

The network was trained with binary cross-entropy loss, and RMSProp optimizer algorithm, over 60 epochs, with a learning rate of 0.001, and early stopping conditions based on the development loss.

The model was trained with 400 examples and tested against 150 examples for both depression and PD.

The performance of this model was measured in F1 score, since there is a significant class imbalance. The fully supervised model obtained an F1 score of 0.69 and 0.67 on the test set for depression and PD, respectively. These results are also shown along with others in the bar charts on Figures 5.8 and 5.9 for depression and PD, respectively, as the leftmost columns in the color red.

These results by themselves are not very meaningful, since the features or the model architecture, among others, could be changed in an attempt to improve the performance of the model, however this is the most fair upper bound to performance of the experiments in the

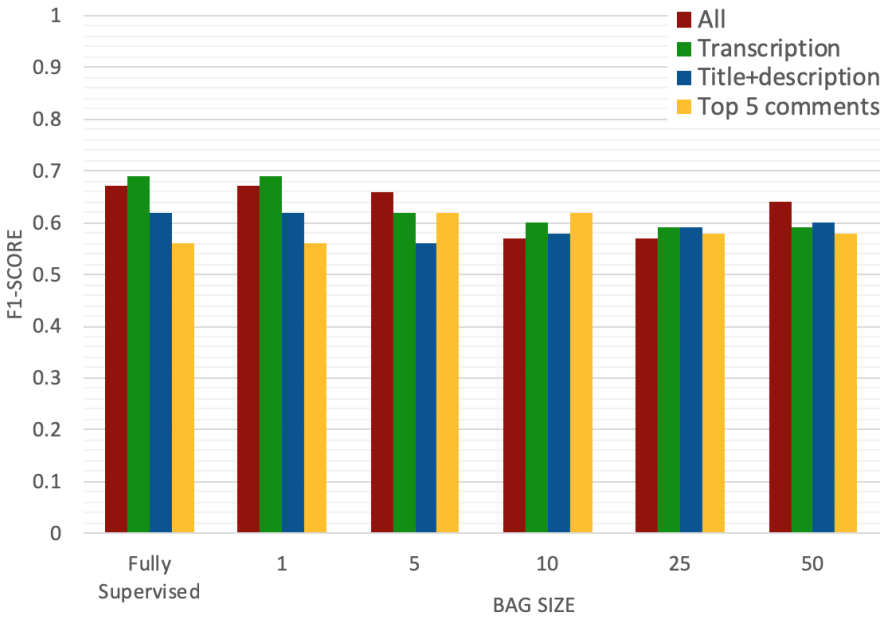


Figure 5.8: Summary of the performance in F1 score of all the models trained to estimate the depression labels of the WSM Corpus, for different bag sizes and sources of textual cues.

following Sections.

We also want to highlight a limitation in all of the experiments presented for the remainder of the Section. In order maximize the number of bags available at train time for both depression and PD, we used the examples retrieved for the query "Parkinson's disease vlog" as negative examples for the self-reported health status of depression. While this is technically accurate, since no individual in the PD videos makes the claim that they are currently affected by depression, it is an error in practice. We use the self-reported health status as a proxy for the true health status, however, since the co-occurrence of depression in PD patients is known to be frequent, this proxy may not be verified for the presence of depression in PD patients.

5.5.4 Deep θ-MIL performance

We proceeded to experimentally verify the main contribution made in this Section, where we tested the deep θ-MIL solution, in labeling the WSM Corpus, without access to any of the manual labels, and having only access to the bag structure and to the bag labels.

As mentioned in Section 4.3, the WSM Corpus v.2 has 11 bags of 50 examples each. We used 8 of these bags, which total 400 instances for training and, the remaining 3 for testing purposes, which contained 150 examples. The distribution of the train and test examples is the same as in the experiments reported in the upper bound experiment, for comparison

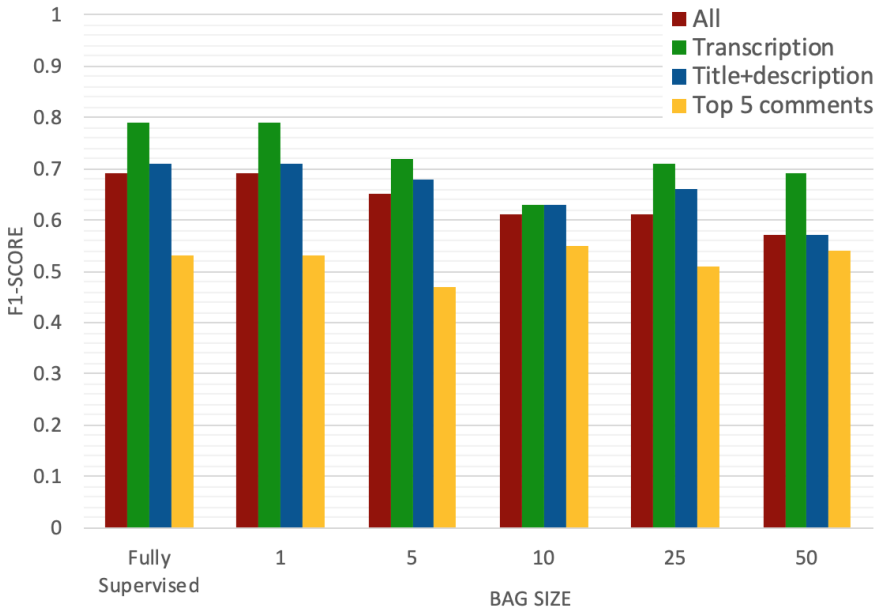


Figure 5.9: Summary of the performance in F1 score of all the models trained to estimate the PD labels of the WSM Corpus, for different bag sizes and sources of textual cues.

purposes.

For each of the instances, we computed the same SBERT embedding for the three available documents: the transcription, the title and description and the top five comments, and use these three 768-dimensional vectors as the input to the network.

The architecture of this network was similar to the one described for the upper bound experiment, where each instance was processed by a three stream network. These streams were then fused and used to generate an instance prediction. The key difference in this case was that this process was repeated for all the instances in the bag, and the prediction of the bag label was made according to Eq. 5.8. Only then the loss was computed according to Eq. 5.9, and the weights are updated through backpropagation. A good strategy to implement this network was to set the batch size to be the bag size and process the instances if the same bag in sequence, so that they were all processed in the same batch. By doing so, the loss was accumulated over the whole bag and the predictions for all the instances in the bag were computed with the same network weights.

We summarize the architecture of the proposed deep-θ-MIL network in Figure 5.10.

The network was trained with similar parameters as the one presented in Section 5.5.3, except θ was set to $2/k$, where k is the size of the bag.

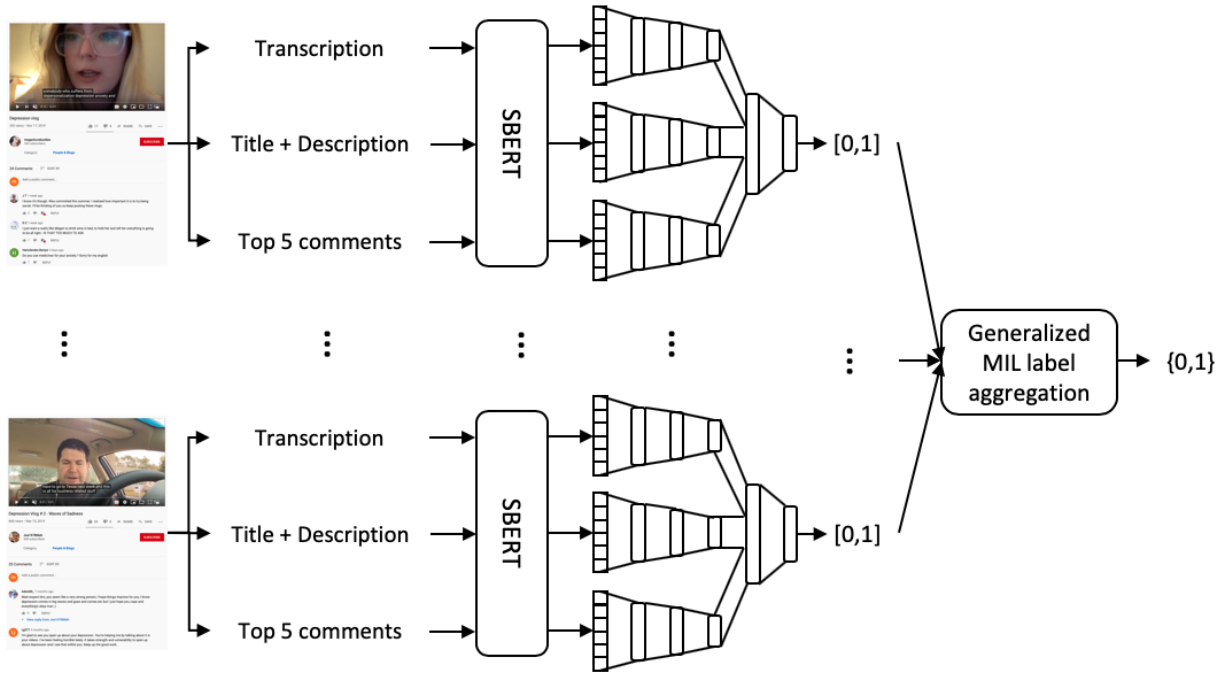


Figure 5.10: Architecture of the proposed deep-θ-MIL solution to automatically annotate the WSM Corpus. This architecture is based on a 3-stream network, where each stream processed one document.

The performance of this network on the test set at instance label level, was an F1-score of 0.64 and 0.57, for depression and PD, respectively. These results are also shown on Figures 5.8 and 5.9, on the set of columns above the label 50, with the color red. We also note that the network was able to achieve an error of zero at the bag level labels, however we consider that this result is not relevant in the scope of this work, which is instance level prediction, so we will not mention it in further experiments.

We note that for the case of PD, there was a significant drop in performance, while for depression the drop in performance was smaller. Nevertheless, in both situations we could experimentally confirm the hypothesis that learning in a (generalized) MIL scenario is a harder problem than a fully supervised one.

5.5.5 Contribution of each type of document

So far, we have only shown experiments where we take advantage of the three types of documents available for each video, the transcription, the title and description, and the top five comments, at the same time. However, it a reasonable assumption that the contribution of each document could vary widely for the final instance label prediction. In this Section, we study the contribution of each document for the performance of the proposed networks.

Table 5.5: Performance in F1 score of the proposed deep MIL network for one type of textual cue at a time, for depression and PD for the original bags of size 50.

Document\Disease	**Depression**	**Parkinson's disease**
Transcription	0.59	0.69
Title + description	0.60	0.57
Top 5 comments	0.58	0.54

We achieve this by making a minor modification to the network described in Section 5.5.4: removing two of the three documents, and their respective streams in the network. Following this change, the only other necessary change was to remove the concatenation layer that merged the three streams into one vector.

We performed the same experiments with the same data partitions as described above, for each one of the three types of documents. The performance in F1-score of the instance level label prediction is summarized in Table 5.5 for depression and PD. For an easier visualization of the results they are also presented in Figures 5.8 and 5.9 on top of the label "50", in the colors green, blue and yellow, for depression and PD, respectively.

As can be seen for PD, the source of data that contains the most useful information for this problem is the transcription, which on its own outperforms the networks trained on the three documents. Intuitively this result was not unexpected, since it was reasonable to assume that the content of the conversation by the subject of the video would be far more important to determine the health status than the title and description of the video, or the comments to the video. However, this was not the case for depression, where the best model was obtained by combining the information from the three documents. The difference in performance depending on the documents used as input to the network was also smaller than with PD.

We note however, that across the two SA diseases, the models trained only with the top five comments were the ones with the poorest average performance. One of the reasons behind this may be that not all the videos have comments, and for these cases random embedding vectors were generated to replace them. At the same time, we observed empirically, through reading some of the comments, that there was little consistency in their content and sentiment across videos with the same label.

5.5.6 Influence of bag size

Another variable in this problem that is interesting to study is the influence of the bag size. Intuitively, one would expect as the size of the bags got bigger, the learning problem would

become harder, since the restrictions on the instance labels are smaller. In this Section we tested this hypothesis for bag sizes of 5, 10, and 25. Additionally we also performed the same experiments by setting the bag size to 1, where we expected to see the same performance as the one obtained by the fully supervised model, since the two problems become equivalent: the instance labels are completely determined by the bag labels.

We note that we also repeated the experiments where we analysed each type of document separately for different bag sizes: 1, 5, 10, and 25, so that we could make a complete report on the influence of these two variables: bag size, and input documents.

The networks were trained with the same parameters as before, as well as trained and tested in the same partitions of the data. However, depending on the bag size, each bag was randomly divided into smaller bags. The instances from the original bags were not mixed. The new bag labels were assigned based on the aggregation of the manually obtained instance level labels for that bag, following Eq. 5.5.

The results for this experiment, for both depression and PD are summarized in Table 5.6. Furthermore, the results are also shown in Figures 5.8, and 5.9, for depression and PD, respectively, above the numbers 1, 5, 10 and 25. Between these two figures it is possible to compare the performance of all the models with different bag sizes and input documents, and the fully supervised upper bound.

Again, in this set of experiments, we are able to confirm the hypothesis that, as a rule of thumb, larger bags are associated to a harder learning problem: we can observe a dropping trend in the performance of all the models, as the bag size increases. This occurs because for bigger bags, the restrictions imposed on the instance labels are smaller. In fact, in the extreme case of a bag with infinite samples, we would be almost in a completely unsupervised learning scenario, other that having the prior of knowing that a fraction θ of the instance belonged to the positive class.

Overall, in the best model for depression, which was obtained using the three documents as the streams inputs, we observed a drop in performance from 0.69 to 0.64, from the fully supervised case to the MIL with bag size of 50. This is an absolute decrease of 5% in performance, that is sacrificed in order to waive the requirement to annotate *any single* sample in the depression subset of the WSM Corpus, going from a fully supervised to a weakly supervised problem.

In the case of PD, the best overall model was the one trained only on the transcription. For this one, the performance obtained in the fully supervised scenario was 0.79, which then dropped to 0.69 when the instance labels were hidden, and the bag size was set to 50.

Table 5.6: Performance in F1 score for the proposed deep MIL network for different sizes of bags, and different types of textual cues, for depression and PD.

Bag size	Document	Depression	Parkinson's disease
1	All	0.67	0.69
	Transcription	0.69	0.79
	Title + description	0.62	0.71
	Top 5 comments	0.56	0.53
5	All	0.66	0.65
	Transcription	0.62	0.72
	Title + description	0.56	0.68
	Top 5 comments	0.62	0.47
10	All	0.57	0.61
	Transcription	0.60	0.63
	Title + description	0.58	0.63
	Top 5 comments	0.62	0.55
25	All	0.57	0.61
	Transcription	0.59	0.71
	Title + description	0.60	0.66
	Top 5 comments	0.58	0.51

The cost of going from a situation where the full dataset was annotated, and the learning scenario was fully supervised, to the generalized MIL framework where only the bag labels are available and no sample of the corpus was manually annotated, was 10% of the absolute performance. Even though the drop in performance was smaller for depression, the absolute performance was better in PD.

Overall this technique shows promising results regarding the automatic annotation of the WSM Corpus without any annotated data available. At the same time, the proposed technique can be improved by finding other more suitable representations for the documents or better network architectures, so that the absolute performance of all the models is improved.

Looking back to the works described in Sections 5.1, 5.3, and 5.4, there would be much room for improvement of all the presented techniques by, for example, leveraging from each other, or by combining them. However, due to time limitations that possibility was not addressed.

At the same time, we can reflect on the work presented in this Chapter and propose other strategies, for future work, that could also be relevant to solve the problem of automating the annotation of the WSM Corpus. Considering the very specific definition of positive self-reported diagnosis, *i.e.* the moment in which the subject of the video claims to be

currently suffering from the target SA disease, it could be argued that a simple rule-based solution could provide a relevant alternative to solve this problem. For example, by searching for and counting sets of target words, to searching for negations, to identifying verb tenses, *etc.* This, and other avenues remain to be explored.

Part III

Detecting speech affecting diseases in-the-wild

Chapter 6

Detecting speech affecting diseases in-the-wild

After describing the collection and manual annotation process of WSM Corpus on Chapter 4, and proposing several automatic annotation strategies to replace its manual annotation process on Chapter 5, we move on to explore the topic of detecting SA diseases in an in-the-wild context and/or with in-the-wild data.

Before moving on to describing in detail the experiments conducted, we want to note that, in this Chapter, we will be focusing on the speech modality alone, to detect SA diseases. In previous Chapters, we have dedicated our attention to the natural language modality (and metadata) to address the automation of the collection and annotation of speech medical corpora. This meant that our focus was on *explicit* cues, collected from *what* was said by the subjects in the videos.

However, we now shift our focus to speech, since we want to be operating in such conditions that detecting SA diseases does not depend on *what* has been said, but only on *how* it was said. At this stage, we will be looking for *implicit* cues that convey a subject's health status, present in the speech signal, rather than the *explicit* ones, present in the content, that were studied in Chapter 5. This decision is especially important in the context of this work, given that the majority of the speech medical corpora, and particularly, the ones used in the scope of this work (DAIC, PC-GITA, and the WSM Corpus), contain subjects who are not only aware of their health status, but where a fraction of the target speakers are also explicitly discussing aspects of their SA disease (including symptoms, medication, treatments, *etc.*). We consider that, in such conditions, it would be redundant to develop a system that would rely on the explicit, natural language based cues, related to the subject's knowledge of their

health status to determine the subject's health status.

In summary, we aim to study speech-based only SA disease diagnosis tools, which have the following advantages:

- Not depending on the individual's knowledge about their health status (*i.e.* the individual does not need to be aware that they suffer from a SA disease, acknowledge it, or even to be talking about it, which was the starting point for the distinction between healthy and non-healthy individuals on Chapter 5
- Not depending on the subject's topic of conversation, which makes the system more flexible to be used in scenarios where the speaker is not discussing their health status or disease
- Providing a layer of privacy to the user, in the sense that the contents of their conversation are never analysed in order to make a prediction about their health status

This Chapter is divided into two Sections. The first one, Section 6.1, is dedicated to describing in detail the solutions we propose to detecting SA diseases. We explore a total of four strategies, of which, one is knowledge based, two are speaker modeling based, and another one is an end-to-end DL based strategy.

The second Section in the Chapter, Section 6.2, details the experiments that were performed to address the problem of detecting SA diseases both in CC and with in-the-wild conditions. For this, we apply the solutions proposed previously, in Section 6.1. The sets of experiments we performed can be broadly classified into one of three categories, regarding the domain of the data used at training and evaluation stages: same domain, cross domain, and multi-domain experiments.

We define as same domain the experiments where the data used for training and evaluation come from the same domain, *i.e.* training with CC data and evaluating with CC data, or training and evaluating with in-the-wild data. Cross domain refers to experiments where the training and evaluation data belong to different domains, *i.e.* training with CC data and evaluating with in-the-wild data and *vice-versa*. Finally, mixed domain experiments are those where data from all available domains is present during training, *i.e.* training is done with both CC and in-the-wild data, and evaluation is done separately for each domain.

More specifically, regarding the same domain experiments, we begin by establishing a baseline for the task of detecting SA diseases in CC, that we will use as a starting point to compare to the performance of several state of the art methods to the detection of SA diseases in in-the-wild conditions, in later Sections of this Chapter. For this baseline, we adopted several

different strategies to model cues that carry health related information. We explore different knowledge based, speaker modeling based, and data driven feature extraction methods. After establishing baseline performances for the task of detecting SA diseases in CC, we move on to studying the same task under in-the-wild conditions, and establish a comparison between the two. When dealing with in-the-wild data, one should consider that it has different characteristics from CC data. One of these is the high heterogeneity of the channel and noise conditions in the recordings. Conversely, this is an aspect that tends to be very well controlled for in CC data, since the recordings are all obtained in similar conditions. Another aspect is the content of the recordings, which, in the case of the data from in-the-wild sources, is completely up to the subject, whilst in CC data it tends to follow a predefined protocol, or be guided in the setting of a clinical interview.

The next experiments that we tackle regard measuring the generalization power of speech medical data from CC when faced with data from in-the-wild. Our claim is that the restricted conditions in which the CC data are collected, particularly in terms of guiding the patients in their spoken interactions with speaking exercises and via clinical interviews, will contribute to reduce the acoustic variability of the examples present in the corpus. Therefore, CC data would provide a less complete representation of the acoustic variability of the cues that characterize speech affected by a target SA disease, than in-the-wild data. We test this hypothesis by using CC data as the train material for an arbitrary model, and the in-the-wild data as the test material.

Then, we move on to what are arguably the most important experiments in the context of this Chapter, which consist of verifying the previous assumption that the self-reported health status present as the labels of the WSM Corpus, is in fact a good proxy for the true health status of the speakers in that corpus. With this in mind, we recall that the labels of the CC data correspond to the individual's true health status, as assessed by the appropriate healthcare professional. On the other hand, the labels of the in-the-wild data in the context of this work, correspond to the individual's self-reported health status, which is not necessarily the same as their true health status. While these two types of labels are different, intuitively we expect the self-reported health status to be mostly the same as the true health status, since there is no incentive for the individual's of the in-the-wild dataset to be deceiving about their health status. However, the experimental verification of this hypothesis has not been done yet. We verify this hypothesis by using in-the-wild data with labels corresponding to self-assessed health status as the training material, and CC data with true health status labels as the test material. A successful model working under these conditions, shows that the examples of self-reported targets (subjects claiming to suffer from either depression or

PD) in the WSM Corpus, display in fact the same acoustic cues that are present in speakers actually diagnosed with those conditions, and therefore, showing that the self-reported health status corresponds the true health status, in the context of the WSM Corpus.

Finally, an additional aspect we want to address in this Chapter, described in more detail in Section 6.2.3, regards combining two different sources of speech medical data - the CC data, and the in-the-wild data - to improve the performance of the detection of SA diseases, leveraging from data from both domains.

6.1 Modeling Strategies for detecting SA diseases

Previous works on the detection and monitoring of SA diseases, particularly depression and PD, as discussed in Section 3.1, tend to describe the acoustic signal using hand-crafted features that are known to capture, to some extent, meaningful information about the health status of the speaker. The advantage of using such features is that, since their extraction follows pre-determined and deterministic algorithms, they do not require data to be trained. Hence, they become useful in situations where the training data are scarce, or datasets are small. This is typically the scenario in speech medical tasks. After extracting the hand-crafted features, they can be passed to an arbitrary machine learning algorithm that will aim to learn relationships between them and some health attribute of the speaker, in our case the presence of a target SA disease. However, one disadvantage of hand-crafted features is that they depend on the existence of domain specific knowledge, necessary to determine the feature extraction algorithms.

Another possible avenue to detect SA diseases is through speaker representation based techniques, including i-vectors and x-vectors, which have been introduced in Section 3.1. Although these techniques were initially developed to tackle tasks related to speaker identification and verification, there has been some evidence that such speaker vectors carry not only speaker related information, but they also model other aspects related to speaker variability, which includes the health status [98] [132] [133].

More recently, there have been attempts at incorporating end-to-end DL based strategies to detect SA diseases, including by extracting information directly from the raw acoustic signal, as described in Section 3.1, also reported in [134], and subsequently others [135]. Such scenarios have several advantages, including that, given their data driven nature, they do not require domain knowledge to be used, and do not make assumptions about the data. However, they typically require large amounts of data to be trained.

In this Section, we will describe four strategies to detect SA diseases, that fall in one of three broad categories mentioned above: generic knowledge based, speaker modeling based, and end-to-end DL based. Along with the strategies, we will describe how they can be included in a complete pipeline to detect SA diseases. The systems presented in this Chapter have been included in parts of prior publications [134] [133].

We note that the strategies that will be discussed in this Section are all disease agnostic. This means that they have not been developed to address any specific SA disease, and that, while our experiments focus on depression and PD, there is nothing in the presented systems and pipelines that is specific to a target SA disease. Therefore, our solutions can be directly used to detect other SA diseases. In the case of the knowledge based features, given their very nature, it can be argued that depending on which ones are used, they can be more suitable for a specific set of SA diseases, but we chose to adopt relatively generic knowledge-based features, as described below.

6.1.1 Generic knowledge based approaches

eGeMAPS with SVMs

The eGeMAPS [77] are a feature set of 88 low-level descriptors (LLDs) and functionals that represent the speech signal in terms of energy, spectral and cepstral features, pitch, voice quality, and micro-prosodic parameters. They were initially proposed in the context of affecting computing, and have showed promising results in tasks such as emotion recognition [136]. More importantly, in the context of this work, they have been used to detect some SA diseases in CC, including depression in recordings of clinical interviews [68], PD [137], autism spectrum disorder [138], among other. In the case of PD, there exist other works that address similar tasks, namely [139] which focuses on detecting medication state in PD patients using, among different features, including eGeMAPS.

The eGeMAPS can then used as input to an arbitrary classifier. In our case, we chose SVM, which will then output the prediction for the health status. The system is the same for both depression and PD.

The described system corresponds to one of the baselines provided in previous INTERSPEECH ComParE Challenges [140], therefore we use it as a starting point to the detection of SA diseases. This baseline is summarized in Fig 6.1.

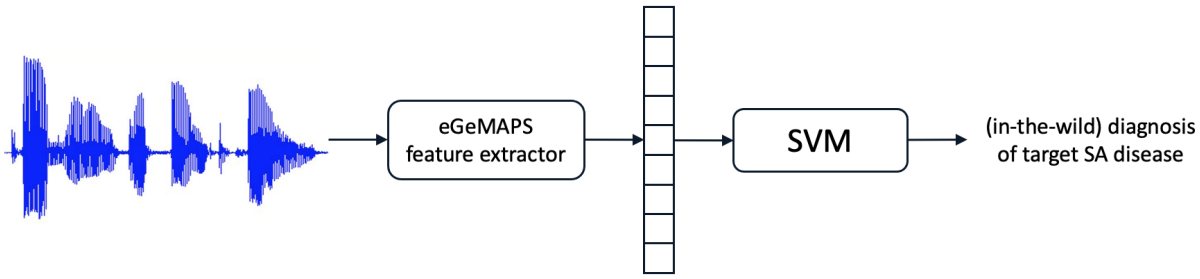


Figure 6.1: Baseline system, using eGeMAPS and SVMs, as proposed in previous INTERSPEECH ComParE Challenges, to detect SA diseases.

6.1.2 Speaker modeling based approaches

i-Vectors with PLDA

The i-vector approach [141] was first introduced in the context of speaker verification and identification. However, it has later been successfully used to solve other tasks related to speech, including emotion recognition [142], language recognition [143], age estimation [144], and acoustic scene classification [145] [146]. More recently, it has also been shown that i-vectors also model some information related to the health status of the speaker [147].

The i-vector approach was motivated by the success of the joint factor analysis (JFA) [148], which models speaker and channel spaces separately. However, unlike JFA, the i-vector approach aims at modeling the total variability, composed of both speaker and channel variability, together in a single low-rank sub-space, called the total variability space. This has been shown to lead to an increase in robustness in channel variations and other sources of signal distortion.

The idea of modeling the total variability space consists of adapting a universal background model (UBM) to a set of given speech frames, based on the speaker-specific subspace adaptation technique in order to estimate the utterance dependent Gaussian mixture model (GMM). The assumption is that all the pertinent variability can be captured by a low rank rectangular matrix T, named the total variability matrix. The GMM supervector (the vector created by stacking all mean vectors from the GMM), M, for a given utterance can be modeled as:

$$M = m + Tw \tag{6.1}$$

where m is the speaker and channel independent component represented by the UBM supervector, and T is the total variability matrix with low rank that maps the high dimensional

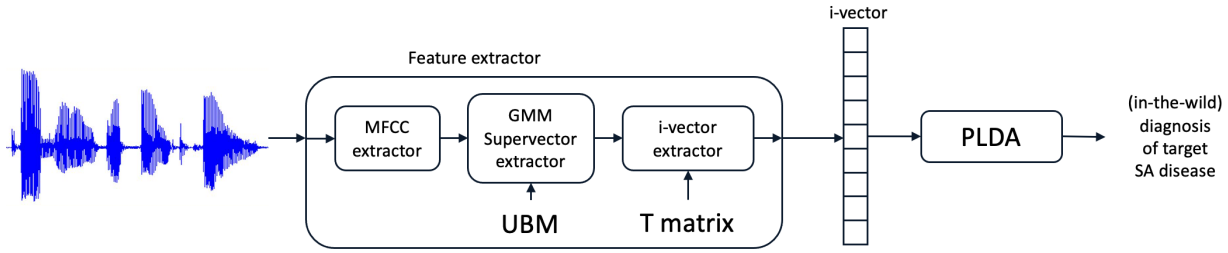


Figure 6.2: Framework, using i-vectors as the front-end and PLDA as the back-end, to detect SA diseases.

GMM supervector into a lower dimensional i-vector, w.

The modeling of the UBM and the total variability matrix can be achieved via expectation maximization (EM), where, in the E-step, w is considered as a latent variable with normal prior distribution $N(0, I)$. The resulting i-vectors are estimated as the mean of the posterior distribution of w as follows:

$$\hat{w}(u) = (I + T^T \Sigma^{-1} N(u) T)^{-1} T^T \Sigma^{-1} S(u), \tag{6.2}$$

where, for an utterance u, the terms $N(u)$ and $S(u)$ represent 0^{th} and centralized 1^{st} order Baum-Welch statistics respectively, and Σ is the covariance matrix of the UBM.

During the evaluation stage, we opted to use the resulting i-vectors as input to a PLDA model [149], which, given two i-vectors $\hat{w}_1$ and $\hat{w}_2$, will compute a score that describes the probability that they belong to the same class. This score is computed as the log-likelihood ratio between the same *versus* different class models.

The described framework, using i-vectors computed from MFCCs as the front-end, and PLDA as the back-end has been a popular choice for several speech based tasks, and it is represented in Figure 6.2.

x-vectors with PLDA

X-vectors [150] are deep neural network based speaker embeddings, that were proposed as an alternative to i-vectors for speaker and language recognition tasks. In contrast with i-vectors, that represent the total speaker and channel variability, x-vectors aim to model characteristics that discriminate between speakers. When compared to i-vectors, x-vectors require shorter temporal segments to achieve good results, and have been shown to be more robust to data variability and domain mismatches [99] [150]. As with i-vectors, x-vectors have also been

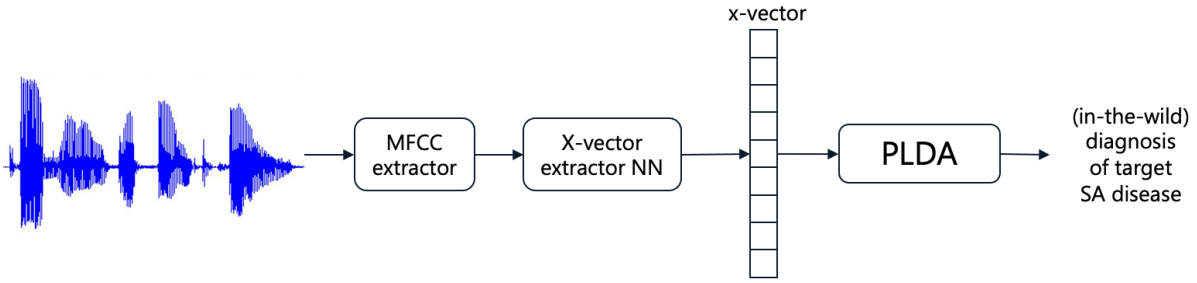


Figure 6.3: Framework, using x-vectors as the front-end and PLDA as the back-end, to detect SA diseases.

used in a variety of tasks related to speech, including language recognition [151], acoustic scene classification [152], speaker diarization [153], emotion recognition [154], among others. Additionally, it had also been shown that x-vectors carry some information about the health status of the speaker, particularly for the detection of PD [98], Alzheimer's disease [155], and OSA [156].

The x-vector extraction process relies on a time-delay NN (TDNN) that computes speaker embeddings from variable length acoustic segments. The typical x-vector network has an architecture as follows: the first few layers have a time-delay architecture [157], that operates on speech frames; then, a statistical pooling layer receives the input of the last TDNN as input, aggregates over the input segment, and computes its mean and standard deviation; following the pooling operation, there are two additional fully connected layers that receive as input the concatenated segment-level statistics obtained as the output of the pooling layer, and compute the speaker embedding; and, finally, a softmax output layer, that is used for training only, and removed when computing new x-vectors. The cost function of this network is based on a multi-class cross entropy. After training the network, the x-vector embeddings are extracted at the level of the first fully connected layer.

For the purpose of detecting SA diseases, we use the x-vectors as inputs to a PLDA model, which will then estimate the health status of the speaker by comparing pairs of x-vectors.

This framework, as the i-vector based one, has also been used before in other speech based tasks, and is summarized in Figure 6.3.

6.1.3 End-to-end DL based approaches

CNN-LSTM with self-attention

The final approach that we have used to detect SA diseases differs from the previous ones in that, in this one, our goal is to use the raw spectrograms as the input to our network, and let the network determine on its own which are the relevant features for this task. This approach blends front- and back-end into a single network that can be trained end-to-end, without inputting into the model any pre-existing domain knowledge. On one hand this strategy relieves the model from the biases that hand-crafted features may impose, on the other hand, such strategies typically required large amounts of data to robustly estimate the relevant features for the given tasks.

The idea of using end-to-end approaches to address problems related to speech has been proposed for tasks such as speaker identification and recognition [158], speech recognition [159] [160] [161], language recognition [162], acoustic localization [163], speech dereverberation [164], emotion recognition [112] [165], among others.

However, at the time of this work, there were, few applications of end-to-end DL based strategies in tasks related to healthcare, particularly to detect depression [73], and PD [166] from speech. In recent years, this type of architectures has received more attention, which resulted in several works that use such strategies to detect depression [167] [168], and PD [101] [169].

The end-to-end architecture chosen for our task of SA diseases diagnosis consists of a combination of CNNs, LSTMs, and a self-attention mechanism to jointly model local spectro-temporal information from the raw spectrograms, producing a high-level representation of the input.

For this, we use fixed length raw mel spectrograms directly as the inputs to the NN. These are fed to several CNN layers which have the purpose of embedding the spectrograms, thus replacing the function of a traditional feature extraction step. These hidden representations are then used as inputs to an LSTM layer, that is used to capture the longer term temporal information present in the input sequence. Finally, we include a self-attention mechanism to capture correlations between the different elements of the input sequence.

Figure 6.4 summarizes the proposed end-to-end system.

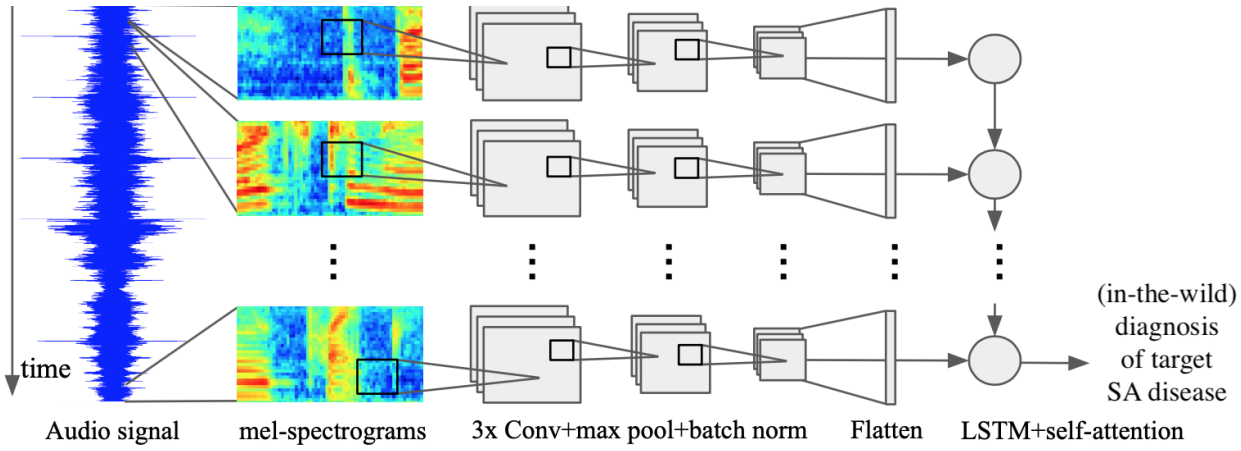


Figure 6.4: Proposed end-to-end model: this model uses mel-spectrograms as inputs to a CNN-LSTM network, where the LSTM layer has a self-attention mechanism.

6.2 Experiments and results

6.2.1 Datasets

To perform the experimental verification of the models described in Section 6.1, which will be presented in Section 6.2.2, we use several datasets from both depression and PD that represent these diseases in both CC and in the-wild domains. Specifically for depression, to portray CC, we use the DAIC-WOZ, a dataset of semi structured clinical interviews, which has been previously described in Section 2.2. Out of the full interviews, we use only the patient side of the conversation, which reduces the dataset to roughly half of its original size. As the in-the-wild dataset for depression, we use two versions of the WSM Corpus: versions 1 and 3, also previously described in Section 4.2 and Section 4.4, respectively. For the experiments performed with the systems based on eGeMAPS, i-vectors and x-vectors, described in Sections 6.1.1, and 6.1.2, we have used version 3 of the corpus. However, for the experiments performed with the end-to-end system described in Section 6.1.3, we have used version 1 of the WSM Corpus, which is a significantly smaller version of the dataset, collected at an earlier stage. This discrepancy between datasets used along different experiments occurred because, at the time when the experiments referring to the system described in Section 6.1.3, based on end-to-end models, were performed, the collection of version 3 of the WSM Corpus had not yet been completed. Subsequently, due to time constraints, the experiments could not be replicated on the larger and more recent version of the WSM Corpus, version 3. We understand that this discrepancy may pose an issue on the direct comparison of the performance of different models, which is something we will keep in mind when discussing the results of Section 6.2.3.

Table 6.1: Summary of the datasets used for the experiments described in Section 6.2, by disease, and recording condition. *All systems except end-to-end, **end-to-end.

Target disease	Condition	Dataset	Size [h]
Depression	CC	DAIC-WOZ (patient side only)	29
	In-the-wild	WSM v.3*	83
		WSM v.1**	28
Parkinson's disease	CC	PC-GITA (monologue, read text, sentences only)	3
	In-the-wild	WSM v.3*	54
		WSM v.1**	13

In the case of PD, to represent CC, we have used the PC-GITA, a small dataset of speaking exercises in Spanish, previously described in Section 2.2. Within the available speaking exercises, we have restricted our experiments to use the subset corresponding to the monologue, read text and sentences tasks, and have excluded the words, vowels and DDK tasks. The motivation for this choice was to guarantee that the tasks present in the CC and in-the-wild data were of similar nature (all running speech). As such, the final subset of the PC-GITA that was used in the experiments, described in Section 2.2.2, included a total of 3 tasks (and 13 sub-tasks) for each of the 100 patients present in the corpus, totaling 1200 recordings. This corresponded to a total of roughly 3 hours of speech.

The data for PD that were used to represent in-the-wild conditions corresponded to the PD dataset of versions 1 and 3 of the WSM Corpus, also previously described in Section 4.2 and Section 4.4, respectively.

Once again, as in the case of depression and for the same reason, all the experiments but the ones referring to the end-to-end system were performed using version 3 of the WSM Corpus, and the remaining experiments used version 1 of the WSM Corpus.

To ensure that the comparison between all the datasets is straightforward, we summarize them in Table 6.1.

6.2.2 Experiments

Intuition

Throughout this document we have hypothesised several times about the differences between the characteristics of CC and in-the-wild speech medical data, and what implications they

may have in the representation of SA diseases. We have argued that in-the-wild speech medical data could resembles real-life scenarios more closely than CC data, and therefore, provide a more robust representation of the acoustic characteristics of a target SA disease. As a consequence, in-the-wild data, could be a valuable source that can enrich our knowledge of SA diseases. However, so far we have not tested our hypothesis, or quantified the differences between speech medical data from the two domains. This Section is dedicated to do so, through a number or experiments based on the systems for detection of SA diseases described in Section 6.1. The experiments that are carried out in this Section were designed to, on one hand, compare the use of hand-crafted features, speaker modeling approaches, and end-to-end DL based approaches to detect SA diseases; and on the other hand, to show the influence of the domain in which the data were collected (CC, or in-the-wild conditions) on the model performance.

As explained in the introduction of this Chapter, the experiments described in this Section, for both depression and PD, can be broadly categorized into one of three categories, depending on the domain of the data used for training and testing the system, and regardless of the modeling strategy used:

- Same domain: Train and test on the same domain (CC vs CC, and in-the-wild vs in-the-wild conditions)
- Crossed domain: Train and test on different domains (CC vs. in-the-wild and *vice-versa*)
- Mixed domain: Train on data from both domains and tested on each domain (CC and in-the-wild data vs CC; and CC and in-the-wild data vs in-the-wild)

Regarding the same domain experiments, both can be interpreted as establishing the baseline for the performance of the detection of SA diseases in either CC or on-the-wild conditions using the four different strategies described in Section 6.1. Additionally, the performance obtained by the same domain experiments using CC data can be compared to other works in the literature, since they use datasets that have been publicly available for several years. The performance obtained in same domain experiments with in-the-wild data can be viewed as a baseline for the remaining in-the-wild experiments. These results, when compared to the results obtained in the cross domain experiments, will allow us to measure the impact of interchanging the domains of either the train or test data.

The results obtained in cross domain experiments have a less straightforward interpretation. In the case of the experiments where the training data come from CC, and the test data come from the in-the-wild conditions, we will be able to observe the limitations that CC data have in detecting SA diseases in conditions similar to those of real life applications. The

Train conditions \ Test conditions	**CC**	**In-the-wild**
CC	Baseline for CC	Limitations of CC data to detect SA diseases in-the-wild
In-the-wild	Validity of proxy between self reported and true health status	Baseline for in-the-wild
CC + In-the-wild	Leveraging from 2 domains to detect SA diseases	Leveraging from 2 domains to detect SA diseases

Figure 6.5: Summary of the intuition behind each experiment, based on the domain of the data used for training and testing

reverse experiment, where the training data come from in-the-wild conditions, and the labels refer to the self-reported health status of the speaker; and the test data come from CC, and the labels refer to the true health status of the speaker, will allow us to verify the validity of the proxy between the self-reported health status and true health status, specifically in the context of the WSM Corpus.

Finally, in the experiments where CC and in-the-wild data are combined during the training stage we will evaluate how to leverage from having data from multiple domains available during training, and whether that is preferable to training in a single domain. We repeat this for two scenarios, one where the test data come from CC, the other from in-the-wild conditions.

All the experiments are summarized in Table 6.5.

We note that all the experiments were performed for the two target SA diseases, depression and PD. Additionally, since all the proposed systems are disease agnostic, *i.e.*, they were not developed to target any specific SA disease, we argue that our systems can be easily replicated for other SA diseases, thus painting a more complete picture of the comparison between CC and in-the-wild conditions for the detection of SA diseases.

Evaluation

We opted for reporting the performance of all the models in UAR. This metric has been extensively used in multiple INTERSPEECH challenges since 2009, including in paralinguistic

tasks. Thus, making this metric a popular option in the speech community.

Following the evaluation based on UAR, the best performing models for each system, were chosen based on the criteria that the performance of the system on the development data were maximized, including any hyperparameter tuning, when applicable. In cases where different sets of hyperparameters yielded the same performance, the tie was broken by selecting the set with the best UAR on the test partition.

The implementation and technical details of each of the systems introduced in Section 6.1 are described below.

Experimental details

eGeMAPS with SVMs

In the experiments related to the eGeMAPS with the SVMs, we extracted the eGeMAPS for each 5 second long segment, using the OPENSMILE toolkit [91]. As such, each 5 second long segment is represented by a feature vector with 88 dimensions. The eGeMAPS are then passed to an SVM. Our SVM classifiers were implemented with python's *Scikit Learn* [170] using the Linear, RBF and Polynomial kernels. The best performing kernel and corresponding parameters were found through a grid search for each task. Classification was performed at the segment level and results at the file level were obtained through a majority voting.

i-vectors with PLDA

This system combined i-vectors as the feature extraction step, with PLDA as the classification step.

The i-vectors used in the scope of this work were extracted with the publicly available tool Kaldi [171], following the recipe *egs/voxceleb/v1*[1] using the pre trained voxceleb i-vector system[2]. As inputs to the i-vector system, we provided 25-dimensional feature vectors composed of 24 MFCCs and log-energy, extracted using a frame length of 25 ms, with 10 ms shift. Each frame was mean normalized over a sliding window of 300 ms and all non-speech frames were removed using energy-based Voice Activity Detection (VAD). The i-vectors were defined as 400-dimensional feature vectors.

The dimensionality of the resulting i-vectors was then reduced employing linear discriminant analysis (LDA), with resulting dimensionalities of $\{25, 50, 100, 200\}$, as well as without any reduction. Naturally, the transformation matrix for the dimensionality reduction corresponding

[1] https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v1
[2] https://kaldi-asr.org/models/m7

to the input dimension is the identity matrix. The LDA transformation matrix was applied to all the i-vectors. The LDA-reduced embeddings from the training partition were then used to train a PLDA classifier to discriminate speech affected by depression or PD from healthy speech. The PLDA provides a log-likelihood ratio (LLR) per audio segment with respect to the two mean LDA-reduced embeddings for the two classes. Afterwards, to assign a class to a PCA-reduced embedding once its LLR was obtained, this score was compared with a decision threshold, λ, which was determined by the Equal Error Rate (EER) operating point computed for the development set.

To obtain file-level predictions, we employed two strategies to combine the LLRs of all segments belonging to the same file. The first approach averages the LLRs of all segments belonging to the same file, and only after computes λ and provides one prediction for each file. The second approach computes λ using the LLR of the segments, makes one prediction for each segment and finally performs a majority vote on these predictions to obtain a final prediction for each file. The best strategy was chosen for each experiment. The performance we report corresponds to the LDA-reduced i-vectors that yielded the highest UAR on the developent set.

x-vectors with PLDA

This system, similarly to the i-vector one, presented in the Section above, combines speaker embeddings with PLDA. In this set of experiments we have replace the i-vectors with x-vectors, which we extracted following the Kaldi recipe *egs/voxceleb/v2*[3] using the pre-trained voxceleb x-vector system[4]. We provided as inputs to the x-vector system 30-dimensional feature vectors composed of 29 MFCCs + log-energy, extracted using a frame length of 25ms, with 10ms shift. Each frame was mean normalized over a sliding window 300 ms and all non-speech frames were removed using VAD. X-vectors were defined as 512-dimensional feature vectors.

After that, we again trained an LDA model with the x-vectors to reduce their dimensionality to $\{25, 50, 100, 200\}$, and also without any reduction. Then the reduced embeddings were used to train a PLDA model, following the same procedure as in the i-vector system, presented in the Section 6.2.2.

Once again we selected the LDA-reduced x-vectors that yielded the best UAR on the development set.

[3]https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2
[4]https://kaldi-asr.org/models/m7

CNN-LSTM with self-attention

As a pre-processing step, the audio is extracted from the interviews, and the silent frames are removed using VAD. Then the audio is divided into 700ms frames, with 50% overlap between consecutive segments. For each segment, the mel-spectrogram is computed by first decomposing the segment with a short time Fourier transform applied over $10ms$. The resulting spectrogram is integrated into 128 mel-spaced frequency bins, and then log-transformed. Hence, each 700ms segment is characterized by a 128 by 70 matrix, which is used as the network input.

The proposed network architecture, as shown previously in Fig 6.4, consists initially of a sequence of 3 convolutional layers that receive the mel-spectrograms as input. They had 32, 64 and 64 filters, and a kernel size of 3×3, 5×5 and 5×5, for the fist second and third layers respectively. Each convolutional layer has a ReLU activation, and is followed by a batch normalization and max-pooling operations. The max-pooling has a pool size of 3×3 and a stride of 2×2. This part of the network works as a feature extractor. The resulting hidden representation of the input is then flattened and passed to a 128-cell LSTM layer followed by a dropout layer with a dropout rate of 0.3, followed by another 128-cell LSTM layer. These recurrent layers of the network capture the temporal dependencies of the sequences of the hidden representations of the spectrograms. The network also includes a self-attention mechanism (also known as intra-attention), after the last LSTM layer, which enables the network to learn correlations between the different elements of the input sequence. Finally, the last layer of the network is simply a classification layer to perform the binary prediction between detecting speech affected by an SA disease vs healthy speech.

The model was trained with the RMSProp optimizer over 50 epochs with a learning rate of 0.0001.

Furthermore, we note that the examples were partitioned into 1 minute sequences which were processed individually. Each of the 1 min sequence inherited the label of the full examples. The predicted labels were then aggregated into a file level prediction using a voting scheme. This choice allowed us to achieve a compromise between the sequence length, processing time, and number of examples in the training dataset.

6.2.3 Results and discussion

In this Section we present the results obtained for the experiments described in Section 2.1. These results are organized not by modeling system but by combinations of train and test data domain.

Table 6.2: Results in UAR of the same domain experiments to detect depression.

Modeling Strategy	Performance for CC vs. CC [UAR%]	Performance for ITW vs. ITW [UAR%]
eGeMAPS + SVM	39.4	64.9
i-vectors + PLDA	50.0	79.7
x-vectors + PLDA	50.0	81.0
CNN-LSTM + Attn	58.0	81.0

Table 6.3: Results in UAR of the same domain experiments to detect PD.

Modeling Strategy	Performance for CC vs. CC [UAR%]	Performance for ITW vs. ITW [UAR%]
eGeMAPS + SVM	80.0	61.1
i-vectors + PLDA	78.3	69.6
x-vectors + PLDA	68.0	72.5
CNN-LSTM + Attn	82.0	73.0

Same domain experiments: Baseline

In this Section we report the performance of the four systems described in Section 6.2.2 to detect depression and PD, where the train and test data originate from the same domain.

Tables 6.2 and 6.3 show the performance in UAR% on the test partition of the data of four systems: eGeMAPS with SVM, i-vectors with PLDA, x-vectors with PLDA, and CNN-LSTM with self-attention, when trained and tested with either data collected in CC or in-the-wild conditions, for depression and PD, respectively. There are a total of 8 combinations of ML systems and data for each SA disease.

Focusing on the results obtained for the experiments relative to depression with CC data, which are shown in the first column of Table 6.2, we can observe that the performance of three out of four systems is chance level or worse. The exception is the CNN-LSTM with self-attention, which obtained a UAR of 58.0%.

The conclusion from this set of experiments is that detecting depression from speech alone in the DAIC-WOZ specifically is a difficult task.

The results relative to the systems trained and tested with data from in-the-wild conditions, the WSM Corpus, for depression can be found on the second column of Table 6.2. In this scenario, we obtain results that are more in line with what we would have expected: The eGeMAPS provide a reasonable baseline for the performance of this task, with a UAR of 64.9%; these are outperformed by the remaining systems, where, although with similar

performances, the CNN-LSTM with self-attention is the best performing model, with a UAR of 81.0%. The result relative to the CNN-LSTM with self-attention was obtained with about one third of the data used in the remaining three models, as referenced in Table 6.1.

We note that these results were obtained with systems trained and evaluated based with labels of self-reported diagnosis, not true diagnosis. A possible interpretation for these results is that examples from the two classes (one where the subjects self-diagnose with having depression, and another where they do not) clearly present different acoustic characteristics and that these were learned by our models. While this interpretation does not confirm our hypothesis that self-reported diagnosis equates to true diagnosis, it is a result that partially supports it.

In the case of PD, trained and tested with CC data, specifically a subset of the PC-GITA, the performance can be found on the first column of Table 6.3. In these experiments we were able to obtain, in the best case scenario, a UAR of 82.0%, with the CNN-LSTM with self-attention. Surprisingly the x-vectors were the worst performing model. We hypothesise that this may be related to the miss-match between the language of the PC-GITA (Spanish) and the language of the pre-trained x-vector extraction model (English).

Regarding the results of the detection of PD in in-the-wild conditions, they can be found on the second column of Table 6.3. In this case, we can observe, as with depression detected in in-the-wild conditions, that the eGeMAPS are the worse performing model, with a UAR of 61.1%, and the x-vectors and the CNN-LSTM with self-attention are the best performing ones, with UARs of 72.5% and 73.0%, respectively. We note that the results obtained with the CNN-LSTM with self-attention used about one quarter of the training data as the remaining models, as was referenced in Table 6.1 These results show that, for both depression and PD, x-vectors and the CNN-LSTM with self-attention are the models that can more robustly deal with variability of recording conditions, out of the studied ones.

Limitations of CC data

Moving on to the experiments where the training data come from CC, and the test data from in-the-wild conditions, the corresponding results can be found in Tables 6.4 and 6.5, for depression and PD, respectively.

As described before, the results of these experiments highlight the limitations that CC data have when dealing with data from in-the-wild conditions. To interpret these results, we should compare them to the in-the-wild baselines from the previous Section, in the second column of Tables 6.2 and 6.3, in the case of depression and PD, respectively. As can be

Table 6.4: Results in UAR of the cross domain experiments to detect depression, where the train data are from CC and the test data from in-the-wild conditions.

Modeling Strategy	**Performance for CC vs. ITW [UAR%]**
eGeMAPS + SVM	40.5
i-vectors + PLDA	52.7
x-vectors + PLDA	50.0
CNN-LSTM + Attn	62.0

Table 6.5: Results in UAR of the cross domain experiments to detect PD, where the train data are from CC and the test data from in-the-wild conditions.

Modeling Strategy	**Performance for CC vs. ITW [UAR%]**
eGeMAPS + SVM	39.3
i-vectors + PLDA	64.3
x-vectors + PLDA	50.0
CNN-LSTM + Attn	55.0

seen for the case of depression, comparing the results from Table 6.4 to the second column of Table 6.2, CC data cannot generalize well to data from a different domain, particularly in-the-wild and the decrease in performance ranges from 24.4 to 31.0 points in UAR. We performed statistical significance tests, specifically paired t-tests, to determine if there was a statistically significant difference between using CC, and in-the-wild data as training material to detect depression in in-the-wild conditions. We observed that using in-the-wild data as training material is significantly better than using CC data ($p<0.05$).

In the case of PD, the results from Table 6.5 should be compared to those on the second column of Table 6.2. As in the case of depression, we can also observe a significant drop in performance between the same and cross domain experiments. In this case, ranging from 5.3 to 22.5 points in UAR. However, considering the performance of all the modeling strategies, the paired t-tests showed that there is no statistically significant difference between using CC or in-the-wild data as training material to detect PD in in-the-wild conditions.

The results from these cross domain experiments, for both depression and PD, also raise the broader issue of whether currently existing models to detect SA diseases, trained on the existing CC data can be successfully used for real life applications. Furthermore, they should also be used as motivation to collect speech medical datasets that resemble real life conditions, such as the WSM Corpus.

Table 6.6: Results in UAR of the cross domain experiments to detect depression, where the train data are from in-the-wild conditions and the test data from CC.

Modeling Strategy	**Performance for ITW vs. CC [UAR%]**
eGeMAPS + SVM	59.8
i-vectors + PLDA	55.0
x-vectors + PLDA	78.0
CNN-LSTM + Attn	62.0

Table 6.7: Results in UAR of the cross domain experiments to detect PD, where the train data are from in-the-wild conditions and the test data from CC.

Modeling Strategy	**Performance for ITW vs. CC [UAR%]**
eGeMAPS + SVM	80.0
i-vectors + PLDA	53.0
x-vectors + PLDA	53.0
CNN-LSTM + Attn	72.0

Validation of proxy between self reported health status and true health status

In this Section, we will discuss the results obtained in the cross domain experiments where the training data come from in-the-wild sources, and the test data from CC. The performance of the four models, reported in UAR, can be found in Tables 6.6 and 6.7, for depression and PD, respectively.

To interpret the results of these experiments, we should keep in mind that, not only the domains of the train and test data are different, but also the labels: During the training stage, the models learn based on labels that correspond to the self-reported health status of the speaker, which we have hypothesised as being a good proxy to the true health status of the speaker. Then, during the evaluation stage, they are scored based on their capability to detect the true health status. In essence, we are assessing which training material, CC data with labels for true health status, or in-the-wild data with labels for self-reported health status, is better at detecting true health status in CC. In such a scenario, if a model trained with in-the-wild data and labels for self-reported health status, can perform close to, or better that the baseline models for CC data, reported in the first column of Tables 6.2 and 6.3, for depression and PD, respectively, we can interpret these results as a verification to our initial hypothesis, *i.e.*: **We verify that, in the context of the WSM Corpus, the self-reported health status is a valid proxy for the true health status**.

By comparing the results obtained for depression, in Table 6.6, with the ones on the first column of Table 6.2, we can see that the performance of all the models improves when training with in-the-wild data. The improvement ranges from 4.0 to 28.0 points in UAR. This is an interesting result that supports not only the validity of our hypothesised proxy, but also another hypothesis that in-the-wild data provide a more complete representation of the acoustic characteristics of a SA disease, in the case of depression. Furthermore, this result is considered statistically significant, as the paired t-test between the same models trained with CC and in-the-wild data yielded a p-value under 0.05.

In the case of PD, we should compare the performances reported in Table 6.7, with the ones on the first column of Table 6.3. Contrary to the case of depression, for PD we observe that, in the best case scenario we are able to obtain the same performance in the two experiments, in the case where the modeling strategy is based on eGeMAPS. In the remaining models, we observe a drop in performance from 10.0 to 25.3 points in UAR. We argue that there may be two reasons that justify this drop, one is the difference in language between the train (English) and test data (Spanish). The second, we hypothesise, relates to the different characteristics of the PD patients in the two corpora, particularly regarding age, where subjects present in the training examples, from CC, had an average age of 61 years old, and in the test examples, from in-the-wild conditions, 45 years old. In turn, this difference in average age, could also have implications in the severity of PD portrayed in the examples present in each of these two corpora.

Nevertheless, given that for the model based on eGeMAPS, the performance remains the same across the two experiments, we consider that this result also supports our hypothesis that self-reported health status is a good proxy for true health status.

In this set of experiments, we once again observed a statistically significant difference between using training material from CC and in-the-wild condition do detect PD in in-the-wild conditions ($p<0.05$).

Mixing data from CC and in-the-wild conditions

In this Section, we explore whether its advantageous to mix data from both domains during training, if available. For this we train the models with data from both CC and in-the-wild conditions, and test separately with data from each domain. The results of these experiments are reported in Tables 6.8 and 6.9, for depression and PD, respectively.

In the case of the detection of depression in CC, we can observe, by comparing the results in the first column of Tables 6.8, 6.2, and Table 6.4, that we only obtain improvements in the

Table 6.8: Results in UAR of the mixed domain experiments to detect depression, where the train data are from CC and in-the-wild conditions.

Modeling Strategy	**Performance for CC+ITW vs. CC [UAR%]**	**Performance for CC+ITW vs. ITW [UAR%]**
eGeMAPS + SVM	47.1	68.9
i-vectors + PLDA	54.0	70.3
x-vectors + PLDA	81.8	77.7
CNN-LSTM + Attn	68.0	69.0

Table 6.9: Results in UAR of the mixed domain experiments to detect PD, where the train data are from CC and in-the-wild conditions.

Modeling Strategy	**Performance for CC+ITW vs. CC [UAR%]**	**Performance for CC+ITW vs. ITW [UAR%]**
eGeMAPS + SVM	79.2	60.3
i-vectors + PLDA	61.0	73.2
x-vectors + PLDA	51.0	69.7
CNN-LSTM + Attn	90.0	75.0

performance of the models by mixing data from both domains during training of the x-vectors based model and the CNN-LSTM with self-attention. The remaining two models, based on i-vectors and eGeMAPS, could not take advantage of having data from both domains during training. The absolute best performing model is the CNN-LSTM with self-attention trained with a mixture of CC and in-the-wild data, which obtained a UAR of 81.8%. We did not observe a statistically significant difference between using either using data from a single domain (CC or in-the-wild) and both domains to detect depression in CC.

The best case scenario to detect depression in in-the-wild conditions, which we can observe by comparing the results shown in the second column of Tables 6.8, 6.2, and Table 6.6, is obtained by only using in-the-wild data during training, and adopting a modeling strategy based on either x-vectors or a CNN-LSTM with self-attention. In both of these cases we were able to obtain a UAR of 81.0%. Looking at each model individually, the only one which showed improvements by mixing data from the two domains during training was the one based on eGeMAPS. Its performance improved 4.0 points in UAR. These results show that the combination of the data from different domains may not be a problem with a straightforward solution. By performing paired t-tests, we observed that there was a statistically significant improvement from using both domains of data, instead of CC only to detect depression in in-the-wild conditions. This difference did not exist when comparing using data from both domains and data from in-the-wild conditions only.

In the case of PD, and specifically its detection in CC, we can compare the results in the first column of Tables 6.9, 6.3, and Table 6.5, to determine that the best strategy to perform this task is to use data from both domains during training and adopt a learning strategy based on CNN-LSTM with self-attention. In this scenario we obtained a UAR of 90.0%, which corresponds to an improvement of 8.0 points to the best single domain model. Once again, only the CNN-LSTM with self-attention was able to take advantage of having data from both domains during training. Considering all of the proposed models, we observed that there was no statistically significant difference between using only one domain of data during training and mixing data from both domains to detect PD in CC.

Finally, to study the impact of mixing data from different domains in the task of detecting PD in in-the-wild conditions, we should compare the results in the second column of Tables 6.3, 6.9, and Table 6.7. From these tables we can observe that, once again, the best strategy is to train a CNN-LSTM with self-attention with data from both domains. With it we can achieve a UAR of 75.0%, which corresponds to an increase in performance of 2.0 points when compared to the best single domain model. The model based on i-vectors was also capable of leveraging from having training data from both domains, which resulted in an improvement in the performance of 3.6 points, to a UAR of 73.2%. Similarly to the case of detecting depression in in-the-wild conditions, we observed that the improvement obtained by mixing the domains of the data during training, compared to using CC data only, was statistically significant ($p<0.05$), but this was not the case when using in-the-wild data only.

Overall we can see that, across most experiments, the leading strategy to detect SA diseases is based on using the CNN-LSTM with self-attention, which is the only strategy that can consistently leverage from combining data from two domains during training. However, in most experiments expect those dedicated to detect PD in in-the-wild conditions, this difference is not statistically significant.

Figures 6.6 and 6.7 summarize the performance in UAR of all the experiments in a more visually intuitive manner.

We should keep in mind, however, that this strategy to combine data is naive, and that should only be used as a starting point to mix data from different domains.

6.2.4 Final considerations

As we conclude this Chapter, it is interesting to look back at this Chapter as well as Chapter 5. In Chapter 5 we have addressed the problem of automating the collection and annotation process of speech medical corpora, particularly in the context of the WSM Corpus. Then, in

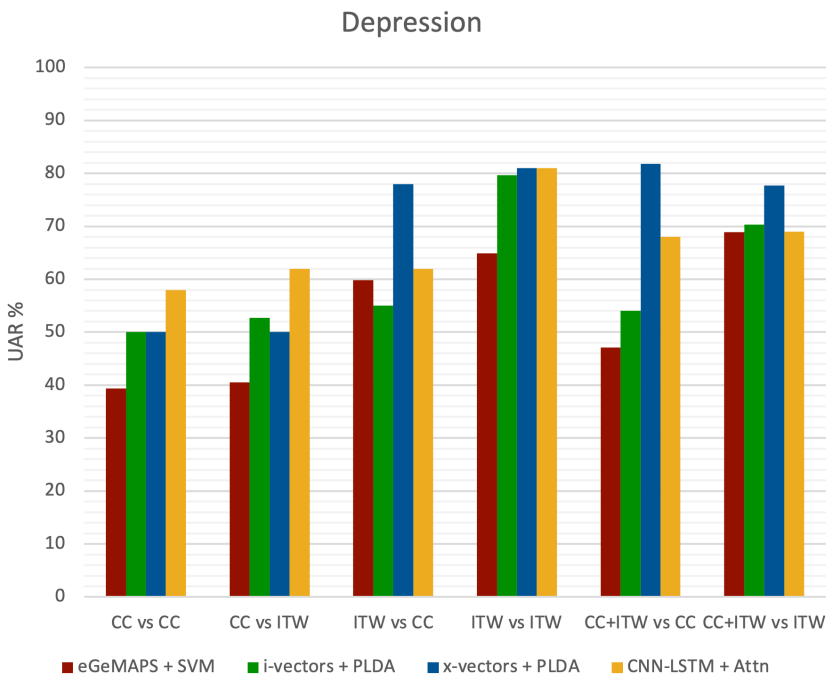


Figure 6.6: Performance in UAR% of the four strategies to detect depression, in both CC and in-the-wild conditions.

this Chapter, we have addressed how we could use the WSM Corpus to detect SA diseases in both CC and in-the-wild conditions. However, we have only looked these problems in isolation. We did not study or propose a complete pipeline that would take an automatically annotated dataset, as opposed to a manually annotated one, as was the case of this Chapter, and use it to detect a target SA disease. This task can prove difficult as the noise resulting from the automatic annotation task can compromise the performance of the SA disease detection task. It would be relevant to quantify the loss in performance due to this integration. This could be accomplished, for example, through an ablation study, where we would repeat the experiments performed in Section 5.5, with varying degrees of noise in the labels of the training data. This would lead to a better understating of the challenges posed by this integration problem, and as a starting point to overcome it.

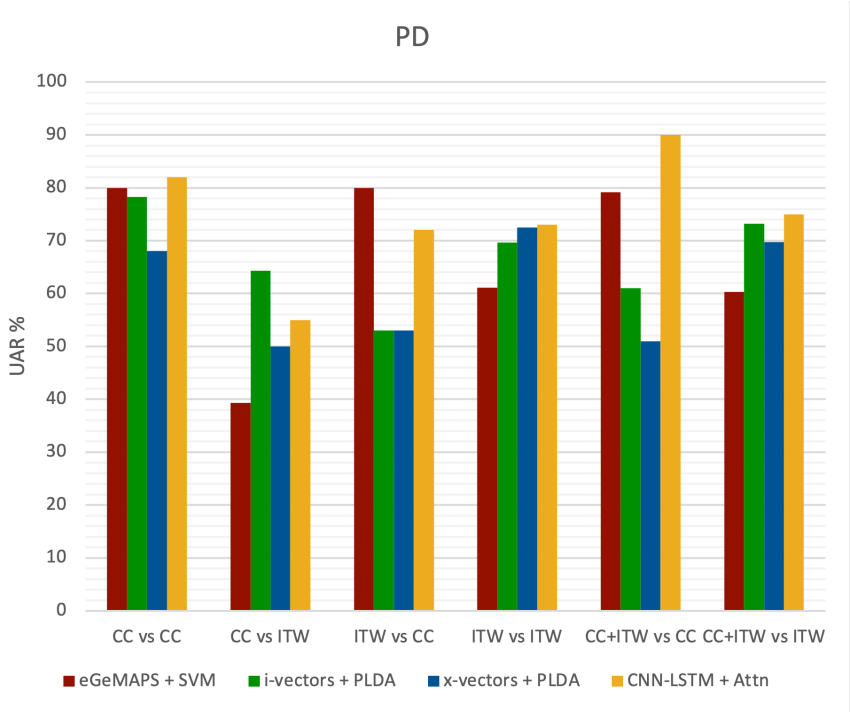


Figure 6.7: Performance in UAR% of the four strategies to detect PD, in both CC and in-the-wild conditions.

Part IV

Conclusion and future work

Chapter 7

Conclusions

In this thesis we have addressed the problem of detecting SA diseases particularly in in-the-wild, or real life conditions, thus addressing a problem that had not previously been tackled. While there exist numerous works related to the detection of several different SA diseases, they are restricted to CC, which, until now, have been the only conditions in which the detection of SA diseases has been studied.

In the scope of this thesis, we adopted two SA diseases as our working examples, depression and PD. We used them to experimentally verify our proposed solutions for the problem of detecting SA diseases in in-the-wild conditions. We were able to address the problem at every stage, from data collection and annotation, to problem modeling.

Over the course of this thesis we were able to make several contributions towards performing in-the-wild detection of SA diseases, of which we highlight the following:

- **WSM Corpus:** The WSM Corpus, which was collected and manually annotated in the scope of this thesis, is a first of its kind corpus of mutimodal recordings collected from online multimedia sources. The corpus currently contains roughy one thousand recordings of subjects affected by depression and PD. This corpus represents what is, to the best of our knowledge, the first publicly available speech medical corpus in in-the-wild conditions, as well as one of the largest speech medical corpora available, regardless of recording conditions.
- **Automatic corpora collection and annotation strategies:** In an effort to reduce the costs associated to the manual collection and annotation of corpora, specifically speech medical corpora, we have proposed several contributions to automate this process, using different amounts of labeled data during training, from 0 to 100%, and using

existing massive online multimedia repositories as the data source. We were able to experimentally test our proposed techniques on the WSM Corpus and verify their validity. At the same time, our contributions remained disease agnostic, allowing them to be easily used to automatically create speech medical datasets for SA diseases other than the two examples studied, depression and PD.

- **Study comparing the detection of SA in CC and in-the-wild conditions:** In the scope of this thesis, we were able to perform what is, to the best of our knowledge, the first study that compares the characteristics of in-the-wild vs CC speech medical data, and, at the same time, compare how current SA disease detection methods perform in both conditions. This was, again to the best of our knowledge, the first work that measured the limitations posed by diagnosis tools trained on CC data when faced with real life scenarios. We also showed that, specifically for the examples of depression and PD, in-the-wild data can be used to create a more complete acoustic representation of the characteristics of these diseases than their CC counterparts.

- **In-the-wild SA diseases detection**: Arguably the most important contribution of this thesis was proposing novel strategies to detect SA diseases that we experimentally demonstrated to be successful in detecting SA diseases in-the-wild conditions. We were able to show this for two very different SA diseases, depression and PD, while adhering to our initial goal of proposing solutions that were disease agnostics, *i.e.* did not take advantage of any domain specific knowledge for the target SA disease. This choice allows future works to leverage from the contributions of this thesis, to create a starting point to the detection of other SA diseases for which there does not need to exist any prior domain knowledge.

The work developed in the scope of this thesis can have several immediate real life applications, all of which contribute to democratizing the access to quality healthcare. They include:

- **Diagnosis aid tools:** These technologies can be used in tandem with appropriate medical specialists, in a hospital scenario, to provide complementary insight on the patient's health status. They would not be restricted to the typical CC where specific speaking exercises and recording conditions are necessary. Our proposed solutions could be implemented simply as an app running in the background of a medical appointment, providing additional real life feedback to the healthcare specialist.

- **Patient monitoring in clinical trials:** Clinical trials are lengthy and costly due to their meticulous nature. The technologies developed in the scope of this thesis could be used to provide better monitoring of the trial subjects, by providing them with, for

example, a smartphone with an SA disease monitoring app. This might contribute to reducing the amount of time necessary to assess the effects of a new drug or treatment, and consequently, reducing the length and cost of clinical trials.

- **Personal health monitoring tool:** Running as background applications on any personal devices, including smartphones, smartwatches, personal computers, digital assistants, *etc.*, to provide frequent monitoring of a subjects health status.

Chapter 8

Future work

In this thesis we have laid the ground for the in-the-wild detection of SA diseases. As one of the first publicly available works dedicated to the detection of SA diseases specifically in real-life scenarios, there is still a lot of room to further develop the work that was presented in this thesis. Some of the future work directions of this thesis include:

- **WSM Corpus:** The WSM Corpus is an ever growing corpus, as such it will continue to be expanded, both in term of the existing datasets (for depression and PD), as well as datasets for other SA diseases. Our short term goals include creating new datasets in the WSM Corpus dedicated to OSA, the flu, and COVID-19, expanding the existing depression and PD datasets, and creating longitudinal datasets where samples from the same speaker are collected over long periods of time. This would, not only result in a larger and more diverse resource of speech medical data, but would also allow the study of more SA diseases, other than the ones contemplated in this thesis.

- **Other modalities:** In the scope of this thesis we have focused on performing the detection of SA diseases from the speech modality alone. However, the WSM Corpus, as a multi-modal corpus, allows the study of these diseases using other modalities. We highlight the visual modality as the natural next step to study diseases such as PD, known to have several motor symptoms, which are already present in the corpus, thus enriching the information used to perform the diagnosis.

- **Other languages:** In this thesis we have focused on detecting SA diseases in English. At the same time, the WSM Corpus is currently also restricted to videos in the English language. It would be relevant to not only collect data in other languages, but also to compare the detection of a given SA across different languages.

- **Longitudinal studies:** An aspect that is of great relevance that was not addressed in this thesis, relates to the study of chronic SA diseases over time. It would be of great relevance to be able to study, for example, how the speech impairments of patients affected by PD evolve over the years, thus establishing a correlation between voice and disease severity.

- **Adopting disease specific strategies:** We have emphasised several times that our solutions to the problem of detecting SA diseases would be disease agnostic. In a way, our goal was to have this thesis be a sort of template for future works in the detection of SA diseases in real life conditions. We have attempted, over the course of this thesis, to present solutions that can be, on one hand used as a starting point for the study of other SA diseases, and on the other hand, easily adapted and specialized to deal with specific SA diseases. With this in mind it would be relevant to adapt the solutions proposed in this thesis so that they would be specific to either depression or PD, among other SA diseases, and perform a study on the impact of such changes. Naturally, we would expect the performance of the disease specific solutions to improve over the solutions presented in this thesis.

- **Leveraging from multiple domains of data to detect SA diseases:** In Chapter 6, we have performed several experiments to detect SA diseases, including experiments where we attempted to leverage from having data from both CC and in-the-wild domains to detect SA diseases. In these experiments it was not clear that we were able to benefit from having data from the two domains. Instead of naively mixing the data, more sophisticated methods should be studied in order to take advantage of having data from both domains.

- **Real world applications:** Finally, we believe it would be very relevant to apply the work proposed in this thesis to real life applications, *e.g.* by having smartphone applications, smartwatches, virtual assistants, *etc.* use the proposed solutions and offering the possibility to monitor and detect SA diseases in real life.

Looking back at the work developed in this thesis, and at the future work we propose, it is important to keep in mind the essence of the problem we are attempting to solve. We are attempting to improve the quality and accessibility of medical diagnosis, thus, to improve humanity's quality of life. This thesis is a small contribution towards this goal.

However, given the sensitive nature of this goal, or any other health related goals, it is paramount to maintain a very conservative perspective when making any claims derived from experimental results.

In the context of this thesis, can we claim that we have solved the problem of detecting SA diseases in in-the-wild conditions? Or are we restricted to a smaller claim, related to depression and PD alone, and even then, only in the context of the studied datasets? And what is the validity of the results we present, given the small amount of data, specifically CC data, that we used in the scope of this thesis, to attempt to answer the research questions posed in Section 1? These, and other questions will remain open to discussion, and our work should be viewed as one more piece of evidence to form consistent scientific consensus around the problem of detecting SA diseases, of any nature, in any conditions.

Bibliography

[1] C. Jiao, X. Du, and A. Zare, "Addressing the inevitable imprecision: Multiple instance learning for hyperspectral image analysis," in *Hyperspectral Image Analysis*. Springer, 2020, pp. 141–185.

[2] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in neural information processing systems*, 2002, pp. 561–568.

[3] M. J. Patel, A. Khalaf, and H. J. Aizenstein, "Studying depression using imaging and machine learning methods," *NeuroImage: Clinical*, vol. 10, pp. 115–123, 2016.

[4] C. Salvatore, A. Cerasa, I. Castiglioni, F. Gallivanone, A. Augimeri, M. Lopez, G. Arabia, M. Morelli, M. Gilardi, and A. Quattrone, "Machine learning on brain mri data for differential diagnosis of parkinson's disease and progressive supranuclear palsy," *Journal of neuroscience methods*, vol. 222, pp. 230–237, 2014.

[5] C. Plant, S. J. Teipel, A. Oswald, C. Böhm, T. Meindl, J. Mourao-Miranda, A. W. Bokde, H. Hampel, and M. Ewers, "Automated detection of brain atrophy patterns based on mri for the prediction of alzheimer's disease," *Neuroimage*, vol. 50, no. 1, pp. 162–174, 2010.

[6] B. Hosseinifard, M. H. Moradi, and R. Rostami, "Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from eeg signal," *Computer methods and programs in biomedicine*, vol. 109, no. 3, pp. 339–345, 2013.

[7] H. Korkalainen, J. Aakko, S. Nikkonen, S. Kainulainen, A. Leino, B. Duce, I. O. Afara, S. Myllymaa, J. Töyräs, and T. Leppänen, "Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea," *IEEE journal of biomedical and health informatics*, vol. 24, no. 7, pp. 2073–2081, 2019.

[8] M. D. Crutcher, R. Calhoun-Haney, C. M. Manzanares, J. J. Lah, A. I. Levey, and S. M. Zola, "Eye tracking during a visual paired comparison task as a predictor of early dementia," *American Journal of Alzheimer's Disease & Other Dementias®*, vol. 24, no. 3, pp. 258–266, 2009.

[9] Y. Zhang, T. Wilcockson, K. I. Kim, T. Crawford, H. Gellersen, and P. Sawyer, "Monitoring dementia with automatic eye movements analysis," in *Intelligent Decision Technologies 2016*. Springer, 2016, pp. 299–309.

[10] A. Y. Meigal, K. S. Prokhorov, N. A. Bazhenov, L. I. Gerasimova-Meigal, and D. G. Korzun, "Towards a personal at-home lab for motion video tracking in patients with parkinson's disease," in *2017 21st Conference of Open Innovations Association (FRUCT)*. IEEE, 2017, pp. 231–237.

[11] R. C. Kessler *et al.*, "Lifetime prevalence and age-of-onset distributions of mental disorders in the world health organization's world mental health survey initiative," *World psychiatry*, vol. 6, no. 3, p. 168, 2007.

[12] T. Üstün, J. Ayuso-Mateos, S. Chatterji, C. Mathers, and C. Murray, "Global burden of depressive disorders in the year 2000," *The British journal of psychiatry*, vol. 184, no. 5, pp. 386–392, 2004.

[13] W. H. Organization, "Depression: A global public health concern," https://www.who.int/mental_health/management/depression/who_paper_depression_wfmh_2012.pdf.

[14] J. Olesen, A. Gustavsson, M. Svensson, H.-U. Wittchen, B. Jönsson, C. S. Group, and E. B. Council, "The economic cost of brain disorders in europe," *European journal of neurology*, vol. 19, no. 1, pp. 155–162, 2012.

[15] R. C. Kessler, S. Aguilar-Gaxiola, J. Alonso, S. Chatterji, S. Lee, J. Ormel, T. B. Üstün, and P. S. Wang, "The global burden of mental disorders: an update from the who world mental health (wmh) surveys," *Epidemiology and Psychiatric Sciences*, vol. 18, no. 1, pp. 23–33, 2009.

[16] P. E. Greenberg, R. C. Kessler, T. L. Nells, S. N. Finkelstein, and E. R. Berndt, "Depression in the workplace: an economic perspective," *Selective Serotonin Reuptake Inhibitors: Advances in Basic Research and Clinical Practice. New York: John Wiley & Sons Ltd*, 1996.

[17] K. Sanderson, E. Tilse, J. Nicholson, B. Oldenburg, and N. Graves, "Which presenteeism measures are more sensitive to depression and anxiety?" *Journal of affective disorders*, vol. 101, no. 1-3, pp. 65–74, 2007.

[18] A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[19] K. Kroenke and R. L. Spitzer, "The phq-9: a new depression diagnostic and severity measure," *Psychiatric annals*, vol. 32, no. 9, pp. 509–515, 2002.

[20] C. Sobin and H. A. Sackeim, "Psychomotor symptoms of depression," *American Journal of Psychiatry*, vol. 154, no. 1, pp. 4–17, 1997.

[21] G. Parker, D. Hadzi-Pavlovic, H. Brodaty, P. Boyce, P. Mitchell, K. Wilhelm, I. Hickie, and K. Eyers, "Psychomotor disturbance in depression: defining the constructs," *Journal of affective disorders*, vol. 27, no. 4, pp. 255–265, 1993.

[22] C. Sobin, L. Mayer, and J. Endicott, "The motor agitation and retardation scale: a scale for the assessment of motor abnormalities in depressed patients," *The Journal of neuropsychiatry and clinical neurosciences*, vol. 10, no. 1, pp. 85–92, 1998.

[23] D. J. Widlöcher, "Psychomotor retardation: clinical, theoretical, and psychometric aspects." *Psychiatric Clinics of North America*, 1983.

[24] H. Ellgring and K. R. Scherer, "Vocal indicators of mood change in depression," *Journal of Nonverbal Behavior*, vol. 20, no. 2, pp. 83–110, 1996.

[25] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. Gailey, and C. Levinton, "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression," *Journal of psychiatric research*, vol. 27, no. 3, pp. 309–319, 1993.

[26] J. K. Darby and H. Hollien, "Vocal and speech patterns of depressive patients," *Folia Phoniatrica et Logopaedica*, vol. 29, no. 4, pp. 279–291, 1977.

[27] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 9, pp. 1530–1540, 2004.

[28] R. Kliper, S. Portuguese, and D. Weinshall, "Prosodic analysis of speech and the underlying mental state," in *International Symposium on Pervasive Computing Paradigms for Mental Health*. Springer, 2015, pp. 52–62.

[29] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, 2010.

[30] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," *Journal of neurology, neurosurgery & psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[31] L. M. De Lau and M. M. Breteler, "Epidemiology of parkinson's disease," *The Lancet Neurology*, vol. 5, no. 6, pp. 525–535, 2006.

[32] M. d. De Rijk, C. Tzourio, M. Breteler, J. Dartigues, L. Amaducci, S. Lopez-Pousa, J. Manubens-Bertran, A. Alperovitch, and W. A. Rocca, "Prevalence of parkinsonism and parkinson's disease in europe: the europarkinson collaborative study. european community concerted action on the epidemiology of parkinson's disease." *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 62, no. 1, pp. 10–15, 1997.

[33] W. Yang, J. L. Hamilton, C. Kopil, J. C. Beck, C. M. Tanner, R. L. Albin, E. R. Dorsey, N. Dahodwala, I. Cintina, P. Hogan *et al.*, "Current and projected future economic burden of parkinson's disease in the us," *npj Parkinson's Disease*, vol. 6, no. 1, pp. 1–9, 2020.

[34] J. Parkinson, "An essay on the shaking palsy," *The Journal of neuropsychiatry and clinical neurosciences*, vol. 14, no. 2, pp. 223–236, 2002.

[35] P. A. Kempster, D. R. Williams, M. Selikhova, J. Holton, T. Revesz, and A. J. Lees, "Patterns of levodopa response in parkinson's disease: a clinico-pathological study," *Brain*, vol. 130, no. 8, pp. 2123–2128, 2007.

[36] A. J. Hughes, S. E. Daniel, L. Kilford, and A. J. Lees, "Accuracy of clinical diagnosis of idiopathic parkinson's disease: a clinico-pathological study of 100 cases." *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 55, no. 3, pp. 181–184, 1992.

[37] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel *et al.*, "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results," *Movement disorders: official journal of the Movement Disorder Society*, vol. 23, no. 15, pp. 2129–2170, 2008.

[38] P. Enderby, "Frenchay dysarthria assessment," *British Journal of Disorders of Communication*, vol. 15, no. 3, pp. 165–173, 1980.

[39] P. M. Enderby and R. Palmer, *Frenchay dysarthria assessment*. Pro-ed, 2008.

[40] L. O. Ramig, C. Fox, and S. Sapir, "Speech treatment for parkinson's disease," *Expert Review of Neurotherapeutics*, vol. 8, no. 2, pp. 297–309, 2008.

[41] J. Müller, G. K. Wenning, M. Verny, A. McKee, K. R. Chaudhuri, K. Jellinger, W. Poewe, and I. Litvan, "Progression of dysarthria and dysphagia in postmortem-confirmed parkinsonian disorders," *Archives of neurology*, vol. 58, no. 2, pp. 259–264, 2001.

[42] K. K. Baker, L. O. Ramig, E. S. Luschei, and M. E. Smith, "Thyroarytenoid muscle activity associated with hypophonia in parkinson disease and aging," *Neurology*, vol. 51, no. 6, pp. 1592–1598, 1998.

[43] S. Perez-Lloret, L. Nègre-Pagès, A. Ojero-Senard, P. Damier, A. Destée, F. Tison, M. Merello, O. Rascol, and C. S. Group, "Oro-buccal symptoms (dysphagia, dysarthria, and sialorrhea) in patients with parkinson's disease: preliminary analysis from the french copark cohort," *European journal of neurology*, vol. 19, no. 1, pp. 28–37, 2012.

[44] N. A. Leopold and M. C. Kagel, "Prepharyngeal dysphagia in parkinson's disease," *Dysphagia*, vol. 11, no. 1, pp. 14–22, 1996.

[45] J. Kalf, B. Bloem, and M. Munneke, "Diurnal and nocturnal drooling in parkinson's disease," *Journal of neurology*, vol. 259, no. 1, pp. 119–123, 2012.

[46] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of speech and hearing research*, vol. 12, no. 2, pp. 246–269, 1969.

[47] J. A. Logemann and H. B. Fisher, "Vocal tract control in parkinson's disease," *Journal of Speech and Hearing Disorders*, vol. 46, no. 4, pp. 348–352, 1981.

[48] R. E. Bartt and J. L. Topel, "Chapter 50 - autoimmune and inflammatory disorders," in *Textbook of Clinical Neurology (Third Edition)*, third edition ed., C. G. Goetz, Ed. Philadelphia: W.B. Saunders, 2007, pp. 1155–1184. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9781416036180100505

[49] R. Matison, R. Mayeux, J. Rosen, and S. Fahn, ""tip-of-the-tongue" phenomenon in parkinson disease," *Neurology*, vol. 32, no. 5, pp. 567–567, 1982.

[50] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews." in *LREC*. Citeseer, 2014, pp. 3123–3128.

[51] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet *et al.*, "Simsensei kiosk: A virtual human interviewer for

healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1061–1068.

[52] E. B. Blanchard, J. Jones-Alexander, T. C. Buckley, and C. A. Forneris, "Psychometric properties of the ptsd checklist (pcl)," *Behaviour research and therapy*, vol. 34, no. 8, pp. 669–673, 1996.

[53] C. D. Spielberger, "State-trait anxiety inventory," *The Corsini encyclopedia of psychology*, pp. 1–1, 2010.

[54] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New spanish speech corpus database for the analysis of people suffering from parkinson's disease." in *LREC*, 2014, pp. 342–347.

[55] M. D. S. T. F. on Rating Scales for Parkinson's Disease, "The unified parkinson's disease rating scale (updrs): status and recommendations," *Movement Disorders*, vol. 18, no. 7, pp. 738–750, 2003.

[56] C. G. Goetz, W. Poewe, O. Rascol, C. Sampaio, G. T. Stebbins, C. Counsell, N. Giladi, R. G. Holloway, C. G. Moore, G. K. Wenning *et al.*, "Movement disorder society task force report on the hoehn and yahr staging scale: status and recommendations the movement disorder society task force on rating scales for parkinson's disease," *Movement disorders*, vol. 19, no. 9, pp. 1020–1028, 2004.

[57] N. Cummins, M. Schmitt, S. Amiriparian, J. Krajewski, and B. Schuller, ""you sound ill, take the day off": Automatic recognition of speech affected by upper respiratory tract infection," in *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*. IEEE, 2017, pp. 3806–3809.

[58] F. Boller and J. Becker, "Dementiabank database guide," *University of Pittsburgh*, 2005.

[59] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[60] F. Ringeval, J. Demouy, G. Szaszak, M. Chetouani, L. Robel, J. Xavier, D. Cohen, and M. Plaza, "Automatic intonation recognition for the prosodic assessment of language-impaired children," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1328–1342, 2010.

[61] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli, P. K. Ghosh, S. Ganapathy *et al.*, "Coswara–a database of breathing, cough, and voice sounds for covid-19 diagnosis," *arXiv preprint arXiv:2005.10548*, 2020.

[62] A. Dhall, R. Goecke, S. Lucey, T. Gedeon *et al.*, "Collecting large, richly annotated facial-expression databases from movies," *IEEE multimedia*, vol. 19, no. 3, pp. 34–41, 2012.

[63] S. Newman and V. G. Mather, "Analysis of spoken language of patients with affective disorders," *American journal of psychiatry*, vol. 94, no. 4, pp. 913–942, 1938.

[64] B. Pope, T. Blass, A. W. Siegman, and J. Raher, "Anxiety and depression in speech." *Journal of Consulting and Clinical Psychology*, vol. 35, no. 1p1, p. 128, 1970.

[65] W. A. Hargreaves, J. Starkweather, and K. Blacker, "Voice quality in depression." *Journal of Abnormal Psychology*, vol. 70, no. 3, p. 218, 1965.

[66] E. Whitman and D. Flicker, "A potential new measurement of emotional state: A preliminary report," *Newark Beth-Israel Hospital*, vol. 17, pp. 167–172, 1966.

[67] N. C. Andreasen, M. Alpert, and M. J. Martz, "Acoustic analysis: an objective measure of affective flattening," *Archives of General Psychiatry*, vol. 38, no. 3, pp. 281–285, 1981.

[68] M. F. Valstar, J. Gratch, B. W. Schuller, F. Ringeval, D. Lalanne, M. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016 - depression, mood, and emotion recognition workshop and challenge," *CoRR*, vol. abs/1605.01600, 2016. [Online]. Available: http://arxiv.org/abs/1605.01600

[69] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

[70] L. Yang, D. Jiang, L. He, E. Pei, M. Oveneke, and H. Sahli, "Decision tree based depression classification from audio video and language information," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 89–96.

[71] J. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H. Kung, C. Dagli, and T. Quatieri, "Detecting depression using vocal, facial and semantic communication cues," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 11–18.

[72] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 43–50.

[73] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 35–42.

[74] A. Pampouchidou, O. Simantiraki, A. Fazlollahi, M. Pediaditis, D. Manousos, A. Roniotis, G. Giannakakis, F. Meriaudeau, P. Simos, K. Marias *et al.*, "Depression assessment by fusing high and low level features from audio, video, and text," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 27–34.

[75] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, and D. Jiang, "Hybrid depression classification and estimation from audio video and text information," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 45–51.

[76] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal measurement of depression using deep learning models," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 53–59.

[77] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 4 2016, open access.

[78] W. Fan, Z. He, X. Xing, B. Cai, and W. Lu, "Multi-modality depression detection via multi-scale temporal dilated cnns," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 73–80.

[79] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level attention network using text, audio and video for depression prediction," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 81–88.

[80] M. Schmitt and B. Schuller, "Openxbow: introducing the passau open-source crossmodal bag-of-words toolkit," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 3370–3374, 2017.

[81] L. Yang, D. Jiang, and H. Sahli, "Feature augmenting networks for improving depression severity estimation from speech signals," *IEEE Access*, vol. 8, pp. 24 033–24 045, 2020.

[82] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, "Hierarchical attention transfer networks for depression assessment from speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7159–7163.

[83] M. Little, P. McSharry, E. Hunter, J. Spielman, and L. Ramig, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *Nature Precedings*, pp. 1–1, 2008.

[84] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *IEEE transactions on biomedical engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[85] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech," *Journal of speech, language, and hearing research*, 2010.

[86] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, "Imprecise vowel articulation as a potential early marker of parkinson's disease: Effect of speaking task," *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2171–2181, 2013.

[87] S. Skodda, W. Grönheit, and U. Schlegel, "Intonation and speech rate in parkinson's disease: General and dynamic aspects and responsiveness to levodopa admission," *Journal of Voice*, vol. 25, no. 4, pp. e199–e205, 2011.

[88] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson's disease," *The journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, 2011.

[89] T. Bocklet, S. Steidl, E. Nöth, and S. Skodda, "Automatic evaluation of parkinson's speech-acoustic, prosodic and voice related cues." in *Interspeech*, 2013, pp. 1149–1153.

[90] A. Bayestehtashk, M. Asgari, I. Shafran, and J. McNames, "Fully automated assessment of the severity of parkinson's disease from speech," *Computer speech & language*, vol. 29, no. 1, pp. 172–185, 2015.

[91] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[92] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The interspeech 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[93] T. Grósz, R. Busa-Fekete, G. Gosztolya, and L. Tóth, "Assessing the degree of nativeness and parkinson's condition using gaussian processes and deep rectifier neural networks," 2015.

[94] M. Tu, V. Berisha, and J. Liss, "Interpretable objective assessment of dysarthric speech based on deep neural networks." in *INTERSPEECH*, 2017, pp. 1849–1853.

[95] M. Cernak, J. R. Orozco-Arroyave, F. Rudzicz, H. Christensen, J. C. Vásquez-Correa, and E. Nöth, "Characterisation of voice quality of parkinson's disease using differential phonological posterior features," *Computer Speech & Language*, vol. 46, pp. 196–208, 2017.

[96] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth, "Convolutional neural network to model articulation impairments in patients with parkinson's disease." in *INTERSPEECH*, 2017, pp. 314–318.

[97] J. Vásquez-Correa, J. Orozco-Arroyave, T. Bocklet, and E. Nöth, "Towards an automatic evaluation of the dysarthria level of patients with parkinson's disease," *Journal of communication disorders*, vol. 76, pp. 21–36, 2018.

[98] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect parkinson's disease from speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1155–1159.

[99] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[100] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 726–733.

[101] J. Mallela, A. Illa, B. Suhas, S. Udupa, Y. Belur, N. Atchayaram, R. Yadav, P. Reddy, D. Gope, and P. K. Ghosh, "Voice based classification of patients with amyotrophic lateral sclerosis, parkinson's disease and healthy controls with cnn-lstm using transfer

learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6784–6788.

[102] A. Ghoshal, P. Ircing, and S. Khudanpur, "Hidden markov models for automatic annotation and content-based retrieval of images and video," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005, pp. 544–551.

[103] R. Segal, T. Markowitz, and W. Arnold, "Fast uncertainty sampling for labeling large e-mail corpora." in *CEAS*. Citeseer, 2006.

[104] Z. Yang, W. W. Cohen, and R. Salakhutdinov, "Revisiting semi-supervised learning with graph embeddings," *arXiv preprint arXiv:1603.08861*, 2016.

[105] H. Kuehne, A. Richard, and J. Gall, "Weakly supervised learning of actions from transcripts," *Computer Vision and Image Understanding*, vol. 163, pp. 78–89, 2017.

[106] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 368–373.

[107] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[108] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educational and psychological measurement*, vol. 33, no. 3, pp. 613–619, 1973.

[109] J. Correia, B. Raj, I. Trancoso, and F. Teixeira, "Mining multimodal repositories for speech affecting diseases," 2018.

[110] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[111] D. Palaz, M. Magimai.-Doss, and R. Collobert, "Analysis of cnn-based speech recognition system using raw speech as input," Idiap, Tech. Rep., 2015.

[112] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[113] S. Oviatt, A. DeAngeli, and K. Kuhn, "Integration and synchronization of input modes during multimodal human-computer interaction," in *Referring Phenomena in a Multimedia Context and their Computational Treatment*. Association for Computational Linguistics, 1997, pp. 1–13.

[114] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[115] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "Identifying human behaviors using synchronized audio-visual cues," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 54–66, 2017.

[116] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.

[117] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2005, pp. 115–124.

[118] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st annual meeting of the association for computational linguistics*, 2003.

[119] J. Correia, B. Raj, and I. Trancoso, "Querying depression vlogs," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 987–993.

[120] J. Correia, I. Trancoso, and B. Raj, "Adaptation of svm for mil for inferring the polarity of movies and movie reviews," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 258–264.

[121] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[122] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE reviews in biomedical engineering*, vol. 10, pp. 213–234, 2017.

[123] O. Z. Kraus, J. L. Ba, and B. J. Frey, "Classifying and segmenting microscopy images with deep multiple instance learning," *Bioinformatics*, vol. 32, no. 12, pp. i52–i59, 2016.

[124] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *arXiv preprint arXiv:1802.04712*, 2018.

[125] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 3, pp. 394–410, 2007.

[126] B. Liu, Y. Xiao, and Z. Hao, "A selective multiple instance transfer learning method for text categorization problems," *Knowledge-Based Systems*, vol. 141, pp. 178–187, 2018.

[127] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *Advances in neural information processing systems*, 2006, pp. 1417–1424.

[128] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3460–3469.

[129] J. Correia, I. Trancoso, and B. Raj, "Automatic in-the-wild dataset annotation with deep generalized multiple instance learning," in *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, 2020, pp. 3542–3550.

[130] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[131] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[132] G. An, D. G. Brizan, M. Ma, M. Morales, A. R. Syed, and A. Rosenberg, "Automatic recognition of unified parkinson's disease rating from speech with acoustic, i-vector and phonotactic features," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[133] J. Correia, F. Teixeira, C. Botelho, I. Trancoso, and B. Raj, "The in-the-wild speech medical corpus," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[134] J. Correia, I. Trancoso, and B. Raj, "In-the-wild end-to-end detection of speech affecting diseases," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 734–741.

[135] N. Srimadhur and S. Lalitha, "An end-to-end model for detection and assessment of depression levels using speech," *Procedia Computer Science*, vol. 171, pp. 12–21, 2020.

[136] F. Povolny, P. Matejka, M. Hradis, A. Popková, L. Otrusina, P. Smrz, I. Wood, C. Robin, and L. Lamel, "Multimodal emotion recognition for avec 2016 challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 75–82.

[137] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Nöth, "Multimodal assessment of parkinson's disease: a deep learning approach," *IEEE journal of biomedical and health informatics*, vol. 23, no. 4, pp. 1618–1630, 2018.

[138] E. Marchi, B. Schuller, S. Baron-Cohen, O. Golan, S. Bölte, P. Arora, and R. Häb-Umbach, "Typicality and emotion in the voice of children with autism spectrum condition: Evidence across three languages," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[139] A. Pompili, R. Solera-Urena, A. Abad, R. Cardoso, I. Guimaraes, M. Fabbri, I. P. Martins, and J. Ferreira, "Assessment of parkinson's disease medication state through automatic speech analysis," *arXiv preprint arXiv:2005.14647*, 2020.

[140] B. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder, E. Bergelson *et al.*, "The interspeech 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity," 2019.

[141] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[142] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "ivectors for continuous emotion recognition," *Training*, vol. 45, p. 50, 2014.

[143] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Twelfth annual conference of the international speech communication association*, 2011.

[144] M. H. Bahari, M. McLaren, D. A. van Leeuwen *et al.*, "Speaker age estimation using i-vectors," *Engineering Applications of Artificial Intelligence*, vol. 34, pp. 99–108, 2014.

[145] M. Dorfer, B. Lehner, H. Eghbal-zadeh, H. Christop, P. Fabian, and W. Gerhard, "Acoustic scene classification with fully convolutional neural networks and i-vectors," *Proceedings of the Detection and Classification of Acoustic Scenes and Events*, 2018.

[146] H. Eghbal-zadeh, B. Lehner, M. Dorfer, and G. Widmer, "A hybrid approach with multi-channel i-vectors and convolutional neural networks for acoustic scene classification," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 2749–2753.

[147] Y. Hauptman, R. Aloni-Lavi, I. Lapidot, T. Gurevich, Y. Manor, S. Naor, N. Diamant, and I. Opher, "Identifying distinctive acoustic and spectral features in parkinson's disease." in *Interspeech*, 2019, pp. 2498–2502.

[148] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM, Montreal,(Report) CRIM-06/08-13*, vol. 14, pp. 28–29, 2005.

[149] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[150] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017, pp. 999–1003.

[151] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors." in *Odyssey*, 2018, pp. 105–111.

[152] H. Zeinali, L. Burget, and J. Cernocky, "Convolutional neural networks and x-vector embedding for dcase2018 acoustic scene classification challenge," *arXiv preprint arXiv:1810.04273*, 2018.

[153] M. Diez, L. Burget, S. Wang, J. Rohdin, and J. Cernockỳ, "Bayesian hmm based x-vector clustering for speaker diarization." in *INTERSPEECH*, 2019, pp. 346–350.

[154] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7169–7173.

[155] S. Zargarbashi and B. Babaali, "A multi-modal feature embedding approach to diagnose alzheimer disease from spoken language," *arXiv preprint arXiv:1910.00330*, 2019.

[156] J. M. Perero-Codosero, F. Espinoza-Cuadros, J. Antón-Martín, M. A. Barbero-Álvarez, and L. A. Hernández-Gómez, "Modeling obstructive sleep apnea voices using deep neural network embeddings and domain-adversarial training," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 240–250, 2019.

[157] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[158] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, vol. 650, 2017.

[159] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[160] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*. PMLR, 2014, pp. 1764–1772.

[161] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.

[162] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," *arXiv preprint arXiv:1804.05160*, 2018.

[163] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates," *Sensors*, vol. 18, no. 10, p. 3418, 2018.

[164] B. Wu, K. Li, F. Ge, Z. Huang, M. Yang, S. M. Siniscalchi, and C.-H. Lee, "An end-to-end deep learning approach to simultaneous speech dereverberation and acoustic modeling for robust speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1289–1300, 2017.

[165] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE international conference on acoustics, speech and*

signal processing (ICASSP). IEEE, 2018, pp. 5089–5093.

[166] H. Zhang, A. Wang, D. Li, and W. Xu, "Deepvoice: A voiceprint-based mobile health framework for parkinson's disease identification," in *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2018, pp. 214–217.

[167] A. Vázquez-Romero and A. Gallardo-Antolín, "Automatic detection of depression in speech using ensemble convolutional neural networks," *Entropy*, vol. 22, no. 6, p. 688, 2020.

[168] M. Muzammel, H. Salam, Y. Hoffmann, M. Chetouani, and A. Othmani, "Audvowelconsnet: A phoneme-level based deep cnn architecture for clinical depression diagnosis," *Machine Learning with Applications*, vol. 2, p. 100005, 2020.

[169] D. Gope and P. K. Ghosh, "Raw speech waveform based classification of patients with als, parkinson's disease and healthy controls using cnn-blstm," 2020.

[170] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[171] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[172] J. Correia, I. Trancoso, and B. Raj, "Detecting psychological distress in adults through transcriptions of clinical interviews," in *International Conference on Advances in Speech and Language Technologies for Iberian Languages (Iberspeech)*. Springer, 2016, pp. 162–171.

[173] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *EMNLP*, vol. 14, 2014, pp. 1532–1543.

[174] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the ACL*, 2004.

[175] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents." in *ICML*, vol. 14, 2014, pp. 1188–1196.

Appendices

Appendix A

Measuring word connotations from word embeddings to detect depression, anxiety and PTSD in clinical interviews

A.1 Motivation

The following Appendix is a short study where we tested the hypothesis that the vocabulary of an individual affected by psychological distress, including depression, anxiety and PTSD, is different from that of those who are not affected, when evaluated in the context of clinical interviews [172].

We consider this work as part of the literature review for the detection of depression and other forms of psychological distress from natural language cues, and one of the contributions accomplished during the thesis towards the state of the art in this field. However, since the experiments do not incorporate in-the-wild data in any capacity, this work is not as relevant as the others presented in the scope of this thesis, therefore it is presented as an Appendix.

In this study, we considered two approaches to evaluate the changes in vocabulary between healthy and distressed speakers. The first took into account small samples of the clinical interview at a time, such as a turn or a sentence, to assess whether that sample belonged to an individual affected by psychological distress. This approach was designed to be used in an online scenario, such as to assist the therapist during the clinical interview. The second was a system that performed an analysis of the full transcription of each interview, taking advantage of larger amounts of information at once. This system was designed to be used in

an offline scenario, after the full interview had been conducted.

In this Appendix we will focus solely on reviewing the offline system when it comes to detecting depression, anxiety and PTSD, as these were the experiments with the most relevant findings, in the context of this thesis proposal.

A.2 Data

This study was performed on the DAIC-F2F, a dataset of audio recordings and transcriptions of clinical interviews of subjects who potentially suffer from psychological distress, conducted by a human therapist. This dataset was previously described in more detail in Chapter 2.2. We randomly partitioned this dataset into a train and development subsets, with 55 and 10 interviews, respectively.

A.3 Proposed approach

The proposed system was based on the computation of a novel type of "conotation" features for each word in the interview based on the relative frequency difference between the use of this word by distressed and healthy individuals. Given a corpus of transcriptions of depressed and healthy individuals the computation of these features was as follows:

- Obtain the vocabulary of the corpus, of size V
- Compute the relative frequency of each word for the subset of the corpus belonging to healthy individuals, and for the one of the depressed people.
- Compute the difference of the relative frequency for each work (the result is a table of size V where a word with a larger absolute value tends to be used more by individuals of one of the classes, healthy or depressed, and a word with a small absolute value is used more or less the same by individuals of the two classes).
- From the table, assign a healthy or distressed "connotation" to each word as a binary label.
- For each word in the vocabulary of the corpus, get their word vector representations.

At test time, for a given interview, for each word:

- Get its word representation vector and find the 20 most similar words from the vocabulary of the training data.

- Refer to the relative frequency difference table and retrieve the "connotation" of the 20 most semantically similar words (semantic similarity is measured as the distance between the global vectors [GloVe] representation of a pair of words, as GloVe maps words into a semantically meaningful space where the distance between words is related to their semantic similarity [173]).
- Establish the "connotation" of the word as the average "connotation"of the k most similar words (here we make $k = 20$).

Finally, average the "connotation" scores for each word over the interview to obtain an interview level score.

The decision threshold could be obtained by using the interview scores to train a simple binary classifier.

A.4 Experiments and results

This approach, as mentioned, was tested on DAIC-F2F, which has manual transcriptions of the interviews. However, we considered that having access to manual transcriptions is not always the most realistic scenario, and that a robust model should be capable of dealing with automatic transcriptions, which contain some errors. To simulate automatic transcriptions in the DAIC-F2F we introduced some noise in the existing manual transcription by replacing 20% of their respective GloVes by a random vector that lies within the same semantic space. In practice this equates to replacing 20% of the correct words in the transcriptions by incorrect ones. We tested the system again under the 20% noise conditions.

The summary of the performance of the system is reported in F1 measure in Table A.1. From it, it can be observed that the model achieves a perfect classification score. It is important to note that since there were only 10 validation interviews there may be some variance in the reported results. Nevertheless, the system performs remarkably well, showing the added value of taking in account longer periods of information at a time. Establishing a parallel with humans, this would resemble how a health care specialist pays attention to the interview as a whole to perform a diagnosis. It can also be seen that this system is robust to noise in the transcriptions, being able to maintain the perfect performance in a 20% noisy scenario.

Finally, by analysing the relative frequency table, computed from the training data, and sorting it by the difference, one can analyze which words correlate most with health and with distress. From Figure A.1, it can be seen that healthy people tend to talk more about casual subjects like school, prom; about relationships and feelings and to laugh more than

Table A.1: Performance in F1 score of the long-term unimodal system with different levels of corruption of the transcription for depression, anxiety and PTSD.

GloVe corruption rate	**Form of distress [acc.]** *PTSD*	*Depression*	*Anxiety*
0%	1.000	1.000	1.000
20%	1.000	1.000	1.000

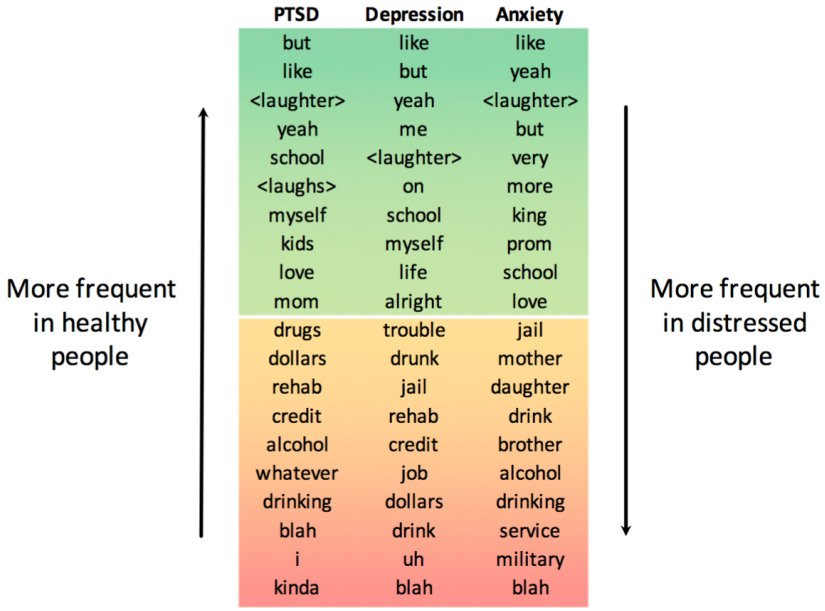


Figure A.1: Examples of words with large relative frequency difference for each label.

distressed people. Conversely, distressed people tend to talk more about traumatic events or topics, such as alcoholism, prison, and drugs, as well as to be generally more uninterested and bored by the conversations, saying "blah" and "uh" much more than healthy people. All these findings correlate well with the general perception of a human of what is a distressed discourse.

Appendix B

Detection of polarity on movie reviews using θ-MIL

B.1 Motivation

In this appendix, we report the experiments and results obtained in the task of inferring the polarity of movie reviews with the IMDb movie review database [174], which was performed in the context of the proposed θ-MIL strategy presented in Section 5.3.

As mentioned, in that Section, the experiments fall out of the scope of this thesis, however it is still relevant to report them in order to demonstrate how the proposed methods work in a practical application.

The motivation for this task was to test the θ-MIL in a scenario where the data would naturally organize into bags, and where the bag label would be determined not by a key positive instance, but by a group of instances.

In the context of movie reviews, the reviews for any given movie are naturally related to each other, as all of them pertain to the same movie. At the same time, the polarity of the movie, whether its considered good or bad overall, is not defined by a single instance, but by a group of them. This scenario fits very well with the θ-MIL framework, hence the choice.

B.2 Data

The polarity dataset is a corpus of movie reviews retrieved from the Internet Movie Database (IMDb) archive [174]. The corpus contains 2000 movie reviews in English, where each review

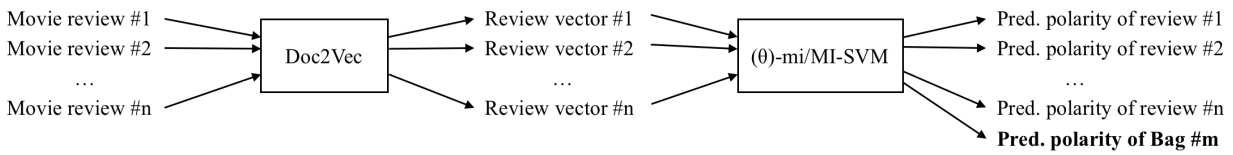


Figure B.1: Proposed θ-MIL framework at test time, to predict the polarity of a movie (bag) and its reviews (instances).

is associated to the movie it refers to, and to a rating expressed by the reviewer in stars or some numerical value. The typical review content is a small text where people summarize the story of the movie and highlight the positive or negative aspects that struck them most.

The movie reviews are determined as positive or negative from their ratting as follows: 1) For ratings specified in 5 star systems, 3.5 stars or more is considered positive, 2 stars or less is considered negative, the remaining, neutral; 2) For 4 stars systems, a rating of 3 stars or more is considered positive, 1.4 or less is considered negative, the remaining, neutral; 3) For letter grade systems, B or above is considered positive, C- or below is considered negative, the remaining, neutral.

The polarity dataset contains 1000 positive reviews, 1000 negative reviews, and no neutral reviews. The 2000 reviews cover 1106 different movies. There are between 1 and 13 reviews per movie.

There are two types of movie-level labels available: the MIL labels, attributed according to the MIL assumption that bags with only negative reviews have a negative bag-level label, and if there is at least one review, then the bag-level label is positive; and the majority labels, attributed to a bag according to which is the most frequent instance label for that bag.

We split the dataset in two subsets, leaving 830 bags (with 1445 instances) for training, and 276 bags (with 555 instances) for testing the models, where the bags are labeled according to the MIL assumption: negative if all instance are negative, and positive if at least one instance is positive. The split guaranteed no movie overlap between the two subsets.

B.3 Features

Each movie review was described by a Doc2Vec features vector. Doc2Vec, or Paragraph Vector [175], emerged as an unsupervised framework that learns a continuous distributed representation for text documents of variable length. From a single sentence to several paragraphs, Doc2Vec can preserve some information related to word ordering. In this

framework, particularly in the distributed memory model, every document is mapped to a unique document vector and every word in the vocabulary is mapped to a unique word vector.

B.4 Experiments and results

In this experimental scenario the bags correspond to sets of movie reviews for a given movie.

The training instances are used to train three classifiers: a fully supervised SVM, a θ-mi-SVM, and a θ-MI-SVM, as described in Section 5.3. The first of the three models aimed at establishing a fully supervised baseline to which we could compare our proposed methods, which in a way, is the upper bound in performance that this approach could obtain if given all the information about the instance labels.

The SVM model will predict the polarity of the test reviews, while the θ-mi-SVM and θ-MI-SVM models will predict both the polarity of the test reviews and their respective bags.

The performance of the three systems against the train and test sets is measured in accuracy and is summarized in Table B.1.

We can see that the performance of the fully supervised SVM for the train and test sets represent the fully supervised upper bound in the performance of these models, since there is no label ambiguity during training. It achieved an accuracy for instance label prediction of 86.8%.

Furthermore, the performance of the θ-mi-SVM and θ-MI-SVM with respect to the accuracy in predicting the instance labels on the test set was 76.9% and 82.9%, respectively. The comparatively poorer performance of θ-mi-SVM method to the remaining methods might be related with the inner workings of the method itself: for each bag, the method selects a fraction of the instances to attribute the bag label to, while the remaining fraction is attributed the opposite label. However, there might be a misclassification of the later set. The alternative, as happens in the θ-MI-SVM method, is to discard the later fraction when estimating the model.

Finally, since the IMDb corpus has bags with different sizes, and many of the bags are of small size, it would be interesting to reevaluate the systems for filtered versions of the corpus, where the smaller bags are discarded. However, we note that since the corpus is small, the subsets of the corpus with large enough bags would become too small to train a robust model.

Table B.1: Performance in accuracy of the the supervised SVM, θ-mi-SVM and θ-MI-SVM for the train and test dataset.

Majority bags	**SVM [acc.%]**	**θ-mi-SVM [acc.%]**		**θ-MI-SVM [acc.%]**	
	inst.	*bags*	*inst.*	*bags*	*inst.*
train	95.85	-	91.06	-	91.62
test	86.85	82.61	76.94	83.33	82.89

Appendix C

Intellectual property and distribution of the WSM Corpus

The WSM Corpus is an audiovisual corpus of videos collected from the online multimedia repository YouTube, mostly featuring recordings in the vlog format, of subjects potentially affected by SA diseases. It contains a total 928 videos, and over 131 hours of speech. Each video in the corpus is accompanied by a crowdsourced annotation for the estimated age, estimated gender, and self-reported health status of the subject in the video. These crowdsourced annotations were obtained through AMT, which were funded by Professor Bhiksha Raj, from Carnegie Mellon University (CMU).

On April 21, 2021, the Center for Technology Transfer and Enterprise Creation (CTTEC), at CMU, has determined that the WSM Corpus' intellectual property, consisting of its annotations (estimated age, estimated gender, and self-reported health status of the subject in the video), and pointers to the videos in the form of URLs, can be distributed for non-commercial purposes, including research and academic purposes, under the creative commons non-commercial license.

With this statement, we include two supporting documents:

- The form submitted to CTTEC, describing, and pointing to the WSM Corpus;
- The email confirmation from Cindy Chepanoske, from CTTEC, this institution's determination regarding the distribution of the WSM Corpus.

Carnegie Mellon University

DISCLOSURE OF INTELLECTUAL PROPERTY SOURCE CODE/ COPYRIGHTS/ APPS

All information requested in this document must be completed in order to expeditiously process this Disclosure. Any missing or incomplete information may delay processing your submission.

1. Title:

the in-the-wild speech medical corpus

2. Creator(s): By signing this form the undersigned Creators acknowledge and agree that they are bound by Carnegie Mellon University's Intellectual Property Policy, on line at http://www.cmu.edu/policies/documents/IntellProp.html. Original signatures for all Creators are required. Therefore, by signing below: (i) if the Policy allows CMU to own this intellectual property and its associated intellectual property rights, you hereby assign to Carnegie Mellon any and all ownership you have in such intellectual property and intellectual property rights; and (ii) if the Policy allows CMU to receive license rights to this intellectual property and its associated intellectual property rights, you hereby grant to CMU any and all such licenses. Original signatures for all Creators are required.

a. Lead Creator:

joana correia	Maria João Rodrigues	03/17/2021
print or type name	signature	date

CS - LTI	(+351) 968029769	joanac@cs.cmu.edu	60
department	phone	current e-mail	% of contribution

Designed and lead the annotation task, reviewed and pre-processed the results.

Nature of contribution to the IP (Briefly explain why this person is a creator)

Student

Carnegie Mellon Employment Status at the time the intellectual property was created (Faculty, Staff, Student, Visitor, Courtesy, etc.)

N/A

Full institutional address (if not affiliated with Carnegie Mellon)

N/A

Full residential address (street, city, state)

N/A

Country of Residence

Portugal

Country of Citizenship

b. Next Creator:

Bhiksha Raj	Bhiksha	14 april 2021
print or type name	signature	date

CS - LTI	412-2686591	bhiksha@cs.cmu.edu	20
department	phone	current e-mail	% of contribution

Oversaw the full process of the dataset creation, from design of annotation task, error corrections, data pre-processing, etc.

Nature of contribution to the IP (Briefly explain why this person is a creator)

Faculty

Carnegie Mellon Employment Status at the time the intellectual property was created (Faculty, Staff, Student, Visitor, Courtesy, etc.)

N/A

Full institutional address (if not affiliated with Carnegie Mellon)

N/A

Full residential address (street, city, state)

USA

Country of Residence

USA

Country of Citizenship

c. Next Creator:

Isabel Trancoso	Isabel Maria Martins Trancoso	
print or type name	signature	date

(external)	(+351) 213100300	isabel.trancoso@inesc-id.pt	20
department	phone	current e-mail	% of contribution

Oversaw the full process of the dataset creation, from design of annotation task, error corrections, data pre-processing, etc.

Nature of contribution to the IP (Briefly explain why this person is a creator)

(external: Instituto Superior Tecnico / INESC-ID)

Carnegie Mellon Employment Status at the time the intellectual property was created (Faculty, Staff, Student, Visitor, Courtesy, etc.)

Rua Alves Redol, n. 9, 1000-029 Lisboa, Portugal

Full institutional address (if not affiliated with Carnegie Mellon)

N/A

Full residential address (street, city, state)

Portugal

Country of Residence

Portugal

Country of Citizenship

d. Next Creator:

print or type name | signature | date

department | phone | current e-mail | % of contribution

Nature of contribution to the IP (Briefly explain why this person is a creator)

Carnegie Mellon Employment Status at the time the intellectual property was created (Faculty, Staff, Student, Visitor, Courtesy, etc.)

Full institutional address (if not affiliated with Carnegie Mellon)

Full residential address (street, city, state)

Country of Residence

Country of Citizenship

Please list additional inventor(s) and relevant information on an additional sheet.

3. Please provide a short description of the function and use of the intellectual property being disclosed in the space below and, if applicable, a copy of the code, link to the code or instructions where it can be accessed.

This corpus consists of a collection of web adresses of youtube videos.
The majority of the youtube videos are vlogs of subjects claiming to suffer from one of several diseases, including depression and Parkinson's disease.
Each web address is accompanied by the following annotations obtained via crowdsourcing:
- an estimate for the subject's age
- an estimate for the subject's gender
- a report of whether the subjects, at any point in the video, claims to suffer from a target disease, such as depression, Parkinson's disease, among others.

The complete corpus, including web addresses and annotations can be found here:
https://www.dropbox.com/sh/f3fff60qhpr1f95/AAASXSJLQtkHOVrSTaigdCyFa?dl=0

4. Intellectual Property Protections

a. This Disclosure describes (please "X" all that apply):

_☐_ Source Code
_☒_ Designs and/or other copyrightable materials
_☐_ Coretech or Coretech improvements (NREC only)

For the question #4b, please indicate the date in the format "Month/Day/ Year" (ex. 01/01/17).

b. State first date of:
a. Completion: 12/31/2020
b. Publication/ Release (outside of CMU): N/A

5. Sponsorship/ Use of 3rd Party Resources

a.

External Sponsor(s)	CMU Oracle #(s) AND	Contract or Grant #(s)

(Your department administrator may be of assistance in identifying funding sources used.)

Have external sponsors been informed of or provided with the intellectual property?

b. If funded under a grant or contract, please describe this intellectual property as one of the following:

_☐_ Background IP (developed prior to the grant but used in research funded by the grant)
_☐_ Background IP Improvement (developed both prior to and during the grant)
_☐_ Foreground IP (developed only during the grant)

c. Internal Sponsor (Department Research Funds, etc.)

d. Was this intellectual property developed in collaboration with any other 3rd parties (companies, universities, etc.) or as a part of a research consortium? Please list below:

Instituto Superior Técnico (University of Lisbon)

e. Have you used any software, libraries, etc. from other internal (e.g., CMU) sources (ex. projects or researchers) in the development of this technology or does the technology otherwise build upon earlier work at CMU? Please list below:

f. Was there any Open Source software, Creative Commons copyrights or other third party material used in the development of this technology or does the technology otherwise build upon earlier work at CMU? Please list below:

6. Have you / do you intennd to release the Source Code pursuant to an Open Source license? If so, please indicate where the code is / will be released (please "X" all that apply):

_☐_ Your website
_☐_ SourceForge, github or other similar hosting provider
_☐_ through a federal agency (e.g. NASA)
_☐_ other (please identify) _____
_☐_ Do NOT intend to release open source

7. How long would it take someone skilled in the art to recreate this copyrightable work? 4-12 mo

Please feel free to attach additional material or data that would provide us with helpful information.

Email the completed electronic copy of this Invention Disclosure form to:

innovation@cmu.edu

If unable to sign electronically, paper copies may be sent to:

Department Administrator
Center for Technology Transfer & Enterprise Creation
4615 Forbes Avenue, Suite 302

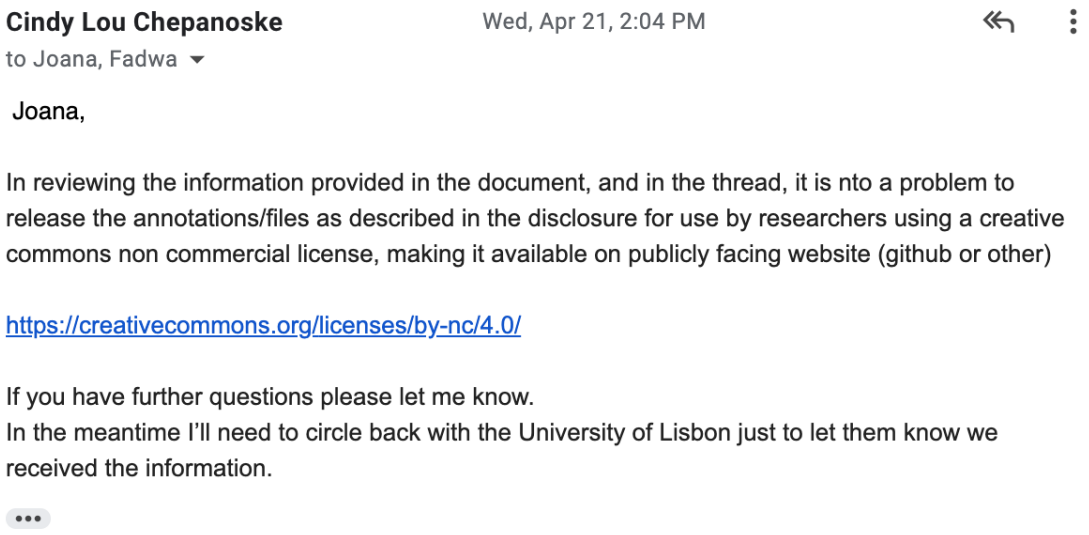


Figure C.1: Email confirmation from the CTTEC regarding the distribution of the WSM Corpus.